Math 331-2: Abstract Algebra
Northwestern University, Lecture Notes

Written by Santiago Cañez

These are notes which provide a basic summary of each lecture for Math 331-2, the second quarter of "MENU: Abstract Algebra", taught by the author at Northwestern University. The book used as a reference is the 3rd edition of *Abstract Algebra* by Dummit and Foote. Watch out for typos! Comments and suggestions are welcome.

Contents

Lecture 1: Introduction to Rings **2**
Lecture 2: Matrix and Polynomial Rings **6**
Lecture 3: More Ring Constructions **8**
Lecture 4: Quotient Rings **14**
Lecture 5: More on Quotients and Ideals **18**
Lecture 6: Maximal Ideals **21**
Lecture 7: Prime Ideals **25**
Lecture 8: Geometry via Algebra **28**
Lecture 9: Rings of Fractions **32**
Lecture 10: More on Fractions **37**
Lecture 11: Chinese Remainder Theorem **43**
Lecture 12: Euclidean Domains **47**
Lecture 13: More on Euclidean **51**
Lecture 14: Principal Ideal Domains **54**
Lecture 15: Unique Factorization **57**
Lecture 16: More on UFDs **60**
Lecture 17: Polynomials Revisited **63**
Lecture 18: Gauss's Lemma **66**
Lecture 19: Eisenstein's Criterion **69**
Lecture 20: Modules over Rings **72**
Lecture 21: Module Homomorphisms **77**
Lecture 22: Sums and Free Modules **80**
Lecture 23: Modules over PIDs **85**
Lecture 24: Structure Theorems **89**
Lecture 25: More on Structure Theorems **92**
Lecture 26: Jordan Normal Form **97**
Lecture 27: More on Normal Forms **103**

Lecture 1: Introduction to Rings

We continue our study of abstract algebra, this quarter focusing on *rings* and (to a lesser extent) *modules*. Rings provide a general setting in which "arithmetic" and the notion of "number" makes sense, and indeed the development of ring theory was borne out of attempts to understand general properties analogous to those of integers. Last quarter we first motivated the study of groups by hinting at their use in describing permutations of roots of polynomials (and in the problem of deciding when a formula for such roots could be found), and rings will now provide the language with which we can begin to discuss these roots themselves. (The full story of these roots belongs to the study of *fields*, which is are special types of rings we will look are more carefully next quarter.) The theory of modules generalizes the subject of linear algebra, and for us will really put to use next quarter when we consider the relation between fields.

Historically, the development of ring theory really took off in the mid-to-late 19th century in the following way. You may have heard of *Fermat's Last Theorem*, which is the result that for $n > 2$ there are no non-trivial solutions to

$$x^n + y^n = z^n.$$

(Solutions such that $x = 1, y = 0, z = 1$ and so on are the *trivial* ones.) This problem has a very rich history, and attempts to provide a valid proof led to much new mathematics. (The final proof was only given a few decades ago, and relies on some very heavy machinery in number theory and algebraic geometry.) One early attempt towards a proof came from trying to exploit a certain factorization: rewrite the given equation as $x^n = z^n - y^n$ and factor the right side:

$$x^n = z^n - y^n = (z - y)(\text{something})(\text{something}) \cdots (\text{something})$$

using something analogous to $z^2 - y^2 = (z - y)(z + y)$ in the $n = 2$ case. The individual factors on the right involve more than just integers, but we are not interested in their precise form here. Then, by trying to "factor" these terms on the right further we can try to come up with something like a "prime factorization" for the x^n on the left, only where we have to be more careful here about what exactly counts as a "prime" in this context. If you could somehow show that there could not be enough "primes" on the right to give an n-th power x^n overall, you would have a proof of Fermat's Last Theorem, and indeed this is what many people in the 19th century thought they had done. (If Fermat did actually have a proof of this result himself, it is very likely that it was phrased along these lines.)

But, it was soon realized that such an argument relied on the *uniqueness* of "prime" factorizations. For instance, in the case of integers we know that $2^4 3^5 7^2$ is not the square of an integer since the 3^5 term is not a square and there are no other primes apart from $2, 3, 7$ we could use to express this number in an alternate way as a potential square, due to the uniqueness of the primes $2, 3, 7$ showing up in this expression. If, however, such a factorization were not unique, we would not be able to derive information about a specific number (such as it not being a square) based solely on the form of one prime factorization alone. This causes a problem for the approach to Fermat's Last Theorem outlined above, since it depends on conclusions made from a single "prime" factorization, and in general in the types of sets of "numbers" (i.e. rings) which show up here, there is no reason to expect that such a uniqueness must hold.

To give an explicit example of where a "prime" factorization can fail to be unique, consider the following set of complex numbers:

$$\mathbb{Z}[\sqrt{-5}] := \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\}.$$

(The notation on the left is pronounced "$\mathbb{Z}$ adjoin $\sqrt{-5}$". We will see what this notation really means soon enough, but it basically describes the structure obtained by taking all possible sums and products of integers and $\sqrt{-5}$ and things built up from these.) Note that adding or multiplying any two elements of $\mathbb{Z}[\sqrt{-5}]$ produces an element still within $\mathbb{Z}[\sqrt{-5}]$, which is essentially why this is a "ring". In this ring, we can factor $6 \in \mathbb{Z}[\sqrt{-5}]$ in the following two ways:

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5}).$$

We claim that both of these are "prime" factorizations of 6! Actually, we should now be more careful with terminology: what really matters is that these are factorizations into *irreducible* elements, where an "irreducible" element is essentially one which cannot be factored any further. (This sounds pretty much like the ordinary definition of "prime" with which you're probably familiar in the context of integers, but we will see later that the term "prime" actually means something different in ring theory; the notions of "prime" and "irreducible" are the same in $\mathbb{Z}$ and many other settings, but not always. Showing that $1+\sqrt{-5}$ and $1-\sqrt{-5}$ are irreducible is a simple brute-force check we will come back to later.) Thus, the factorization of elements of $\mathbb{Z}[\sqrt{-5}]$ into irreducible elements is not unique, so care has to be taken when working with this ring.

Understanding this failure to have unique factorization in certain contexts played an important role in the development of ring theory. It turns out that this lack of uniqueness can be salvaged in many (not all) cases by considering factorizations of what were historically called "ideal numbers", and what we now call *ideals*, instead. The factorization of ideals is not something we will study in this course, since it belongs more properly to a course in (algebraic) number theory, but ideals in general are of central importance in ring theory and this is the setting in which they were first introduced. For our purposes, the point will be that ideals play a role in ring theory analogous to the one played by normal subgroups in group theory.

Rings. We now come to formally defining the notion of a ring. A ring will have two operations, as opposed to a group which only had one, but the simplest way to start is with a group itself. Indeed, a *ring* is an abelian group $(R, +)$—where we think of the group operation $+$ as "addition"—equipped with a second binary operation $\cdot : R \times R \to R$ (which we think of as "multiplication", and usually write simply ab instead of $a \cdot b$) satisfying:

- associativity: $(ab)c = a(bc)$ for all $a, b, c \in R$, and
- distributivity: $a(b + c) = ab + ac$ and $(b + c)a = ba + ca$ for all $a, b, c \in R$.

The distributivity requirement is important since it is the one which expresses a compatibility between the two ring operations—without it, addition and multiplication would just be two random operations with no relation to each other, and that's no fun.

Note that both orders $a(b+c)$ and $(b+c)a$ have to be given explicitly in the definition since the ring multiplication is not assumed to be commutative. Addition *is* assumed to be commutative, simply because every type of "addition" we've ever come across is commutative, which is not the case for "multiplication". (In fact, as the book suggests, that $+$ is commutative can be derived from the ring axioms alone, at least in the case where the ring has a multiplicative identity. We will define this formally in a bit.)

Examples. Many examples of groups we saw last quarter can also be viewed as rings. For instance, $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{R}$ under their usual operations. Also, $\mathbb{Z}/n\mathbb{Z}$ is a (finite) ring under addition and multiplication mod n. The set $\mathbb{Z}[\sqrt{-5}]$ of complex numbers of the form $a+b\sqrt{-5}$ with $a, b \in \mathbb{Z}$ we mentioned in our introductory discussion is also a ring, and indeed it is a *subring* of $\mathbb{C}$, meaning that it is a subgroup with respect to addition which is closed under multiplication.

The group of invertible matrices $GL_n(\mathbb{R})$ over $\mathbb{R}$ is not a ring under usual matrix addition (it is not closed under addition and does not contain zero), but the larger set $M_n(\mathbb{R})$ of *all* $n \times n$ matrices over $\mathbb{R}$ is a ring. We can replace $\mathbb{R}$ here by $\mathbb{Z}$ or $\mathbb{Z}/n\mathbb{Z}$ for instance and still get examples of rings.

Types of rings. We can impose various requirements on the structure of a ring, each encoded via some new terminology. For instance, a ring R is *commutative* if the ring multiplication is commutative:

$$ab = ba \text{ for all } a, b \in R.$$

We say that R has an *identity*, or is a ring with *unity*, if it has a multiplicative identity:

$$\text{there exists } 1 \in R \text{ such that } 1a = a1 = a \text{ for all } a \in R.$$

Note we use "1" to denote the multiplicative identity, even though it might not be literally the number "1". The additive identity (the identity for the abelian group structure under $+$) is usually denoted by 0, as you might expect. One word of caution here is that many sources *assume* as part of the definition of "ring" that it has an identity element, but this varies. Our book does not assume so, and I do not like to either, for reasons we will see later. If a ring we consider is meant to have an identity we will explicitly say so.

If a ring R has unity, we can then talk about multiplicative inverses. A *unit* $a \in R$ is an element with a multiplicative inverse in R:

$$\text{there exists } b \in R \text{ such that } ab = ba = 1.$$

The set of units, denoted $R^\times$, is then a group under the multiplication operation, and is called the *group of units* of R:

$$R^\times := \{a \in \mathbb{R} \mid a \text{ is a unit}\}.$$

In particular, this explains the notation $(\mathbb{Z}/n\mathbb{Z})^\times$ we used last quarter for the multiplicative group of integers mod n—it is precisely the group of units in the ring $\mathbb{Z}/n\mathbb{Z}$.

A nonzero element $a \in R$ is called a *zero divisor* of R if it "divides" zero in a nontrivial way, in the sense that 0 can be obtained as a product of a and some other nonzero element:

$$\text{there exists } 0 \neq b \in R \text{ such that } ab = 0.$$

Note that a zero divisor can never be a unit, since if a has an inverse then $ab = 0$ implies $b = 0$ after multiplying on the left by the inverse of a. An *integral domain* is a commutative ring without zero divisors; the term "integral" comes from the idea that such a ring should be viewed as a generalization of $\mathbb{Z}$.

More examples. The standard examples of $\mathbb{Z}, \mathbb{Q}$, and $\mathbb{R}$ are commutative, have identities, and are integral domains. The ring $\mathbb{Z}/n\mathbb{Z}$ of integers mod n is commutative and has unity, but is not an integral domain unless n is prime: for n not prime, we can find $0 \neq a \in \mathbb{Z}/n\mathbb{Z}$ which is not relatively prime to n, and if d then denotes $gcd(a, n)$ (so in particular $\frac{n}{d}$ and $\frac{a}{d}$ are integers), we have:

$$a(\tfrac{n}{d}) = (\tfrac{a}{d})n \equiv 0 \mathrm{mod},$$

so that a is a zero divisor. (For instance, $2 \cdot 3 = 0 \bmod 6$ and $4 \cdot 3 = 0 \bmod 6$, so $2, 3, 4$ are all zero divisors in $\mathbb{Z}/6\mathbb{Z}$.) For p prime, every nonzero element of $(\mathbb{Z}/p\mathbb{Z})^\times$ is a unit (as we saw last quarter), so there are no zero divisors and $\mathbb{Z}/p\mathbb{Z}$ is an integral domain.

The ring $M_2(\mathbb{R})$ of 2×2 matrices over $\mathbb{R}$ is a noncommutative ring with unity. It has zero divisors, since for instance

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

The group of units of $M_2(\mathbb{R})$ is the group $GL_2(\mathbb{R})$ of invertible matrices we saw last quarter. Now, consider the *subring* S of $M_2(\mathbb{R})$ defined by

$$S := \left\{ \begin{bmatrix} a & 0 \\ 0 & 0 \end{bmatrix} \;\middle|\; a \in \mathbb{R} \right\}.$$

(Again, to be a subring just means it is a subgroup under addition which is also closed under multiplication. It is a straightforward check to see that sums, differences, and products of matrices of the form above are still in S.) The identity element $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ of $M_2(\mathbb{R})$ is not an element of S, but nonetheless S *does* have an identity, namely $\left[\begin{smallmatrix} 1 & 0 \\ 0 & 0 \end{smallmatrix}\right]$. The point is that (multiplicative) identities do not necessarily behave so nicely with respect to taking subrings: a subring might not have an identity even if the larger ring does; a subring might have an identity even if the larger ring does not; and both a subring and the larger ring might have identities which are different. (This is one reason why not requiring that rings have identities is a good convention in my mind.)

Division rings and fields. A *division ring* is a ring with unity in which every nonzero element is a unit: $R^\times = R - \{0\}$. A commutative division ring is called a *field*, and is thus the "nicest" type of algebraic structure possible: both addition and multiplication are commutative, and everything has an additive inverse and (expect for 0) a multiplicative inverse. We should point out that it is common here to assume that $0 \neq 1$, so that the additive and multiplicative identities are distinct. (It is common to assume this even at the level of integral domains as well. Essentially, for technical reasons, we want to exclude the *trivial ring* $\{0\}$ from consideration as a field or integral domain.)

Standard examples of fields are $\mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$. Also, $\mathbb{Z}/p\mathbb{Z}$ for p prime is a (finite) field. We will study fields in more detail next quarter, where we will see that in general roots of a polynomial should be taken to lie in some field. Groups will then entire the picture again when considering certain types of field automorphisms.

Quaternions revisited. The standard example of a division ring which is not a field is the *quaternion ring* $\mathbb{H}$. Last quarter we saw the quaternion group $Q_8 = \{\pm 1, \pm i, \pm j, \pm k\}$, which is the starting point for the quaternion ring. Note that from now on (during this quarter at least) the term "quaterions" alone will refer to elements of this ring instead of the group from last quarter.

As a set we have:

$$\mathbb{H} := \{a + bi + cj + dk \mid a, b, c, d \in \mathbb{R}\}.$$

We define addition in the obvious way (just add as normal and group like-terms), and multiplication using distributivity and the relations among i, j, k we saw last quarter. Since $i^2 = j^2 = k^2 = -1$, we view the quaternions as a generalization of complex numbers, only with more "imaginary" parts. This ring has unity but is non-commutative, since for instance $ij = k$ but $ji = -k$.

However, it is true that every nonzero quaternion has a multiplicative inverse, so that $\mathbb{H}$ is a division ring. If $a + bi + cj + dk$ is nonzero (meaning that at least one of a, b, c, d is nonzero), a direct computation will show that

$$(a + bi + cj + dk)(a - bi - cj - dk) = a^2 + b^2 + c^2 + d^2$$

(view $a - bi - cj - dk$ as something like the "quaternionic conjugate" of $a + bi + cj + dk$), which means that

$$\frac{a - bi - cj - dk}{a^2 + b^2 + c^2 + d^2}$$

is the multiplicative inverse of $a + bi + cj + dk$. (Again, this is similar to what happens for complex numbers, where $\frac{a-ib}{a^2+b^2}$ is the inverse of $a + ib$.)

Lecture 2: Matrix and Polynomial Rings

Warm-Up 1. Suppose R is a ring. We show that $0a = 0$ for all $a \in R$, and that if R has unity, then $(-1)a = -a$ for all $a \in R$. These are not as obvious as they might seem, since of course "0" and "-1" do not necessarily denote the numbers 0 and -1 we are used to, but rather specific elements arising from the ring structure at hand. Indeed, 0 is defined by an *additive* property, whereas $0a = 0$ is a claim about a *multiplicative* one, and similarly -1 is the *additive* inverse of the *multiplicative* identity, so the fact that *multiplying* it by a produces the *additive* inverse of a really is saying something noteworthy. Ultimately, the key point in both of these results is to highlight the utility of the distributive properties.

First, for $a \in R$, we have:

$$0a = (0 + 0)a = 0a + 0a.$$

(Note $0 = 0 + 0$ solely from the definition of 0 as an additive identity, irrespective of whether it is anything like the "number" 0.) Whatever $0a \in R$ is, it has an additive identity by the definition of "ring", and adding this additive identity to both sides results in $0 = 0a$ as desired. (From now on, we will use the phrase "subtraction" to refer to "addition of an additive inverse".) Next, since $1 + (-1) = 0$ and $1a = a$, we have:

$$a + (-1)a = 1a + (-1)a = [1 + (-1)]a = 0a = 0,$$

which shows $(-1)a$ satisfies the property required of $-a$. Since additive inverses are unique, we thus get $(-1)a = -a$ as claimed.

Warm-Up 2. Suppose R is a ring such that $a^2 = a$ for all $a \in R$. (Such rings are called *Boolean rings*.) We show that $x + x = 0$ for all $x \in R$, and then that R is commutative. If $x \in R$, then $(x + x)^2 = x + x$ by the given property of R, and expanding this out (using distributivity) gives:

$$x^2 + x^2 + x^2 + x^2 = x + x.$$

But $x^2 = x$, so this becomes $x + x + x + x = x + x$, and subtracting x on the left and right gives $x + x = 0$ as claimed.

Now, let $x, y \in R$. Then $(x + y)(x + y) = x + y$ by the property which R has, and so

$$x^2 + xy + yx + y^2 = x + y.$$

(Note that order matters here: when we distribute in $(x+y)(x+y)$, the x from the first factor and y from the second gives xy, whereas the y in the first and x in the second gives yx, and these are not a priori the same in general.) Again $x^2 = x$ and $y^2 = y$ reduces the equality above to

$$x + xy + yx + y = x + y,$$

and subtracting gives $xy + yx = 0$, which says that $yx = -(xy)$. But also, $xy + xy = 0$ by the first part of this problem, so $xy = yx$ since additive inverses are unique. Hence R is commutative.

Cancellation in integral domains. An important observation about integral domains is that they have a certain cancellation property, which we will see come up in the following result. Suppose R is a finite integral domain. We claim that R is then actually a field, meaning that every nonzero element should have a multiplicative inverse.

Concretely, let us list all the elements of R as:

$$R = \{0, 1, a_1, \ldots, a_n\}.$$

Pick some nonzero $r \in R$, and consider the products

$$r,\ ra_1, , \ldots,\ ra_n$$

obtained by multiplying r by each nonzero element of R. We claim that these elements are all distinct: if $ra_i = ra_j$, then $a_i = a_j$. This is the "cancellation" property referred to above, since we are in effect "cancelling" out the r factors in $ra_i = ra_j$. This is certainly true if r already has a multiplicative inverse, since multiplying by r^{-1} will achieve this cancellation, but the point is that this works even without an inverse by the lack of zero divisors alone. Indeed, we can rewrite $ra_i = ra_j$ as $r(a_i - a_j) = 0$, and since r is nonzero it must be the case that $a_i - a_j = 0$ since otherwise r would be a zero divisor. Hence $a_i = a_j$ as claimed.

Thus, the elements $r, ra_1, \ldots, ra_n \in R$ are $n + 1$ many distinct nonzero elements of R, and so must be precisely (as a set) equal to $1, a_1, \ldots, a_n$ in some order. In particular, $ra_i = 1$ for some i, so r does have a multiplicative inverse, and hence we conclude that R is a field. This still works for a non-commutative unital (meaning it has unity) finite ring without zero divisors, only that we get a division ring in that case.

Matrix rings. Just as we can consider matrices with real entries, we can more generally consider matrices with entries in a ring. If R is a commutative ring with unity, then $M_n(R)$ denotes the ring of $n \times n$ matrices with entries in R. Addition and multiplication in this ring are defined in the ordinary way, where the point is that the fact we can add and multiply elements of R together using the given ring operations is what makes this work out. In other words, computing a product of matrices requires us to use expressions of the form $a_{i1}b_{1j} + a_{i2}b_{2j} + \cdots + a_{in}b_{nj}$ (imagine a_{ij} are the entries of one matrix and b_{ij} those of another), and such an expression makes sense in any ring. (The restriction that R be commutative and have an identity is just a common convention, and the definition would work just as well for other rings. But, we will stick with this convention. In particular, if we want there to be an "identity" matrix, R had better have an identity itself.)

The group of units of $M_n(R)$ is denoted $GL_n(R)$, and generalizes some matrix group we saw last quarter. Concretely, in order for $A \in M_n(R)$ to be invertible requires that its determinant (computed using the usual methods) be a *unit* in R:

$$GL_n(R) = \{A \in M_n(R) \mid \det A \in R^\times\}.$$

This comes from the fact that if A^{-1} exists *within* $M_n(R)$, then taking determinants of both sides of $AA^{-1} = I$ and using standard determinant properties gives

$$(\det A)(\det A^{-1}) = 1,$$

which means that $\det A \in R$ has multiplicative inverse $\det A^{-1} \in R$. (Recall that the value of $\det A$ in the end is some expression made solely of sums and products of the entries of A, which is why it is an element of R itself.) The fact that $\det A \in R^\times$ is enough to guarantee invertibility of A *over* R comes from the fact that there is a general formula for A^{-1} analogous to the one in the

2×2 case which involves multiplying by $\frac{1}{\det A} = (\det A)^{-1}$. The precise formula needed will not be important for us.

For instance, even though $\left[\begin{smallmatrix} 1 & 2 \\ 3 & -1 \end{smallmatrix}\right] \in M_2(\mathbb{Z})$ has an inverse in $M_2(\mathbb{Q})$, it does not have an inverse in $M_2(\mathbb{Z})$ itself since the candidate inverse has non-integer entries. The determinant of $\left[\begin{smallmatrix} 1 & 3 \\ 2 & 2 \end{smallmatrix}\right]$ over $\mathbb{Z}/6\mathbb{Z}$ is 2 (since $-4 = 2 \bmod 6$), which is not a unit in $\mathbb{Z}/6\mathbb{Z}$, so this matrix is not invertible in the ring $M_2(\mathbb{Z}/6\mathbb{Z})$.

Polynomial rings. Another general class of rings we will consider—arguably more important than matrix rings—are *polynomial rings*. Again, we start with a commutative ring R with unity. The ring $R[x]$ (pronounced "R adjoin x") consists of polynomials with coefficients in R:

$$R[x] := \{a_0 + a_1 x + \cdots + a_n x^n \mid n \geq 0 \text{ and each } a_i \in R\}.$$

We define addition and multiplication of these polynomials as you normally would (in particular by "expanding things out" when multiplying), only that we use the rings operations of R when computing sums and products of the coefficients.

Here we are considering the symbol x to denote simply a "formal variable" with no other meaning nor interpretation. Think of it simply as a placeholder used to distinguish between the coefficient a_1 and the coefficient a_2, between a_2 and a_4, and so on. In other words, we are not necessarily treating a polynomial as a *function* where x with x something to be substituted in for. This distinction is important: the polynomial $x^2 + x$ in $(\mathbb{Z}/2\mathbb{Z})[x]$ is not the zero polynomial (the zero polynomial is the one which has all coefficients being zero), but if did treat this as a function on $\mathbb{Z}/2\mathbb{Z}$ we would have $x^2 + x = 0$ for all $x \in \mathbb{Z}/2\mathbb{Z}$, so that it would represent the zero function. We should consider polynomials to be (formal) objects of study in their own right, irrespective of other interpretations we might be able to give them in other settings.

We can also consider multivariable polynomial rings:

$$R[x_1, \ldots, x_n] := \{\text{polynomials in the variables } x_1, \ldots, x_n \text{ with coefficients in } R\}.$$

For instance, $2x^2 + 4xy^3 + 3x^4y^2$ is an element of $\mathbb{R}[x, y]$. Again, addition and multiplication are defined in the usual way.

Lecture 3: More Ring Constructions

Warm-Up 1. Suppose R is an integral domain. We show that $R[x]$ is also an integral domain, and then more generally that $R[x_1, \ldots, x_n]$ is an integral domain. First, in the single-variable case, write $p(x), q(x) \in R[x]$ as

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n \quad \text{and} \quad q(x) = b_0 + b_1 x + \cdots + b_m x^m.$$

A first approach is to suppose $p(x)q(x) = 0$ with $p(x) \neq 0$, in which case we need to show that $q(x) = 0$. This can be done by writing out $p(x)q(x)$:

$$p(x)q(x) = a_0 b_0 + (a_0 b_1 + a_1 b_0)x + (a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + \cdots + a_n b_m x^{n+m}.$$

If this is to be zero, every coefficient must be zero, so in particular the leading coefficient (i.e. coefficient of the highest-order term) must be zero:

$$a_n b_m = 0.$$

We can assume $a_n \neq 0$ since $p(x) \neq 0$ and otherwise the $a_n x^n$ would not be present in the expression for $p(x)$, so since R is an integral domain we must have $b_m = 0$. And so on, by figuring out the expressions for all the other coefficients and working our way down—looking next at the coefficient of the x^{n+m-1} term—we can show that $b_{m-1} = 0$, then $b_{m-2} = 0$, etc.

This works, but is a bit tedious since it requires knowing the precise expressions for each coefficient. A second approach is to rephrase the integral domain condition: if $p(x)$ and $q(x)$ are both nonzero, then $p(x)q(x)$ should be nonzero. This is simpler: assume $a_n \neq 0$ and $b_m \neq 0$ in the expressions above for $p(x)$ and $q(x)$ (otherwise don't include these terms in these expressions), so that then the coefficient $a_n b_m$ of x^{n+m} in the product $p(x)q(x)$ is nonzero since R is an integral domain. Thus $p(x)q(x) \neq 0$ since at least $a_n b_m x^{n+m}$ is present, so $R[x]$ is an integral domain.

Now, the multivariable case can be approached similarly, but it requires some care in dealing with coefficients of multivariable products. But, this is unnecessary, since the point is that we can view $R[x_1, \dots, x_n]$ as being built out of $R[x_1]$ by adjoining successive variables, one at a time. For instance, we can consider $R[x_1, x_2]$ alternatively as the ring of polynomials in the variable x_2 *over* the coefficient ring $R[x_1]$:

$$(R[x_1])[x_2] = \{c_0 + c_1 x_2 + c_2 x_2^2 + \cdots + c_t x_2^t \mid t \geq 0 \text{ and each } c_i \in R[x_1]\}.$$

If we write each c_i as a polynomial in the variable x_1, we can then turn the polynomial expression above into a polynomial in the variables x_1, x_2; for instance,

$$\underbrace{(1 + 2x_1)}_{\in \mathbb{Z}[x_1]} + \underbrace{(2x_1 - x_1^2)}_{\in \mathbb{Z}[x_1]} x_2 + \underbrace{(1 - x_1^4)}_{\in \mathbb{Z}[x_1]} x_2^2 = 1 + 2x_1 + 2x_1 x_2 - x_1^2 x_2 + x_2^2 - x_1^4 x_2^2 \in \mathbb{Z}[x_1, x_2].$$

In general, $R[x_1, \dots, x_n] = (((R[x_1])[x_2]) \cdots)[x_n]$ is a polynomial ring over a polynomial ring over a polynomial ring, etc.

Thus, since R is an integral domain, so is $R[x_1]$. But then so is $(R[x_1])[x_2]$ by applying the single-variable case to the integral domain $R[x_1]$, and then so is $((R[x_1])[x_2])[x_3]$, and so on. In other words, by induction we may assume that $R[x_1, \dots, x_{n-1}]$ is an integral domain, and then by the single-variable case (which is the base case) we have that $(R[x_1, \dots, x_{n-1}])[x_n] = R[x_1, \dots, x_n]$ is also an integral domain.

Warm-Up 2. The converse of the result above is true—namely that if $R[x]$ is an integral domain, so is R—since R is a subring of $R[x]$ (constant polynomials) and subrings of integral domains are integral domains. Thus, since $\mathbb{Z}/6\mathbb{Z}$ is not an integral domain, neither is $(\mathbb{Z}/6\mathbb{Z})[x]$. We will give a few examples of zero divisors in this polynomial ring.

First, any zero divisors in $\mathbb{Z}/6\mathbb{Z}$ are also zero divisors in $(\mathbb{Z}/6\mathbb{Z})[x]$ by considering them as constant polynomials:

$$2 \cdot 3 = 0 \qquad \text{and} \qquad 4 \cdot 3 = 0.$$

If we want some nonconstant examples, we can consider something like

$$2x \cdot 3x = 0 \quad \text{and} \quad 4x \cdot 3x = 0$$

since $6x^2$ and $12x^2$ are both zero over $\mathbb{Z}/6\mathbb{Z}$ since they have coefficients equal to zero. If we want zero divisors of the form, say, $a + bx$, we can expand out something like

$$(a + bx)(c + dx) = 0$$

and see what conditions this imposes on $a, b, c, d \in \mathbb{Z}/6\mathbb{Z}$. We get that ac, bd, and $ad + bc$ all have to be zero, so in particular each of a, b, c, d is necessarily a zero divisor in $\mathbb{Z}/6\mathbb{Z}$. From this we can find an example like

$$(3x + 3)(2x + 4) = 0 \text{ over } \mathbb{Z}/6\mathbb{Z}.$$

And of course we can find higher-order zero divisors as well.

One thing to note is that the zero divisor $3x + 3$ can simply be multiplied by 2 to get 0, and $2x + 4$ can be multiplied by 3, so that the equality $(3x + 3)(2x + 4) = 0$ is not the simplest way to see that these are zero divisors. In fact, we can describe all zero divisors in $(\mathbb{Z}/6\mathbb{Z})[x]$ in general as those polynomials $p(x)$ for which there exists a *constant* $a \in \mathbb{Z}/6\mathbb{Z}$ such that $ap(x) = 0$. An analogous result is true for other coefficient rings, which we will leave for you to think about.

Related constructions. We now give a sense for why considering polynomial rings will be a nice way to think about other types of constructions more generally. For instance, back on the first day we used the notation $\mathbb{Z}[\sqrt{-5}]$ to denote the subring of $\mathbb{C}$ consisting of all complex numbers of the form $a + b\sqrt{-5}$ with $a, b \in \mathbb{Z}$. The bracket notation used in $\mathbb{Z}[\sqrt{-5}]$ is no accident, since this can be considered to be something analogous to a polynomial ring. Namely, if we take $\mathbb{Z}[x]$ but replace x by $\sqrt{-5}$, we get:

$$\mathbb{Z}[\sqrt{-5}] = \{a_0 + a_1\sqrt{-5} + a_2\sqrt{-5}^2 + \cdots + a_n\sqrt{-5}^n \mid n \geq 0 \text{ and each } a_i \in \mathbb{Z}\}.$$

The difference here—as opposed to $\mathbb{Z}[x]$—is that we *can* give some meaning to $\sqrt{-5}$ and its powers by interpreting them not only as formal "symbols" but indeed as complex numbers. Then $\sqrt{-5}^2 = -5, \sqrt{-5}^3 = -5\sqrt{-5}$, and so on, so that all higher-order powers of $\sqrt{-5}$ in

$$a_0 + a_1\sqrt{-5} + a_2\sqrt{-5}^2 + \cdots + a_n\sqrt{-5}^n$$

reduce down to a simpler one, which is why we can more compactly describe these elements via $a + b\sqrt{-5}$ alone. We can think of both $\mathbb{Z}[x]$ and $\mathbb{Z}[\sqrt{-5}]$ as rings *generated* by $\mathbb{Z}$ and a single non-integer element, only that in the latter case the usual interpretation we give to $\sqrt{-5}$ allows us to describe the resulting expressions in a simpler way.

Along the same lines, $\mathbb{Q}[\sqrt{2}]$ ("$\mathbb{Q}$ adjoin $\sqrt{2}$") denotes the subring of $\mathbb{R}$ consisting of real numbers of the form

$$a_0 + a_1\sqrt{2} + a_2\sqrt{2}^2 + \cdots + a_n\sqrt{2}^n$$

with $a_i \in \mathbb{Q}$ (i.e. "polynomials" in $\sqrt{2}$), but which can be more simply rewritten as $a + b\sqrt{2}$ for some $a, b \in \mathbb{Q}$. Elements of $\mathbb{Q}[\pi] \subseteq \mathbb{R}$, on the other hand, cannot be written in any way simpler than

$$a_0 + a_1\pi + a_2\pi^2 + \cdots + a_n\pi^n.$$

For an example with two "generators", consider

$$\mathbb{Q}[\sqrt{2}, \sqrt{3}] = \{a + b\sqrt{2} + c\sqrt{3} + d\sqrt{6} \mid a, b, c, d \in \mathbb{Q}\}.$$

Here, we take "polynomials" in the two "variables" $\sqrt{2}, \sqrt{3}$, and simplify the resulting expressions using the fact that $\sqrt{2}^2 = 2$ and $\sqrt{3}^2 = 3$. The need to include $\sqrt{6}$ in the expression comes from products like $\sqrt{2}\sqrt{3}$, and no other terms are needed since for instance $\sqrt{2}\sqrt{6}$ can be written solely in terms of $\sqrt{3}$, and $\sqrt{3}\sqrt{6}$ can be written in terms of $\sqrt{2}$.

If R is an integral domain, we use $R(x)$ to denote the *field of rational functions* over R, which consists of quotients of polynomials over R;

$$R(x) := \left\{ \frac{p(x)}{q(x)} \;\middle|\; p(x), q(x) \in R[x],\ q(x) \neq 0) \right\}.$$

For instance, $\frac{x^2+1}{x^3-2x+1}$ is an element of $\mathbb{R}(x)$, although it can also be viewed as an element of $\mathbb{Q}(x)$ or of $\mathbb{Z}(x)$. Addition and multiplication on $R(x)$ are defined as they usually are for ordinary fractions. In general, $R(x)$ is indeed a field, where the inverse of $\frac{p(x)}{q(x)}$ is $\frac{q(x)}{p(x)}$. We will discuss the relation between $R[x]$ and $R(x)$ more carefully later, where the point is that $R(x)$ is essentially the "smallest" field which contains $R[x]$.

We can also consider "substituting" in for x in such fields of fractions, as we did with polynomials above. For instance, $\mathbb{Q}(\sqrt{2})$ is the field whose elements are of the form

$$\frac{a+b\sqrt{2}}{c+d\sqrt{2}}$$

with $a, b, c, d \in \mathbb{Q}$, which come from taking quotients of elements of $\mathbb{Q}[\sqrt{2}]$. In this case, however, we can actually rewrite such a quotient to make it look like exactly like an element of $\mathbb{Q}[\sqrt{2}]$, so that $\mathbb{Q}[\sqrt{2}]$ and $\mathbb{Q}(\sqrt{2})$ actually mean the same thing. (Not true of $R[x]$ and $R(x)$ in general. For instance, $\mathbb{Q}[\pi]$ is not the same as $\mathbb{Q}(\pi)$, since quotients of polynomials in π cannot be expressed as single polynomials in π.) Indeed, we have:

$$\frac{a+b\sqrt{2}}{c+d\sqrt{2}} = \frac{a+b\sqrt{2}}{c+d\sqrt{2}}\left(\frac{c-d\sqrt{2}}{c-d\sqrt{2}}\right) = \left(\frac{ac-2bd}{c^2+2d^2}\right) + \left(\frac{bc-ad}{c^2+2d^2}\right)\sqrt{2} \in \mathbb{Q}[\sqrt{2}].$$

It is also possible to show that $\mathbb{Q}(\sqrt{2}, \sqrt{3})$—quotients of elements of $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$—is just $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ itself (i.e. $\mathbb{Q}[\sqrt{2}, \sqrt{3}]$ is already a field), although this takes some more work. We will come back to this from a more general perspective next quarter; the key point is that things like $\sqrt{2}, \sqrt{3}$ are "algebraic" (a term we will define next quarter) over $\mathbb{Q}$, as opposed to something like π for instance, which is not.

Formal series. A polynomial only involves finitely many terms by definition, but we can discuss "infinite polynomials" using the language of *formal power series*. A formal power series over R is an expression of the form

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots$$

where each $a_i \in R$. To be clear, you may have seen power series before in a calculus (or analysis) course, where you dealt with questions of convergence, but here we are not concerned with convergence at all (which might not make sense over an arbitrary ring) and are treating $\sum_{n=0}^{\infty} a_n x^n$ solely as a "formal" expression, and *not* one which is meant to possibly represent a function. We add formal power series in the usual way by adding matching coefficients:

$$(a_0 + a_1 x + a_2 x^2 + \cdots) + (b_0 + b_1 x + b_2 x^2 + \cdots) = (a_0 + b_0) + (a_1 + b_1)x + (a_2 + b_2)x^2 + \cdots,$$

and we multiply formal power series by using the distributive property and grouping together like terms:

$$(a_0 + a_1 x + a_2 x^2 + \cdots)(b_0 + b_1 x + b_2 x^2 + \cdots) = a_0 b_0 + (a_0 b_1 + a_1 b_0)x + (a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + \cdots.$$

Explicitly, the coefficient of x^n in the product $(\sum_{n=0}^{\infty} a_n x^n)(\sum_{n=0}^{\infty} b_n x^n)$ is

$$\sum_{k=0}^{n} a_k b_{n-k}.$$

The resulting ring is denoted by $R[[x]]$ and is called the *ring of formal power series* over R. Note that we can view the polynomial ring $R[x]$ as a subring of $R[[x]]$ by considering power series whose coefficients are all eventually zero.

Just as quotients of polynomials gave a field (the field of rational functions), we can take quotients of power series to get a field as well. (We will assume here that we are working over a field F, not just a random commutative ring; we will see why this matters later.) Such a quotient is called a *formal Laurent series*, and the resulting field is denoted by $F((x))$:

$$F((x)) := \left\{ \frac{\sum_{n=0}^{\infty} a_n x^n}{\sum_{n=0}^{\infty} b_n x^n} \;\middle|\; a_i, b_i \in R, \sum_{n=0}^{\infty} b_n x^n \neq 0 \right\}.$$

We will say more about this field a bit later, and in particular argue that we can express an element in it more simply as a "power series" where we allow negative exponents on x:

$$\sum_{n=-N}^{\infty} a_n x^n = \frac{a_{-N}}{x^N} + \frac{a_{-N+1}}{x^{N-1}} + \cdots + \frac{a_{-1}}{x} + a_0 + a_1 x + \cdots$$

where $N \geq 0$. (If you have had a course in *complex analysis* before, you would have seen the notion of a Laurent series there, and indeed our notion of a formal Laurent series is related, only that, again, here we care nothing about convergence nor divergence.) One reason why we are introducing these types of rings now is to setup a nice way to describe some other examples of rings later on. An underlying goal of this course is to give you a sense of how widespread ring theory is, in that rings show up in most other areas of mathematics: algebra (duh), analysis, geometry, topology, probability, and so on. The language we are building up now will make describing such examples of rings later on easier.

Ring homomorphisms. To give one way in which we can think about the structure encoded in a power series ring, and an alternate way of viewing polynomial rings, it is now appropriate to introduce the notion of a ring homomorphism. The definition is, of course, exactly analogous to what we had for groups before, only now that we have two operations we would like to preserve:

> A function $\phi : R \to S$ between rings R, S is a *ring homomorphism* if it preserves both the additive and multiplicative structures:
>
> $$\phi(a+b) = \phi(a) + \phi(b) \quad \text{and} \quad \phi(ab) = \phi(a)\phi(b)$$
>
> for all $a, b \in R$. As usual, take note that the addition and multiplication on the left of both equalities takes place in R, whereas the addition and multiplication on the right sides takes place in S.

Moreover, ϕ is a *ring isomorphism* if in addition it is injective and surjective, in which case we saw that R and S are *isomorphic* as rings and write $R \cong S$.

Examples are plentiful, and indeed many examples we saw last quarter in the context of groups are *also* examples of ring homomorphisms, particularly in the case of $\mathbb{Z}/n\mathbb{Z}$ or $\mathbb{Z}$. For instance, we know that if we consider the additive group structure, we have

$$\mathbb{Z}/mn\mathbb{Z} \cong \mathbb{Z}/n\mathbb{Z} \times \mathbb{Z}/m\mathbb{Z}$$

when m, n are relatively prime, but in fact this also holds when considering both sides as *rings*. (The product of rings is defined in an analogous way to the product of groups, which we will clarify

in the example below.) That is, not only can we come up with a bijective function between the two structures above which preserves addition, we can also find one which will preserve multiplication too. We will postpone a proof of this until we can derive it from a more general result (the Chinese Remainder Theorem) later. One thing to note is that, although the problem of classifying groups up to isomorphism played a big role last quarter, that will not be the case in our study of rings. For sure we will in certain instances care about whether one ring is isomorphic to another, but not to the extend we did with groups. One reason for this is that this is perhaps only likely possible to do for finite rings, and finite rings will not be an example we will care much about, except for in the case of finite fields. (We will in fact classify all finite fields.)

Back to series. Coming back to the ring of formal power series over a ring R, we might note that the data of a specific formal power series

$$\sum_{n=0}^{\infty} a_n x^n = a_0 + a_1 x + a_2 x^2 + \cdots$$

is fully encoded in the collection of coefficients $a_0, a_1, a_2, \ldots \in R$. That is, we can define a bijection between $R[[x]]$ and the *product* of countably-infinite many copies of R:

$$R \times R \times R \times \cdots := \{(a_0, a_1, a_2, \ldots) \mid a_i \in R\}.$$

This product is analogous to products we saw before, only that here we allow infinitely-many components, and can think of $(a_0, a_1, a_2, \ldots)$ as an infinitely-long "vector". The bijection

$$R[[x]] \to R \times R \times R \times \cdots$$

simply sends a formal power series to its vector of coefficients. By considering the usual direct product *additive* group structure on $R \times R \times R \times \cdots$ (where addition is performed componentwise), we get that $R[[x]]$ and $R \times R \times R \times \cdots$ are isomorphic as *additive groups*.

However, the difference between the two rings is apparent in their multiplications. Multiplication in the direct product ring is also defined componentwise:

$$(a_0, a_1, a_2, \ldots)(b_0, b_1, b_2, \ldots) = (a_0 b_0, a_1 b_1, a_2 b_2, \ldots),$$

which is *not* how multiplication is defined in the power series ring. Thus, $R[[x]]$ and $R \times R \times R \times \cdots$ are *not* isomorphic as rings (at least under the map which sends a power series to its sequence of coefficients), even though they are isomorphic as additive groups. This highlights that, in some sense, what makes a ring a "ring" is really the multiplication, not so much the addition. Now, we *can* define an alternate multiplication on the additive group $R \times R \times R \times \cdots$ so that it explicitly does mimic the multiplication on the power series ring by setting:

$$(a_0, a_1, a_2, \ldots)(b_0, b_1, b_2, \ldots) = (a_0 b_0, a_0 b_1 + a_1 b_0, a_0 b_2 + a_1 b_1 + a_2 b_0, \ldots),$$

in which case the resulting ring will be isomorphic to $R[[x]]$. (We would not call this the "direct product" ring structure however, which is used only considering the componentwise operations.) Essentially, the entire point of introducing power series at all is to give a simpler way to encode *this* specific ring structure, since remembering

$$(a_0, a_1, a_2, \ldots)(b_0, b_1, b_2, \ldots) = (a_0 b_0, a_0 b_1 + a_1 b_0, a_0 b_2 + a_1 b_1 + a_2 b_0, \ldots)$$

is not as straightforward as remembering

$$(a_0 + a_1x + a_2x^2 + \cdots)(b_0 + b_1x + b_2x^2 + \cdots) = a_0b_0 + (a_0b_1 + a_1b_0)x + (a_0b_2 + a_1b_1 + a_2b_0)x^2 + \cdots,$$

which really just comes from the usual way of multiplying "polynomials".

Since we can view $R[x]$ as a subring of $R[[x]]$, we can ask to what subring of $R \times R \times R \times \cdots$ the polynomial ring $R[x]$ should bijectively correspond. The answer is given by what's called the *direct sum* $R \oplus R \oplus R \oplus \cdots$, which is the set of elements of the direct product which have only finitely many nonzero components. Said another way, it is the the set of "vectors" whose entries are eventually zero:

$$R \oplus R \oplus R \oplus \cdots := \{(a_0, a_1, a_2, \ldots) \mid a_i \in R \text{ and } \exists N \text{ such that } a_k = 0 \ \forall k > N\}.$$

This corresponds to the fact that the coefficients of a polynomial, viewed as a power series, will be all zero past some index. As an additive group, the direct sum of countably-infinite many copies of R is isomorphic to $R[x]$, but again *not* as a ring if we take the multiplicative structure into account.

We can more define more general direct products

$$\prod_{i=1}^{\infty} R_i := R_1 \times R_2 \times R_3 \times \cdots$$

(if you have never seen the product $\prod$ notation before, simply view it as analogous to the summation notation Σ) and direct sums

$$\bigoplus_{i=1}^{\infty} R_i := R_1 \oplus R_2 \oplus R_3 \oplus \cdots$$

of rings $R_1, R_2, R_3, \ldots$ in analogous ways. We will use these structures from time to time, but will not say much more about their "real" meaning. The distinction that in the direct sum case we should only consider elements which have finitely many nonzero components is a concept that shows up elsewhere in mathematics, such as in the notion of the "product topology" you would learn about in a topology course such as MATH 344-1. There is an underlying idea that ties all of these concepts together, but that belongs to the subject of *category theory* and is far beyond the scope of this course. We only mention this here to give the sense that what we are studying does not exist in a vacuum, and absolutely has connections to other areas of mathematics.

Lecture 4: Quotient Rings

Warm-Up 1. We determine all rings homomorphisms $\phi : \mathbb{Z} \to \mathbb{Z}$. When first considering only the additive structure, since 1 generates $\mathbb{Z}$ as an additive group, ϕ is completely determined by the value of $\phi(1) \in \mathbb{Z}$. Without any other restrictions, $\phi(1)$ can be anything, which is why there there are infinitely many *group* homomorphisms $\mathbb{Z} \to \mathbb{Z}$, one for each possible integer value of $\phi(1)$.

But now throwing in the multiplicative structure imposts more restrictions. In particular, the fact that $1^2 = 1$ implies that

$$\phi(1)^2 = \phi(1),$$

so $\phi(1)$ must be an element of $\mathbb{Z}$ which squares to itself. Only 1 and 0 have this property, where one "ring-theoretic" way to see this is as follows: $\phi(1)^2 = \phi(1)$ is the same as $\phi(1)(\phi(1) - 1) = 0$, and $\mathbb{Z}$ being an integral domain implies that either $\phi(1) = 0$ or $\phi(1) - 1 = 0$ as claimed. Thus there are two ring homomorphisms $\mathbb{Z} \to \mathbb{Z}$, corresponding to $\phi(1) = 1$—in which case $\phi(n) = n$ for all n—and $\phi(1) = 0$, in which case $\phi(n) = 0$ for all n.

Warm-Up 2. Let $\phi : \mathbb{Q}[x] \to \mathbb{R}$ be the homomorphism defined by sending any rational number to itself and setting $\phi(x) = \pi$. This determines all of ϕ using the preservation of addition and multiplication properties, as we will see. We determine the image and kernel of ϕ, which are defined just as they were for groups:

> The *image* of $\phi : R \to S$ is the subset $\phi(R)$ of S consisting of all elements obtained by applying ϕ to elements of R, and the *kernel* of ϕ is the set $\ker \phi$ of all things in R which are mapped to zero:
>
> $$\phi(R) := \{\phi(a) \in S \mid a \in R\}, \ \ker \phi := \{a \in R \mid \phi(a) = 0\}.$$
>
> It is a straightforward to check that $\phi(R)$ is subring of S, and $\ker \phi$ a subring of R.

We can determine all values of ϕ explicitly as follows:

$$\begin{aligned}\phi(a_0 + a_1x + a_2x^2 + \cdots + a_nx^n) &= \phi(a_0) + \phi(a_1x) + \phi(a_2x^2) + \cdots + \phi(a_nx^n) \\ &= \phi(a_0) + \phi(a_1)\phi(x) + \phi(a_2)\phi(x)^2 + \cdots + \phi(a_n)\phi(x)^n \\ &= a_0 + a_1\pi + a_2\pi^2 + \cdots + a_n\pi^n.\end{aligned}$$

(Recall we are specifying that ϕ should send any rational to itself.) Thus the image of ϕ is the subring $\mathbb{Q}[\pi]$ of $\mathbb{R}$ we described last time, namely the subring of "polynomial expressions" in π. The kernel of ϕ consists of polynomials $p(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$ over $\mathbb{Q}$ such that

$$\phi(p(x)) = a_0 + a_1\pi + a_2\pi^2 + \cdots + a_n\pi^n = 0,$$

and in fact that only polynomial with this property is the zero polynomial! This is not at all obvious, and reflects the fact that π is what's called a *transcendental* real number, which literally means that it is not the root of a nonzero polynomial with rational coefficients. This is also the underlying reason why there is no way to simplify an expression such as $a_0+a_1\pi+a_2\pi^2+\cdots+a_n\pi^n$ as opposed to the case where we take, say, "polynomials" in $\sqrt{2}$, and is something we will come back to next quarter when we say more about transcendental elements.

If instead we considered the ring map $\mathbb{Q}[x] \to \mathbb{R}$ sending x to $\sqrt{2}$ (and any rational to iself), the image would be $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$, and the kernel would be the set of all polynomial multiples of $x^2 - 2$. Certainly, $x^2 - 2$ is a rational polynomial which has $\sqrt{2}$ as a root, and the fact is that *any* polynomial with $\sqrt{2}$ as a root must be a multiple of this one. This is something we will come back to proving later, when we think about $\mathbb{Q}[x]$ as what's called a *Euclidean domain*.

Towards quotient rings. We now seek to describe quotient constructions in ring theory, where, analogously to what we did for groups, we take a ring and imagine introducing some new "relations". If R is a ring and $I \subseteq R$ a subring, we start with the *additive* quotient group R/I, which is the set of additive cosets:

$$R/I := \{r + I \mid r \in I\}$$

under the operation of addition defined as

$$(r + I) + (r' + I) := (r + r') + I.$$

Note that we were use additive notation $r + I$ to describe a coset instead of the multiplicative notation gH use usually used last quarter, simply because the group operation we are using here

is addition. As before, often we will not use coset notation at all, and simply think about elements of R/I as elements R only with additional equalities imposed. In particular,

$$r = r' \text{ in } R/I \text{ (i.e. } r + I = r' + I) \text{ if and only if } r - r' \in I.$$

That is, r and r' determine the same quotient element when they differ by an element of I, since elements of I are the ones we are declaring to be "zero" in the quotient. ($r - r' \in I$ is simply the additive version of $g^{-1}k \in H$ we saw last quarter for determine when $g = k$ in G/H.) Note that R/I is indeed a true (additive) group since I is an additive normal subgroup of R simply because R is abelian with respect to addition.

Now we want to introduce a multiplication in order to get a ring structure. The obvious attempt—especially if we are considering elements of R/I as if they were simply elements of R—is to use the multiplication we already have on R:

$$(r + I)(r' + I) := rr' + I.$$

But the thing to watch out for—as we saw for groups—is the issue of whether or not this multiplication is well-defined. That is, if we have $r = r'$ in R/I, we want to make sure that $rs = r's$ in R/I for any $s \in R$, and that $sr = sr'$ in R/I as well. Otherwise, multiplication would be dependent on which "equal" elements we used when performing the operation, which is no good. In the case of groups, this is what led us to uncover the notion of a "normal subgroup", and so we seek the analogous thing for rings.

To say that $r = r'$ in R/I means that $r - r' \in I$. For $s \in R$, to say that $rs = r's$ in R/I would require that $rs - r's \in I$. This final element can be written as $(r - r')s$, so we need to know that:

$$r - r' \in I \text{ implies } s\underbrace{(r - r')}_{\in I} \in I \text{ for all } s \in R.$$

Similarly, $sr = sr'$ in R/I means $sr - sr' \in I$, so we need to know that:

$$r - r' \in I \text{ implies } \underbrace{(r - r')}_{\in I}s \in I \text{ for all } s \in R.$$

In other words, the property we need in order to ensure that multiplication on the quotient is well-defined is that multiplying any element of I by any element of R on either side always produces an element of I itself. As with groups (where we got "normal" subgroups), this property is important enough that we give it its own name: we say that I is an *ideal* of R.

Ideals. To be clear, a subring I of R is an *ideal* of R if $ra \in I$ and $ar \in I$ for all $a \in I$ and $r \in R$. This can be expressed more compactly as:

$$rI \subseteq I \quad \text{and} \quad Ir \subseteq I$$

for all $r \in R$, where by rI we mean the set of all products ra with $a \in I$, and Ir is the set of all products ar with $a \in I$. So, I is an ideal when multiplying its elements by those of R keeps you within I. When this holds, then R/I is a ring under the operations described above, which we call the *quotient ring* "R mod I".

Ideals will be important objects in our study of rings. It is often useful to separate the two requirements in the definition above, to say that I is a *left-ideal* of R when $rI \subseteq I$ holds for all $r \in R$, and that I is a *right-ideal* when $Ir \subseteq I$ holds for all r. These two properties are independent

of one another, since it is possible for a left-ideal to not be a right-ideal, and vice-versa. When *both* properties hold, we often say that I is a *two-sided* ideal, or simply an "ideal" as in the definition we gave about. (That is, the term "ideal" on its own will always refer to the two-sided case.) Quotient rings only make sense in general for two-sided ideals.

We will look at plenty of examples of ideals in the coming days, but here is a first general example. Suppose R is commutative and $a \in R$. The *principal ideal* generated by a is denoted (a) is the ideal consisting of all multiples of a by elements of r:

$$(a) := \{ra \mid r \in R\}.$$

It is straightforward to check that this is a subring of R, and the ideal condition (in the commutative case, left and right ideals are the same) follows from associativity $s(ra) = (sr)a$. (Defining the notion of a "principal ideal" in the non-commutative case is a little more involved, so we'll do that after we have some more comfort with ideals built-up.)

Example. Consider the quotient ring $\mathbb{Q}[x]/(x^2 - 2)$, where $(x^2 - 2)$ is the principal ideal of $\mathbb{Q}[x]$ generated by $x^2 - 2$. We claim that this is quotient is isomorphic to a more familiar ring we have worked with before. Indeed, consider a rational polynomial

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

as an element of this quotient. In the quotient, $x^2 - 2 = 0$ since all elements in the ideal $(x^2 - 2)$ we are quotienting by are declared to be zero. Thus, $x^2 = 2$ in the quotient, so we can simplify powers of x beyond x^1 in the expression for $p(x)$ above:

$$x^2 = 2,\ x^3 = (x^2)x = 2x,\ x^4 = (x^2)(x^2) = 4,\ x^5 = (x^4)x = 4x,\ \text{etc.}$$

Thus, in the end, the expression for $p(x)$ can be simplified to

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n = b + cx$$

for some $b, c \in \mathbb{Q}$. Hence, concretely, $\mathbb{Q}[x]/(x^2 - 2)$ consists of polynomials $b + cx$ with $b, c \in \mathbb{Q}$, under the multiplication given by:

$$(b + cx)(d + fx) = (bd + 2cf) + (bf + cd)x,$$

where we use the fact that $cfx^2 = 2cf$ when "expanding" the product on the left.

But the point is that this is precisely how multiplication in $\mathbb{Q}(\sqrt{2})$ works! Indeed, here we get:

$$(b + c\sqrt{2})(d + f\sqrt{2}) = (bd + 2cf) + (bf + cd)\sqrt{2},$$

so the map $\mathbb{Q}[x]/(x^2 - 2) \to \mathbb{Q}(\sqrt{2})$ defined by sending any rational to itself and $x \mapsto \sqrt{2}$ is an isomorphism. By taking the quotient by the ideal $x^2 - 2$ and hence setting $x^2 - 2 = 0$ to be a true equality in the quotient, we are essentially declaring x to be a "square root of 2", so it makes complete sense that the resulting ring should be the same as $\mathbb{Q}(\sqrt{2})$. Note that, thus, in this case the quotient $\mathbb{Q}[x]/(x^2 - 2)$ is actually a field, and in fact this is no accident: that the quotient is a field reflects a certain property of the ideal $(x^2 - 2)$—maximality—which we will discuss later.

Lecture 5: More on Quotients and Ideals

Warm-Up 1. We verify that the quotient ring $(\mathbb{Z}/3\mathbb{Z})[x]/(x^2+1)$ is a field by explicitly finding the multiplicative inverse of each nonzero element. In this ring the equality $x^2+1=0$ holds, so $x^2=-1$ allows us to simplify the description of elements. (In a sense, x plays the role of a "square root of -1"—or "i"—in this quotient). In other words, a polynomial

$$a_0 + a_1x + a_2x^2 + \cdots + a_nx^n$$

will be equal in the quotient to one of the form $a+bx$, since $x^2=-1$, $x^3=-x$, $x^4=1$, $x^5=x$, and so on. Thus, elements in the quotient can be uniquely characterized as being of the form

$$a+bx \text{ with } a,b \in \mathbb{Z}/3\mathbb{Z},$$

and no two such elements are the same. This shows that the quotient has $3^2=9$ elements, coming from the 3 possible choices for each of a and b. (We will show next quarter that in fact any *finite* field must have prime-power many elements.)

Now, since $x^2=-1$, we have $x(-x)=-x^2=1$ so that the inverse of x is $-x=2x$. (Recall that computations are done mod 3). The inverse of $x+1$ should satisfy

$$(x+1)(a+bx)=1,$$

and we can determine the values of a,b from this condition alone. The product on the left equals:

$$ax+bx^2+a+bx = ax+b(-1)+a+bx = (a-b)+(a+b)x$$

where we use $x^2=-1$ in the first step. In order for this to equal 1 requires that $a-b=1$ and $a+b=0$ in $\mathbb{Z}/3\mathbb{Z}$, and checking all possible values of $a,b\in\mathbb{Z}/3\mathbb{Z}$ shows that only $a=2, b=1$ works. Thus the inverse of $x+1$ is $x+2$. Similarly, the inverse of $2x+1$ is found by solving

$$(2x+1)(c+dx)=1,$$

which after expanding the left becomes

$$(c-2d)+(2c+d)x=1.$$

This gives $c-2d=1, 2c+d=0$, so $c=2, d=2$. Hence $(2x+1)^{-1}=2x+2$. Thus, $(\mathbb{Z}/3\mathbb{Z})[x]/(x^2+1)$ is commutative, has unity, and every nonzero element has an inverse:

$$1^{-1}=1,\ 2^{-1}=2,\ x^{-1}=2x,\ (x+1)^{-1}=x+2,\ (2x+1)^{-1}=2x+2,$$

so $(\mathbb{Z}/3\mathbb{Z})[x]/(x^2+1)$ is a field.

Warm-Up 2. We find some zero divisors in the quotient $(\mathbb{Z}/3\mathbb{Z})[x]/(x^2+x+1)$, thereby showing that this ring is not an integral domain. Note in this case that 1 is a root of x^2+x+1 over $\mathbb{Z}/3\mathbb{Z}$. Thus if we expect that polynomials over $\mathbb{Z}/3\mathbb{Z}$ should behave in a way similar to polynomials over $\mathbb{R}$ (we will clarify the extent to which this is true later—it is key that $\mathbb{Z}/3\mathbb{Z}$ is in fact a field), it should in theory be possible to factor x^2+x+1 in a way which has $x-1=x+2$ as a factor. If we do then have

$$(x+2)p(x)=x^2+x+1$$

for some polynomial $p(x)$, then $x+2$ and $p(x)$ will be zero divisors in the quotient, because the equality above becomes $(x+2)p(x)=0$ in the quotient.

In the quotient we have $x^2+x+1=0$, so $x^2=-x-1=2x+2$. This again allows us to rewrite an arbitrary polynomial in the quotient as one of the form $a+bx$ with $a,b\in\mathbb{Z}/3\mathbb{Z}$. (For instance, $x^3=xx^2=x(2x+2)=2x^2+2x=2(2x+2)+2x=6x+4=1$.) Thus we seek $p(x)=a+bx$ which satisfies

$$(x+2)(a+bx)=x^2+x+1=0 \text{ in } (\mathbb{Z}/3\mathbb{Z})[x]/(x^2+x+1).$$

The left side is:

$$ax+bx^2+2a+2bx=ax+b(2x+2)+2a+2bx=(2a+2b)+(a+4b)x,$$

so the conditions we need are $2a+2b=0$ and $a+4b=a+b=0$. Picking $a=1,b=2$ works, so we have

$$(x+2)(1+2x)=0$$

in the quotient, and hence $x+2$ and $1+2x$ are zero divisors. Note that multiplying $1+2x$ by 2 gives $(x+2)[2(1+2x)]=0$ as well, so $2(1+2x)=2+x$ is also a zero divisor.

Isomorphism theorems. A key property of normal subgroups last quarter was that they are precisely the things which arise as kernels of group homomorphisms. The analogous thing is true in the setting of rings: ideals are kernels and kernels are ideals. That is, if $\phi: R\to S$ is a ring homomorphism, then $\ker\phi=\{x\in \mid \phi(x)=0\}$ is an ideal of R, and if I is an ideal of R, I is the kernel of the "projection map" $R\to R/I$ which sends every element of R to the element it represents in the quotient (i.e. essentially itself). That $\ker\phi$ is an ideal of R is a straightforward check: in particular, the ideal condition comes from $\phi(rx)=\phi(r)\phi(x)$ (and the other way around), so that if $\phi(x)=0$ then $\phi(rx)=0$ as well.

As in the case of group homomorphisms, we can also see that two elements $r,r'\in R$ give the same output $\phi(r)=\phi(r')$ precisely when $r-r'\in\ker\phi$, so that cosets of the kernel correspond to elements of the image. This gives the ring-theoretic analog of the *First Isomorphism Theorem*:

$$R/\ker\phi\cong\phi(R).$$

Of course, since a ring is an abelian additive group, we already know that this isomorphism holds with respect to the *group* structure on both sides by the First Isomorphism Theorem from last quarter, so the real point now is that it still holds at the level of *rings* when we incorporate the multiplications as well.

Similarly, there are direct ring-theoretic analogs of the *Second, Third*, and *Fourth* Isomorphism Theorems we previously saw for groups. The statements and proofs are exactly the same (only that we use additive notation, so $A+B$ instead of AB for instance), so we will not repeat them in full here. The fourth of these, if you recall, gave the relation between subgroups of a quotient and subgroups of the original group, so the ring case this says things like: "subrings in R/I are in one-to-one correspondence with subrings of R which contain I", and so on.

Examples. In fact we already saw one instance of the First Isomorphism Theorem in action. Last time we argued that the quotient $\mathbb{Q}[x]/(x^2-2)$ should be isomorphic to $\mathbb{Q}(\sqrt{2})$, and now we can phrase this as follows: we have a ring homomorphism

$$\mathbb{Q}[x]\to\mathbb{R} \text{ defined by } x\mapsto\sqrt{2}, \text{rational}\mapsto\text{itself}$$

whose image is $\mathbb{Q}[\sqrt{2}]=\mathbb{Q}(\sqrt{2})$ and whose kernel is (x^2-2). (Again, we have not yet shown that anything in the kernel is a polynomial multiple of x^2-2, but we will do so later. The fact is that if F is a field, any ideal in $F[x]$ is *principal*.) Thus we have

$$\mathbb{Q}[x]/(x^2-2)\cong\mathbb{Q}(\sqrt{2})$$

by the First Isomorphism Theorem for rings.

Imagine we lived in a world where we knew nothing about $\mathbb{R}$ or $\sqrt{2}$, and all that we had to work with were rational numbers and rational numbers alone. Over $\mathbb{Q}$, the polynomial $x^2 - 2$ has no root, so the question is whether we can in some sense "enlarge" $\mathbb{Q}$ to a field over which $x^2 - 2$ *does* have a root, using only rational numbers alone? The answer we can now see is "yes": the quotient $\mathbb{Q}[x]/(x^2 - 2)$ is indeed a field which "contains" $\mathbb{Q}$ (as the subfield of constant polynomials) over which $x^2 - 2$ does have a root, namely "x" itself! The point is that by taking the quotient by $x^2 - 2$ and thus declaring $x^2 - 2 = 0$ to be true, we are literally forcing x to become "$\sqrt{2}$" in the quotient. This type of construction will be an important use of quotients of polynomial rings next quarter.

Similarly, considering the ring homomorphism $\mathbb{R}[x] \to \mathbb{C}$ which sends $x \mapsto i$ and any real number to itself leads to the isomorphism

$$\mathbb{R}[x]/(x^2 + 1) \cong \mathbb{C}$$

via the First Isomorphism Theorem. Thus, the left side gives a way to construct something like "$\mathbb{C}$" but using only real numbers alone. (Indeed, one approach to giving a formal definition of the set of complex numbers is to literally define it as the quotient $\mathbb{R}[x]/(x^2 + 1)$.) Again, by setting $x^2 + 1 = 0$ in the quotient we are declaring that x itself should be a "square root of -1".

Principal ideals. The coming days will be dedicated to studying ideals in more detail. A first observation is that if R has unity and $I \subseteq R$ is an ideal which contains 1, then $I = R$, since $r = r \cdot 1$ must be in I for all $r \in R$ by the ideal condition $rI \subseteq I$. More generally, if I contains any *unit*, it must be all of R: if $u \in U$ is a unit, then $u^{-1}u = 1$ is in I by the ideal condition, so $I = R$.

It will be useful to have a nice way of describing various examples of ideals. Perhaps the simplest example is that of an ideal "generated" by a single element. Suppose R has unity. We define the *principal* ideal generated by $a \in R$ to be the *smallest* ideal (a) of R which contains a in the sense that if $I \subseteq R$ is an ideal which contains a, then $(a) \subseteq I$. (This is analogous to how we first defined subgroups generated by elements in a group last quarter.) But, we can be much more explicit about what elements in this "smallest" ideal (a) actually look like. First, (a) should contain a, and so must contain all things of the form ra for $r \in R$ if it is to satisfy the (left) ideal condition. But then, if $ra \in (a)$, we must have $rar' \in (a)$ for all $r' \in R$ by the (right) ideal condition. Ideals should be subrings, meaning they should in particular be closed under addition, so (a) must also contain sums of elements of the form rar'. Thus, we get that explicitly (a) looks like:

$$(a) = \{r_1 a r_1' + \cdots + r_n a r_n' \mid r_i, r_i' \in R\}.$$

You can verify directly that this is indeed an ideal of R, which we will do explicitly next time in the case where we replace the single element a by a more general subset $A \subseteq R$. Principal ideals, in some sense, are ring-theoretic analogs of cyclic subgroups.

A few remarks are in order. First, note that if we want all elements of (a) to look like $\sum r_i a r_i'$, in particular a itself should look like this, which means that we should allow for the possibility that $r_i = r_i' = 1$—this is why we assumed that R has unity when defining the notion of a principal ideal, since otherwise we might not be able to express a in the required form and elements of (a) would be uglier to describe in general. Second, when R is commutative, $r_i a r_i' = (r_i r_i')a$, so we can greatly simplify the description of elements of (a):

$$r_1 a r_1' + \cdots + r_n a r_n' = r_1 r_1' a + \cdots + r_n r_n' a = (r_1 r_1' + \cdots + r_n r_n')a,$$

so that $(a) = \{ra \mid r \in R\}$. The point is that in the commutative case, left and right ideals are the same, so we need only care about multiplying a on the *left* by elements of R.

Principal ideals are the simplest types of ideals to work with. We will see that in many familiar examples, such as $\mathbb{Z}$ or the ring of polynomials over a *field*, all ideals are in fact principal. (Rings with this property will be called *principal ideal domains*, or "PIDs" for short.)

Lecture 6: Maximal Ideals

Warm-Up 1. Suppose R is a ring with unity. Given $A \subseteq R$, we describe explicitly the ideal generated by A, which is denoted by (A). (Principals ideals are the case where A has a single element.) By definition, this is the *smallest* ideal of R which contains all of A, and in fact, the work we did last time for principal ideals suggests what the answer should be:

$$(A) = \{r_1 a_1 r_1' + \cdots + r_n a_n r_n' \mid r_i, r_i' \in R, a_i \in A\}.$$

The only difference between this and (a) is that here the "middle" terms a_i coming from A do not have all be the same. The point is that if $a_i \in A$, then (A) should contain all things of the form $r_i a_i$ by the left ideal condition, then all things of the form $r_i a_i r_i'$ by the right ideal condition, and then all sums of such things by the subring condition. We assume R has unity in order to ensure that (A) as described explicitly here does contain A, by taking the r_i, r_i' to be 1.

Let us verify that (A) as above is indeed an ideal of R. (The fact that it is the smallest ideal of R containing A, after we know it is an ideal, comes from noting that if I is an ideal of R which contains A, I will have to contain all of the expressions above by the ideal and subring criteria, so that I will have to contain this (A).) First, (A) is closed under addition, since

$$(r_1 a_1 r_1' + \cdots + r_n a_n r_n') + (s_1 b_1 s_1' + \cdots + s_m b_m s_m')$$

where $r_i, r_i', s_i, s_i' \in R$ and $a_i, b_i \in A$, is still of the required form since it too is a sum of elements of the form (ring element)(element of A)(ring element). The fact that (A) is closed under multiplication is a consequence of the ideal condition alone (take $r \in I, a \in I$ in the ideal condition), so we need only verify the ideal condition: if $r \in R$, then

$$r(r_1 a_1 r_1' + \cdots + r_n a_n r_n') = (rr_1) a_1 r_1' + \cdots + (rr_n) a_n r_n'$$

and

$$(r_1 a_1 r_1' + \cdots + r_n a_n r_n') r = r_1 a_1 (r_1' r) + \cdots + r_n a_n (r_n' r),$$

which are each of the required form. Hence (A) as described above is closed under addition and closed under multiplication by elements of R, so it is an ideal. (The fact that (A) contains additive inverses is a consequence of the ideal condition by taking $r = -1$: $(-1)x = -x \in (A)$ if $x \in (A)$.)

When R is commutative, by using $r_i a_i r_i' = (r_i r_i') a_i$, the description can be simplified a bit to:

$$(A) = \{r_1 a_1 + \cdots + r_n a_n \mid r_i \in R, a_i \in A\}.$$

You can think of $r_1 a_1 + \cdots + r_n a_n$ as analogous to "linear combinations" in linear algebra, and so (A) as analogous to a "span". We will explore this more carefully later on.

Warm-Up 2. We show that the ideal $(4, x+3) \subseteq \mathbb{Z}[x]$ generated by 4 and $x+3$ is not principal. (So, $(4, x+3)$ is alternate notation for (A) when $A = \{4, x+3\}$.) Note that if view this as an ideal of $\mathbb{Q}[x]$ instead, $(4, x+3)$ *is* principal, and is indeed equal to all of $\mathbb{Q}[x] = (1)$: in this case, $4 \in (4, x+3)$ is a unit, so an observation from last time immediately gives $(4, x+3) = \mathbb{Q}[x]$. In the case of $\mathbb{Z}[x]$, 4 is not a unit, so this argument does not apply.

Suppose $(4, x+3) \subseteq \mathbb{Z}[x]$ could be generated by a single element $p(x) \in \mathbb{Z}[x]$. (Aside: explicitly, $(4, x+3)$ consists of all polynomials of the form $4q(x) + (x+3)r(x)$ with $q(x), r(x) \in \mathbb{Z}[x]$.) Then in particular $4 \in (p(x))$ and $x+3 \in (p(x))$. The idea is to use these conditions to determine what $p(x)$ would have to look like, and argue that no such $p(x)$ could exist. If $4 \in (p(x))$, then

$$4 = p(x)\ell(x) \text{ for some } \ell(x) \in \mathbb{Z}[x].$$

This forces $p(x), \ell(x)$ to be constant polynomials, since otherwise their product would necessarily have degree at least 1. (This depends on the fact that $\mathbb{Z}$ is an integral domain.) But then the only possibilities for $p(x)$ are $\pm 1, \pm 2, \pm 4$. We will show in a second that $p(x)$ cannot be ± 1, so that $p(x) = \pm 2$ or $p(x) \pm 4$. But in any of these cases, the polynomial $p(x)\ell(x) \in (p(x))$ has coefficients which are all even:

$$\pm 2 \sum a_i x^i = \sum \pm 2a_i x^k \text{ or } \pm 4 \sum a_i x^k = \sum \pm 4a_i x^k,$$

so such a product can never equal $x+3$, so we would not have $x+3 \in (p(x))$, meaning that $p(x)$ alone could not in fact generate all of $(4, x+3)$.

To rule out $p(x) = \pm 1$ (we did not do this explicitly in class since it is a little involved), suppose this was true and consider the equality

$$4q(x) + (x+3)r(x) = \pm 1,$$

which would hold for some $q(x), r(x)$. (Recall we are assuming $(4, x+3) = (p(x))$. Note $r(x)$ must have degree one less than that of $q(x)$, since the degree of $(x+3)r(x)$ is $\deg r(x) + 1$ and the highest degree term here has to cancel out with the highest degree term of $q(x)$ if we want the expression above to equal a constant. Write the polynomials $q(x), r(x)$ as

$$q(x) = a_0 + a_1 x + \cdots + a_n x^n \quad \text{and} \quad r(x) = b_0 + b_1 x + \cdots + b_{n-1} x^{n-1}.$$

The coefficient of x^n in $4q(x) + (x+3)r(x)$ is then

$$4a_n + b_{n-1}.$$

This coefficient should equal 0 if we want $4q(x) + (x+3)r(x) = \pm 1$ to hold, so that $b_{n-1} = -4a_n$ must in particular be a multiple of 4. The coefficient of x^{n-1} in $4q(x) + (x+3)r(x)$ is

$$4a_{n-1} + b_{n-2} + 3b_{n-1},$$

which should again be zero, implying that $b_{n-2} = -4a_{n-1} - 3b_{n-1}$ is a multiple of 4. And so on, the coefficient of x^k in $4q(x) + (x+3)r(x)$ is

$$4a_k + b_{k-1} + 3b_k,$$

which must be zero for $k > 0$, which recursively gives that b_{k-1} is a multiple of 4 for $k > 0$; in particular, b_0 is a multiple of 4. But then the constant terms in $4q(x) + (x+3)r(x) = \pm 1$ give

$$4a_0 + 3b_0 = \pm 1,$$

which would imply that ± 1 is also divisible by 4 since b_0 is, which is not true. Thus $p(x) \neq \pm 1$ as claimed, so $(4, x+3)$ is not principal. (Whew!)

Arithmetic of ideals. Back on the first day when providing some motivation for ring theory, we mentioned the idea that, historically, ideals arose from the desire to have a notion of an "ideal

number" as a way to get around the lack of "unique factorization" in certain contexts. The modern definition of an ideal is a bit removed from the historical definition of an "ideal number", but the underlying concepts and goals are the same.

We will not go into what an "ideal number" was originally defined as historically, but *will* give a sense in which ideals can be treated as "numbers" in a way, mainly because we can make sense of "adding" and "multiplying" them. Given two ideals $I, J \subseteq R$, we define their *sum* $I + J$ to be the ideal consisting of all sums of an element of I and an element of J:

$$I + J := \{a + b \mid a \in I, b \in J\}.$$

You can check that this is indeed an ideal of R, and is in fact the smallest ideal which contains both I and J; in other words, this is precisely the ideal generated by $I \cup J$. The *product* ideal IJ takes a bit more care to define. The quick definition is that this should be the ideal generated by all products ab with $a \in I, b \in J$. The issue is that this ideal contains more than such products themselves, since the set of such products is not closed under addition:

$$ab + a'b' \text{ is not necessarily of the form (element of } I\text{)(element of } J\text{)}.$$

To get a true ideal then we have to take sums of such products, and we get:

$$IJ := \{a_1b_1 + \cdots + a_nb_n \mid a_i \in I, b_i \in J\}.$$

Again, it is straightforward to check that this is an ideal of R; in particular, for the (left) ideal condition, note that

$$r(a_1b_1 + \cdots + a_nb_n) = (ra_1)b_1 + \cdots + (ra_n)b_n$$

is of the required form since each ra_i is in I, because I is itself an ideal. Note also that since I and J are both ideals, a product like ab is in both I and J, so that IJ is in fact a subset of $I \cap J$.

We will only briefly discuss the idea of doing "arithmetic" with ideals in this course, mainly in the context of "unique factorization" later on. This is a much important notion in areas such as number theory and algebraic geometry, which we will only give a flavor of.

Revisiting fields. Coming back to quotients, we previously saw examples where a quotient R/I was actually a field, such as in $\mathbb{Q}[x]/(x^2-2)$ or $(\mathbb{Z}/3\mathbb{Z})[x]/(x^2+1)$. We now seek to understand the condition on the ideal I which accounts for this. As a first observation, note that the only ideals of a field F are the zero ideal $\{0\}$ and the ideal $(1) = F$. Indeed, if $I \subseteq F$ is any nonzero ideal, then any nonzero element $u \in I$ is a unit, and we have seen that containing a unit implies that $I = F$. A ring whose only ideals are the zero ideal and the entire ring is called a *simple* ring, mimicking the definition of "simple group" from last quarter. So, fields are examples of simple rings. (Simple rings will not play a big role for this in this course, certainly not to the extent to which simple groups played a role last quarter and will do so again next quarter.)

In fact, in the commutative case, fields are the *only* examples of simple rings with unity. That is, suppose R is a commutative ring with unity which is simple. Then for any nonzero $x \in R$, the principal ideal (x) is either $\{0\}$ or R, so it must be R because $x \neq 0$. But if $(x) = R$, there exists $r \in R$ such that $rx = 1$, so x is a unit and R is thus a field. The upshot is that fields can be characterized precisely as the commutative simple rings with unity.

Characterizing non-commutative simple rings is not as straightforward, and in particular there are more of these than just division rings, which are the non-commutative analogs of fields. For instance, for $n > 2$, the ring of matrices $M_n(F)$ over a field F is simple (one can show that the ideals of a matrix ring $M_n(R)$ in general are of the form $M_n(I)$ where I is an ideal of R) but not a division

ring. The issue with trying to carry out the proof that "commutative and simple implies field" in the non-commutative case comes from the fact that principal ideals in the non-commutative case have more elements than simply multiples of the generator: if R is simple with unity and $x \in R$ is nonzero, it is still true that (x) must be all of R, so that $1 \in (x)$, but this means that

$$r_1 x r_1' + \cdots + r_n x r_n' = 1$$

for some $r_i, r_i' \in R$, which says nothing about whether or not x has a multiplicative inverse. (In the commutative case, the expression above reduces to $rx = 1$ for some r.)

Maximal ideals. So, suppose R is commutative with unity and that I is a proper (meaning $I \neq R$) ideal. (The proper condition is needed to guarantee that $1 \neq 0$ in the quotient, as we require to be true in a field.) As explained above, R/I is a field (it is already commutative since R is) if and only if its only ideals are $I/I = \{0\}$ and R/I. Any ideal of R/I is of the form J/I for some ideal $J \subseteq R$ containing I according to the Fourth Isomorphism Theorem for rings. (The J with this property is the pre-image of the quotient ideal in question under the projection map $R \to R/I$.) Such a J fits into a chain $I \subseteq J \subseteq R$. Thus, the condition that any ideal of R/I is either the zero ideal or everything says that J/I above would have to be either $J/I = I/I$, in which case $I = J$, or $J/I = R/I$, in which case $J = R$. Hence, we see that R/I is a field if and only if whenever we have an ideal J sitting between I and R as in $I \subseteq J \subseteq R$, we must have $I = J$ or $J = R$.

Proper ideals with this property are important enough that we give them a name: *maximal ideals*. (To be clear, the condition that "$I \subseteq J \subseteq R$ implies $I = J$ or $J = R$" is the one which defines maximality.) Thus, a maximal ideal is one which is not the entire ring and not contained in any larger ideal apart from the entire ring. The conclusion we have from the work above is that if R is commutative with unity, then R/I is a field if and only if I is a maximal ideal of R. The idea behind the proof we gave above of this result is perhaps obscured by the fact that we phrased it in terms of the behavior of the ideals of the quotient and the characterization of a field as a simple commutative ring, and indeed it is not so clear at first sight how the maximality of the ideal says anything about the existence of inverses in the quotient. We will give an alternate proof next time which avoids the use of ideals of the quotient and makes clearer where the inverses come from. (The proof-to-come and the proof above are really the same in the end, just that one is phrased using ideals and the other using elements, but the proof using elements gives perhaps better intuition.)

Examples. We finish by looking at some examples of maximal ideals. First, the maximal ideals of $\mathbb{Z}$ are precisely those which are generated by single primes: $(p) = p\mathbb{Z}$ is maximal if and only if p is prime. (Recall, as we will prove later, that every ideal of $\mathbb{Z}$ is principal.) Indeed, we have $(a) \subseteq (b)$ for $a, b \in \mathbb{Z}$ if and only if a is divisible by b (since $a \in (b)$ means $a = kb$ for some $k \in \mathbb{Z}$), so the maximality condition "$(a) = (b)$ or $(b) = \mathbb{Z}$" says that either $b = \pm a$ or $b = \pm 1$ (i.e. the only divisors of a are $\pm a$ and ± 1), which holds if and only if a is prime. For instance, (4) is not maximal since $(4) \subseteq (2)$ and $(2) \neq \mathbb{Z}$.

An example from before showed that $\mathbb{Q}[x]/(x^2-2)$ is a field, so in fact (x^2-2) must be maximal in $\mathbb{Q}[x]$. A previous Warm-Up showed that $(\mathbb{Z}/3\mathbb{Z})[x]/(x^2+1)$ is a field, so (x^2+1) is maximal in $(\mathbb{Z}/3\mathbb{Z})[x]$. What ties these examples together is the polynomials generating these maximal ideals are in fact *irreducible*, which essentially means that they cannot be factored. We will come back to this later. Finally, we can ask whether the ideal $(4, x+3)$ from the second Warm-Up is maximal, which we will decide at the start of next time.

Lecture 7: Prime Ideals

Warm-Up 1. We give an alternate proof of the fact that if R is a commutative ring with identity, then an ideal $I \subseteq R$ is maximal if and only if R/I is a field, which avoids the characterization of a field as a commutative simple ring with unity. This proof focuses directly on elements, and shows more concretely where an inverse for a nonzero element in the quotient comes from. (As stated last time, if you "unravel" the previous proof we gave to focus more on elements, you get this proof.) A key observation is that to say a nonzero r in the quotient R/I has an inverse concretely means that there exists $x \in R$ such that $rx - 1 \in I$ (so $rx - I = m$ for some $m \in I$), since this is the condition which gives $rx = 1$ in R/I.

Suppose I is a maximal ideal of R, and let $r \in R - I$. (The elements of the complement $R - I$ are the ones which give nonzero elements in R/I.) Since $r \notin I$, the ideal (I, r) generated by I and r is strictly larger than I, so since I is maximal we must have $(I, r) = R$. Thus, we can express $1 \in R$ as

$$m + rx = 1$$

for some $m \in I$ and $x \in R$. (Keep the existence of such an equality in mind going forward—it is a common type of thing which shows up in arguments involving maximality!) This becomes $rx = 1$ in R/I, so r has an inverse in R/I, and hence R/I is a field.

Conversely, suppose R/I is a field and suppose J is an ideal of R such that $I \subseteq J \subseteq R$. If $I \neq J$, there exists $r \in J - I$. Then r is nonzero in R/I, so there exists $x \in R$ such that $rx = 1$ in R/I. This is equivalent to $rx - 1 = m$ for some $m \in I$, or $m + rx = 1$. Since $m + rx \in J$, we have $1 \in J$ so $J = R$, and hence I is maximal.

Warm-Up 2. We show that the ideal $(4, x + 3) \subseteq \mathbb{Z}[x]$ is not maximal, both directly and also by showing that the corresponding quotient is not a field. The latter is perhaps simpler: in $\mathbb{Z}[x]/(4, x + 3)$, we impose the relations

$$4 = 0 \quad \text{and} \quad x + 3 = 0$$

on $\mathbb{Z}[x]$, where the second is equivalent to $x = -3 = 1$ since $4 = 0$. Thus any polynomial in $\mathbb{Z}[x]$ is equivalent to a constant one in the quotient (replace every instance of x by 1), and constant polynomials are subject to $4 = 0$. Thus the quotient $\mathbb{Z}[x]/(4, x + 3)$ is isomorphic to $\mathbb{Z}/4\mathbb{Z}$, so since this is not a field we have that $(4, x + 3)$ is not maximal.

To show that $(4, x + 3)$ is not maximal directly, simply note that

$$(4, x + 3) \subseteq (2, x + 3)$$

since any polynomial multiple of 4 is also a multiple of 2. This containment is a proper one since $2 \in (2, x + 3)$ is not in $(4, x + 3)$, so if we know that $(2, x + 3) \neq \mathbb{Z}[x]$ then we will know that $(4, x + 3)$ is not maximal. To know that $(2, x + 3) \neq \mathbb{Z}[x]$, we can again consider a quotient: in $\mathbb{Z}[x]/(2, x + 3)$, $2 = 0$ and $x = 1$ again, so we find that $\mathbb{Z}[x]/(2, x + 3) \cong \mathbb{Z}/2\mathbb{Z}$. Since this is non-trivial, $(2, x + 3) \neq \mathbb{Z}[x]$. (Here the quotient is a field, so $(2, x + 3)$ is in fact a maximal ideal.) Showing directly that $(2, x + 3) \neq \mathbb{Z}[x]$ requires knowing that $1 \in \mathbb{Z}[x]$ cannot be expressed as

$$2p(x) + (x + 3)q(x) = 1$$

with $p(x), q(x) \in \mathbb{Z}[x]$, which can be done through a brute-force computations along the lines of the one we gave when showing that $(4, x + 3)$ is not a principal ideal of $\mathbb{Z}[x]$. This is certainly doable, but the quotient method is quicker. (That $1 \neq 0$ in the quotient essentially comes down to the fact that there is no way to derive $1 = 0$ solely from $2 = 0$ and $x = 1$ alone.)

Another example. Here is one more example of a maximal ideal. Take F to be a field and consider the ring $F[[x]]$ of formal power series over F. The ideal generated by (x) consists of all formal power series without a constant term:

$$x(a_0 + a_1x + a_2x^2 + \cdots) = a_0x + a_1x^2 + a_2x^3 + \cdots .$$

Quotienting out by this ideal sets $x = 0$ and thus reduces any formal power series to a constant one, so $F[[x]]/(x) \cong F$. Since F is a field, the ideal (x) is maximal.

In fact, it turns out that (x) is the *only* maximal ideal of $F[[x]]$. (Note that something like $(x-1)$ is *not* a maximal ideal, since, according to a problem on the first homework, $x-1$ is actually a unit, and so $(x-1)$ is all of $F[[x]]$. The same is true of $(x-a)$ for any nonzero $a \in F$.) More generally, you will show on a homework problem that any *discrete valuation ring*—of which $F[[x]]$ is an example—has a unique maximal ideal. Rings with this property show up often in number theory and algebraic geometry, which we will give a glimpse of later.

Existence of maximal ideals. One more observation is the fact that any ring R with unity not only has a maximal ideal, but in fact there is a maximal ideal which contains any given proper ideal: if $I \subseteq R$ is a proper ideal, there exists a maximal ideal $M \subseteq R$ such that $I \subseteq M$. Thus, any proper ideal in a ring with unity can be "enlarged" to one which is maximal.

We mention this now to highlight that this is an application of Zorn's Lemma, which we briefly introduced last quarter. As before, we can show existence without having any idea as to how to actually "construct" the said object. Here the argument runs as follows. Consider the collection $\mathcal{P}$ of all proper ideals of R which contain I:

$$\mathcal{P} := \{J \mid J \text{ is a proper ideal of } R \text{ and } I \subseteq J\}.$$

This is nonempty since $I \in \mathcal{P}$. Let $\mathcal{C}$ be a chain in $\mathcal{P}$, which if you recall means a subset such that for any $A, B \in \mathcal{C}$, either $A \subseteq B$ or $B \subseteq A$. The union $\bigcup \mathcal{C}$ of all elements of $\mathcal{C}$ is then an ideal of R: checking all the requirements is straightforward and uses the chain condition $A \subseteq B$ or $B \subseteq A$ to do all necessary sum and products computations inside of something (A or B) we already know to be an ideal. Moreover, $\bigcup C$ is a *proper* ideal of R: if not, $1 \in R$ would be in $\bigcup C$, so that $1 \in A$ for some $A \in \mathcal{C}$, which would imply that $A = R$ and hence that A would not have been proper. Thus, $\bigcup C$ is an upper bound of C in $\mathcal{P}$, so by Zorn's Lemma $\mathcal{P}$ has a maximal element, which is precisely a maximal ideal containing I. (Without the assumption that R has unity, what goes wrong is the argument that $\bigcup C$ is a proper ideal.)

Prime ideals. We motivated the consideration of maximal ideals by asking for the scenario under which a quotient R/I would be a field. Now we ask, if we do not get a field, when do we get an integral domain, which is perhaps the next best thing? That is, if R is commutative with identity, what condition on a proper ideal $I \subseteq R$ guarantees that R/I is an integral domain?

Here the answer is perhaps simpler to come by as opposed to the field/maximal case. To say that R/I is an integral domain means that

$$\text{if } ab = 0 \text{ in } R/I, \text{ then } a = 0 \text{ or } b = 0 \text{ in } R/I.$$

But to be zero in the quotient means to belong to the ideal, so this becomes:

$$\text{if } ab \in I, \text{ then } a \in I \text{ or } b \in I.$$

A proper ideal of R with this property is called a *prime* ideal. Thus, the answer to our question is that R/I is an integral domain if and only if I is a prime ideal. Note that the definition of an

integral domain itself, in terms of a lack of zero divisors, can now be rephrased precisely as the statement that the zero ideal (0) is a prime ideal, since $ab = 0$ just means $ab \in (0)$, so the integral domain condition becomes the prime condition $ab \in (0) \implies a \in (0)$ or $b \in (0)$.

Example. The main example of a prime ideal, and indeed the example which explains the use of the term "prime" in this context, comes from $\mathbb{Z}$, where a nonzero ideal $(p) = p\mathbb{Z}$ is a prime ideal if and only if p is a prime number. Indeed, the point is that prime numbers numbers do have the property that whenever p divides ab (which is what $ab \in (p)$ means), p divides a or p divides b. This is not the usual definition of "prime number", but is in fact equivalent to it, as we will see later in a general setting. (In general, a non-unit element of a commutative ring having this divisibility property is called a *prime element*, whereas an element having the property that it cannot be "factored"—i.e. more like the usual definition of prime—is called *irreducible*. We will clarify these definitions later, and consider the question as to when they are equivalent, as they are in $\mathbb{Z}$. Note that in $\mathbb{Z}$, the nonzero prime ideals are precisely the maximal ideals.)

The fact that a (usual) prime number has the property that whenever it divides a product it must divide at least one of the factors follows from the fact that $gcd(a, b)$ can be characterized as the smallest positive integer which can be expressed as $ax + by$ for $x, y \in \mathbb{Z}$. We will not give the explicit proof here, since it is essentially the same as the proof we will give in a bit that maximal ideals are always prime.

Another example. As another example, consider the ideal $(x + 3) \subseteq \mathbb{Z}[x]$. This is not maximal since, for instance, $(x + 3) \subseteq (2, x + 3)$. In the quotient $\mathbb{Z}[x]/(x + 3)$ we set $x = -3$, so that any polynomial in the quotient is equal to a constant one, which implies that

$$\mathbb{Z}[x]/(x + 3) \cong \mathbb{Z}.$$

Since this is an integral domain, $(x+3)$ is a prime ideal. (We could also be more precise in describing this quotient by using the First Isomorphism Theorem: the ring homomorphism $\mathbb{Z}[x] \to \mathbb{Z}$ defined by sending any integer to itself and x to -3 is surjective and has kernel equal to $(x + 3)$, so the isomorphism above follows.)

To show directly (without quotients) that $(x+3)$ is a prime ideal requires knowing that whenever $p(x)q(x)$ is a polynomial multiple of $x + 3$, then $p(x)$ or $q(x)$ is a multiple of $x + 3$; that is,

$$p(x)q(x) = (x + 3)r(x) \text{ for some } r(x) \Rightarrow p(x) = (x + 3)s(x) \text{ or } q(x) = (x + 3)s(x) \text{ for some } s(x).$$

Proving this is an efficient way requires some facts about polynomials (division algorithm and Gauss's Lemma) we will discuss later, so we skip it for now. Of course, the argument via quotients above works just fine. (Well, actually, the argument we gave above is in fact incomplete, since we have not shown that the kernel of the map $\mathbb{Z}[x] \to \mathbb{Z}$ defined above is indeed the ideal generated by $x + 3$; proving this amounts to the same amount of work as proving that $(x + 3)$ is prime is directly, dependent again on some polynomial facts we have yet to discuss.)

Maximal implies prime. Finally, we come to the fact that maximal ideals are always prime, in commutative rings R with unity at least. This is immediate from a consideration of quotients: if $I \subseteq R$ is maximal, then R/I is a field, which is thus an integral domain, so I is prime. But, as in the first Warm-Up, it is useful to give a non-quotient proof of this, in order to illustrate some general techniques. In particular, the proof relies on the ability to express $1 \in R$ in a certain way, did the proof in the Warm-Up.

Suppose $I \subseteq R$ is maximal, and that $ab \in I$ for some $a, b \in R$. Assuming $a \notin I$, we must then show that $b \in I$ if we want to conclude that I is a prime ideal. If $a \notin I$, then the ideal (I, a) strictly

contains I, and so must equal $(I, a) = R$ since I is maximal. Thus there exist $x \in I$ and $r \in R$ such that

$$x + ra = 1.$$

Multiplying through by b gives $xb + rab = b$. But now, xb and rab are both in I by the ideal condition, since $x, ab \in I$. Hence $b = xb + rab \in I$ as well, as was to be shown.

As mentioned earlier, this is essentially the proof that prime numbers have the divisibility property mentioned before: p divides ab implies p divides a or p divides b. Indeed, if $p \nmid a$, then $gcd(p, a) = 1$, so we can write

$$px + ay = 1$$

for some $x, y \in \mathbb{Z}$. Then $pxb + ayb = b$, so since p divides the left side (using $p \mid ab$), p divides the right side b as well.

Lecture 8: Geometry via Algebra

Warm-Up 1. Suppose R is a commutative ring with unity with the property that every proper ideal is prime. We show that R is in fact a field. As a first observation, (0) is prime, so R is at least an integral domain. Let $x \in R$ be nonzero, so that our goal is to produce an inverse for x. Consider the principal ideal (x^2) generated by x^2. If this ideal is not proper, so $(x^2) = R$, then there exists $y \in R$ such that $yx^2 = 1$, in which case yx is an inverse or x.

Otherwise, (x^2) is proper, in which case our assumptions imply that it is prime. Since $xx \in (x^2)$, the prime condition thus gives $x \in (x^2)$. Hence there exists $y \in R$ such that $x = yx^2$. This gives $x - yx^2 = 0$, so $x(1 - yx) = 0$. Since R is an integral domain and $x \neq 0$, this implies that $1 - yx = 0$, so $yx = 1$ and y is the multiplicative inverse of x. Hence R is a field.

Warm-Up 2. Suppose F is a field and consider the two-variable polynomial ring $F[x, y]$. We claim that the principal ideal (x) generated by x is prime in $F[x, y]$. As a first approach, the quotient

$$F[x, y]/(x) \cong F[y]$$

can be seen to be isomorphic to the single-variable polynomial ring $F[y]$, since setting $x = 0$ in a two-variable polynomial leaves a polynomial in y alone. (More rigorously, apply the First Isomorphism Theorem to the surjective map $F[x, y] \to F[y]$ defined by sending x to 0 and y and elements of F to themselves; the kernel consists of those polynomials which can be written as $xf(x, y) \in (x)$. We will essentially prove this in our second direct approach below.) Since F is a field, $F[y]$ is an integral domain (a previous Warm-Up we did back in the first week), so $F[x, y]/(x)$ is an integral domain, and hence $(x) \subseteq F[x, y]$ is prime as claimed. Note that this quotient is not a field, so (x) is not maximal, which we can see directly from the fact that $(x) \subseteq (x, y) \subsetneq F[x, y]$. (The ideal (x, y) itself is maximal, as you can check.)

For a second more direct approach, suppose $f(x, y)g(x, y) \in (x)$. (This more direct approach is essentially the same as that above, only without using quotients.) We claim that $f(x, y) \in (x)$ or $g(x, y) \in (x)$. By thinking of the ring $F[x, y]$ as $(F[y])[x]$ instead, we can express both $f(x, y)$ and $g(x, y)$ in the following way:

$$f(x, y) = a_0(y) + a_1(y)x + a_2(y)x^2 + \cdots + a_n(y)x^n$$
$$g(x, y) = b_0(y) + b_1(y)x + b_2(y)x^2 + \cdots + b_m(y)x^m$$

where $a_i(y), b_k(y) \in F[y]$. Then

$$f(x, y)g(x, y) = a_0(y)b_0(y) + (\text{terms involving } x).$$

In order to have $f(x,y)g(x,y) \in (x)$, we must have $a_0(y)b_0(y) = 0$, since a multiple of x cannot have any "constant" terms with respect to x, meaning terms involving only y:

$$x(c_0(y) + c_1(y)x + \cdots) = c_0(y)x + c_1(y)x^2 + \cdots.$$

Since $F[y]$ is an integral domain, $a_0(y)b_0(y) = 0$ implies that $a_0(y) = 0$ or $b_0(y) = 0$, which means that

$$f(x,y) = x(a_1(y) + a_2(y)x + \cdots + a_n(y)x^{n-1}) \text{ or } g(x,y) = x(b_1(y) + b_2(y)x + \cdots + b_m(y)x^{m-1}),$$

so $f(x,y) \in (x)$ or $g(x,y) \in (x)$ as required. Hence $(x) \subseteq F[x,y]$ is prime.

Rings of continuous functions. Now we will go off on a bit of a tangent in order to briefly introduce the idea of studying geometric objects solely via algebraic means. This is a very important concept in modern mathematics, and allows one to greatly broaden the scope to which "geometry" applies. Now, this is not something which will play a crucial role in our course, and we introduce this now mainly to give an example of how rings and ideals show up in other areas. We might come back to this point of view one or two more times just to provide some more context, but this is not something you should expect to see on an exam for instance. Hopefully you find it to be interesting (and perhaps crazy!) nonetheless.

We start by considering the ring $C[0,1]$ of continuous real-valued functions on the interval $[0,1]$:

$$C[0,1] := \{f : [0,1] \to \mathbb{R} \mid f \text{ is continuous}\}.$$

(In what follows, we will state various claims without justification, since proving them requires knowing more about continuous functions. Not everyone will have had a course in real analysis yet, so we cannot provide these justifications here, but this will not be a roadblock towards understanding the basic idea we are trying to get across. I am happy to talk about this in more detail in office hours if you'd like to see more of the actual analysis.) This is a ring (in fact commutative with unity) under the usual addition and multiplication of functions, and contains elements such as $1, e^x, \sin x$, etc. Now, for $p \in [0,1]$, we define m_p to be the subset consisting of continuous functions which vanish at p:

$$m_p := \{f \in C[0,1] \mid f(p) = 0\}.$$

It is straightforward to show that this is a subring of $C[0,1]$, and in fact an ideal since $f(p) = 0$ implies $g(p)f(p) = 0$ for any $g \in C[0,1]$. Moreover, we claim that m_p is a *maximal* ideal, which is simplest to prove by considering quotients via the First Isomorphism Theorem: the "evaluation at p" map $\phi : C[0,1] \to \mathbb{R}$ defined by $\phi(f) = f(p)$ is a ring homomorphism, which is surjective since the constant functions can be used to obtain any element of $\mathbb{R}$ in the image, and which has $m_p = \ker \phi$ as its kernel by definition of m_p; we thus get

$$C[0,1]/m_p \cong \mathbb{R},$$

which is a field, so m_p is maximal. (Showing that m_p is maximal directly without quotients amounts to showing that if $g \notin m_p$, the constant function 1 can be written as $1 = f(x) + h(x)g(x)$ for some $f(x) \in m_p$ and $g(x) \in C[0,1]$, which takes some work but is not too difficult. Draw a picture!)

And now the crucial point: in fact, *any* maximal ideal of $C[0,1]$ is of the form m_p for some $p \in [0,1]$. (Proving this is one thing that requires knowing more about continuity, and properties of the interval $[0,1]$. The key property needed is that $[0,1]$ is *compact*, and in fact an analogous statement is true if we consider maximal ideals of the ring of continuous functions $C(X)$ on an

arbitrary compact space X instead.) Thus, the maximal ideals of $C[0,1]$ are in bijection with the elements of $[0,1]$ (via the map which associates to a maximal ideal the point at which all functions inside of it vanish), so that we can essentially *recover* the interval $[0,1]$ from knowledge of the maximal ideals of $C[0,1]$ alone:

$$[0,1] \leftrightsquigarrow \{\text{maximal ideals of } C[0,1]\}.$$

The goal then becomes to study the "geometry" of the space $[0,1]$ by studying the maximal ideals of the ring $C[0,1]$ instead. In this case it is not so clear as to *why* one would want to do this, but this sets the stage for similar ideas in the setting of other rings.

Derivatives via algebra. Consider now the smaller ring $C^\infty[0,1]$ of *infinitely-differentiable* functions on $[0,1]$. Again, it is true that the maximal ideals of this ring are precisely the same m_p from before. We claim now that one can study the concept of a derivative solely from the perspective of these maximal ideals. If so, then this should give some insight for why this approach might be useful: we can study derivatives without having to introduce the concept of a "limit". In settings where the notion of a "limit" does not make sense, we can still hope to talk about "derivatives" via algebraic means.

Let $f \in m_p \subseteq C^\infty[0,1]$. According to a version of Taylor's Theorem (again, we are using some analysis here), we can express f as

$$f(x) = f(p) + f'(p)(x-p) + \frac{1}{2}f''(p)(x-p)^2 + h(x)(x-p)^2$$

for some function $h(x) \in C^\infty[0,1]$. Here, $f(p) = 0$ since $f \in m_p$. Now consider the element f gives in the quotient m_p/m_p^2, where $m_p^2 = m_p m_p$ is the product of the ideal m_p with itself, which is the set of all finite sums of elements of the form $a(x)b(x)$ with $a(x), b(x) \in m_p$. In particular, since $x - p \in m_p$, $(x-p)^2 \in m_p^2$. Thus, in m_p/m_p^2 we get:

$$f(x) = 0 + f'(p)(x-p) + \frac{1}{2}f''(p)\underbrace{(x-p)^2}_{0} + h(x)\underbrace{(x-p)^2}_{0} = f'(p)(x-p).$$

Hence, the only piece of data which characterizes what f gives in the quotient is $f'(p)$, so that we can recover the derivative of f at p purely algebraically from $f \in m_p/m_p^2$. (No limits needed! If all we have available is f and m_p, we can still essentially find $f'(p)$.) For a function g which did not vanish at p, we can apply this reasoning to the function $g(x) - g(p) \in m_p$ to still recover $g'(p)$. Higher-order derivatives can also be recovered algebraically, using quotients such as m_p^2/m_p^3 to find $f''(p)$, m_p^3/m_p^4 to find $f'''(p)$, and so on.

Ideals in polynomial rings. Now consider $\mathbb{C}[x]$, the ring of polynomials over $\mathbb{C}$. Any ideal of the form $(x-a)$ for $a \in \mathbb{C}$ is maximal, and in fact these give all of the maximal ideals. (This uses the fact that $\mathbb{C}$ is what's called *algebraically closed*, which is a concept we will study next quarter. The key point is that any polynomial with complex coefficients has a root within $\mathbb{C}$ itself.) Thus, as before, we can characterize a maximal ideal using a single complex number, so we can recover $\mathbb{C}$ from knowledge of the maximal ideals of $\mathbb{C}[x]$ alone:

$$\mathbb{C} \leftrightsquigarrow \{\text{maximal ideals of } \mathbb{C}[x]\} \quad \text{via} \quad a \leftrightsquigarrow (x-a).$$

Visually, we literally think of this set of maximal ideals as being the complex plane itself.

Similarly, we can describe all maximal ideals of $\mathbb{C}[x, y]$ as those of the form $(x - a, y - b)$, generated by two polynomials $x - a$ and $y - b$ with $a, b \in \mathbb{C}$. Thus, a maximal ideal here is characterized completely by a pair $(a, b) \in \mathbb{C}^2$, and hence we have a correspondence:

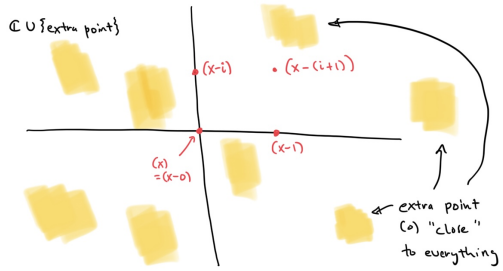
$$\mathbb{C}^2 \leftrightsquigarrow \{\text{maximal ideals of } \mathbb{C}[x, y]\}.$$

We can keep going, and interpret $\mathbb{C}^n$ in general as obtained from the set of maximal ideals of $\mathbb{C}[x_1, \ldots, x_n]$. For the most part, this works pretty well for something like $\mathbb{R}[x]$ as well, except that there are more maximal ideals now than just those generated by degree 1 polynomials.

Algebraic Geometry. But why consider maximal ideals alone? Why not consider prime ideals as well? Doing so leads to the subject of modern *algebraic geometry*, where this all really takes off. Any maximal ideal is prime, so we have already described many prime ideals of $\mathbb{C}[x]$ above. It turns out there is just one more prime ideal: the ideal (0) generated by zero. The set of prime ideals of a commutative ring with unity is called its *spectrum*, so

$$\text{spectrum of } \mathbb{C}[x] = \{(x - a) \mid a \in \mathbb{C}\} \cup \{0\}.$$

We have indicated above that we should visualize the prime ideals $(x - a)$ as points $a \in \mathbb{C}$, so how do we visualize the extra "point" (0)? (Note, (0) should not be visualized as the complex number 0, since 0 is what corresponds to the ideal $(x - 0)$ instead.) The answer is that we visualize (0) as an extra point that exists outside of the usual picture of $\mathbb{C}$, but which in some sense is "everywhere" and somehow "infinitely close" to any other point (!!!)

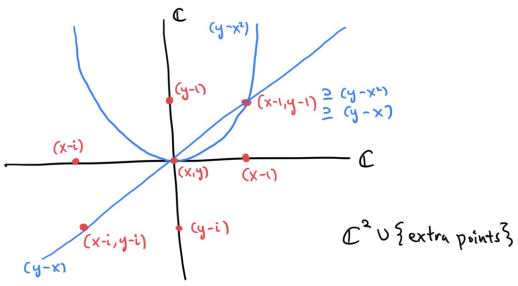


This seems pretty crazy at first, but is really the type of insight which makes algebraic geometry such a powerful tool. (We will not attempt to define what "infinitely close" or "everywhere" mean in this context. Doing so requires the language of *topology*, where the technical fact is that the *closure* of (0) in the spectrum is the entire spectrum.) The basic idea behind considering all this is that anything which is true of (0) should be true of *all* points, since (0) in a way "approximates" any other point. (Again, this all sounds incredibly vague here, but can in fact be made precise.)

What about the spectrum of $\mathbb{C}[x, y]$? Again we have the maximal ideals—visualized as ordinary points in $\mathbb{C}^2$—and (0), and it turns out the only other prime ideals are those generated by *irreducible* (i.e. cannot be factored) polynomials in $\mathbb{C}[x, y]$. For instance, $y - x^2 \in \mathbb{C}[x, y]$ generated a non-maximal prime ideal. How should we visualize this "point" in the spectrum? Any prime ideal is contained in a maximal one (Zorn's lemma), so we can ask for when we have

$$(y - x^2) \subseteq (x - a, y - b)$$

for some $a, b \in \mathbb{C}$, since $(x - a, y - b)$ describes all maximal ideals. It turns out that this is true precisely when $a, b \in \mathbb{C}$ satisfy the equation of the given irreducible polynomial, so in this case when $b - a^2 = 0$. That is, the maximal ideals containing $(y - x^2)$ correspond to the usual points on the parabola $y = x^2$ when we draw the spectrum of $\mathbb{C}[x, y]$ as if it were $\mathbb{C}^2$:



(Due to the limitations of our living in a three-dimensional world, here we have collapsed $\mathbb{C}$—ordinarily drawn as a plane—to be visualized as a line, so that we can draw $\mathbb{C}^2$ as a "plane".) The "point" $(y-x^2)$ is then something which exists outside of this plane picture of $\mathbb{C}^2$, but which is in a sense "infinitely close" to all the usual points on the parabola $y = x^2$; that is, we think of the entire parabola *itself* as also being a "point" of the spectrum of $\mathbb{C}[x, y]$. As another example, the prime ideal $(y - x)$ is visualized as the entire line $y = x$ (as one single object), containing on it all the points (a, b) satisfying $a = b$ corresponding to the maximal ideals $(x-a, y-b)$ which contain $(y-x)$. So, our picture is now complete: the "space" given by the prime ideals of $\mathbb{C}[x, y]$ looks like $\mathbb{C}^2$, with extra "points" included corresponding to curves defined by irreducible polynomial equations (such "points" are "infinitely close" to all of the usual points on that curve), and an extra point (0) which is "everywhere" included as well. (Again, the blue curves in the picture above actually present *single* points in the spectrum. The "fuzzy" yellow point (0) from the previous picture which is "everywhere" is omitted in this new picture, but it still exists in this space.)

Crazy as this all sounds, it really turns out to be extremely useful. For instance, one can now try to make sense of a type of "derivative" in this setting, by considering quotients of certain maximal ideals by their squares in a way analogous to the infinitely differentiable function case. All of this takes place without having any type of "limit" available, which perhaps hints at why this subject might be worth considering.

Lecture 9: Rings of Fractions

Warm-Up, kind of. (This not the typical type of "Warm-Up" we usually see, but rather just a fun thought experiment meant to further clarify the "geometry via algebra" approach we outlined as time.) Consider the polynomial ring $\mathbb{C}[x]$, and the corresponding "space" $\mathbb{C}$ formed by taking its maximal ideals. If we are really meant to treat this as an honest "space", then in particular we should be able to evaluate functions at the "points" of this space and get numbers as a result. So, we ask: given $f(x) \in \mathbb{C}[x]$, how can we treat this as a function on the space of maximal ideals, or, more precisely, how can we "evaluate" this function at a "point" of this space and get "value" in $\mathbb{C}$ as result? That is, we want to treat f like the following type of object:

$$f : \{\text{maximal ideals of } \mathbb{C}[x]\} \to \mathbb{C},$$

just as you can interpret (if you want) a polynomial as a function $f : \mathbb{C} \to \mathbb{C}$.

So, we need to know how to produce an element of $\mathbb{C}$ from the data of f, $\mathbb{C}[x]$, and a maximal ideal $M \subseteq \mathbb{C}[x]$ alone. But we in fact already know how to do this: since M is maximal, $\mathbb{C}[x]/M$ is a field, which is actually isomorphic to $\mathbb{C}$ itself, and thus, the "projection map":

$$\mathbb{C}[x] \to \mathbb{C}[x]/M \cong \mathbb{C}, \;\; f \mapsto \overline{f},$$

where $\overline{f}$ is the element which f gives in the quotient, is precisely what we want! This associates to f and M an element of $\mathbb{C}$, which we interpret as "f evaluated at M". So, we can indeed treat

elements of the ring $\mathbb{C}[x]$ as if they were functions on the space of maximal ideals. It is fair to ask in what way this mimics the usual way of thinking about a polynomial as a function, and the answer is that it is exactly the same: if we express the maximal ideal in question as $M = (x - a)$ for some $a \in \mathbb{C}$, then the "value" of f at $(x - a)$ obtained in the way above is indeed $f(a)$!. To see this, write f as a polynomial centered at a using a Taylor expansion:

$$f(x) = f(a) + f'(a)(x - a) + \cdots + \frac{f^{(n)}(a)}{n!}(x - a)^n.$$

When we pass to the quotient $\mathbb{C}[x]/(x - a) \cong \mathbb{C}$, all of the terms with a factor of $(x - a)$ in them become zero, so all we are left with is $f(a) \in \mathbb{C}$. In other words, the map

$$\mathbb{C}[x] \to \mathbb{C}[x]/(x - a) \cong \mathbb{C}$$

which sends a polynomial to its image in the quotient is precisely "evaluation at a": $f \mapsto f(a)$. The upshot is that, not only can we treat elements of $\mathbb{C}[x]$ as functions on the space of maximal ideals, doing so reproduces the ordinary way of treating polynomials as functions. (Higher-order derivatives like $f^{(k)}(a)$ can, as in the infinitely-differentiable function example from last time, be recovered using quotients like $M/M^2 = (x - a)/(x - a)^2$, M^2/M^3, and so on.)

This sets up one of the basic ideas of algebraic geometry, that *any* commutative ring with identity can be treated as the ring of functions on some space, namely the space of maximal ideals of that ring. Treating such an arbitrary ring in this way opens up new approaches to both algebra and geometry, but we will have to leave to another (likely graduate) course to develop. One more thing to point out is that so far we have only spoken about viewing functions defined on the set of maximal ideals, but last time we indicated that prime ideals should also be included as points in the "space" we are constructing. So, how should we "evaluate" an element of a ring on a *prime* ideal instead, in order to still get an element in some field as a result? For instance, how can we use $f(x, y) \in \mathbb{C}[x, y]$ and the prime ideal $(y - x^2) \subseteq \mathbb{C}[x, y]$ in order to get "the value of f at $(y - x^2)$"? The same idea as above does not quite work, since the quotient $\mathbb{C}[x, y]/(y - x^2)$ is no longer a field, which is bad because "functions" of the type we want should take values in a *field*, not just a random ring. There is a way to fix this using material we will soon develop, and indeed one reason we introduce this question now is to provide at least some context for some definitions and constructions which are still to come.

Functions on a hydrogen atom. Let us give one more example of these ideas, which I have see elsewhere referred to as giving a model of the "functions on a hydrogen atom". Now, this is not at all meant to be a phrase which should be interpreted in a literal way, but merely a nice analogy to clarify what is going on.

Consider the ring $\mathbb{C}[[x]]$ of formal power series over $\mathbb{C}$. (What follows would also work with any field in place of $\mathbb{C}$.) We pointed out before that $\mathbb{C}[[x]]$ has a unique maximal ideal, namely the one (x) generated by x. Thus, the "space of maximal ideals" has only a single point. But, if we include prime ideals as well, there is one more "point": the prime ideal (0). This extra "point" is distinct from the maximal one (x), but is still in some sense arbitrary "close" to (x), and exists "all around" it without any one "definite" location:

If you have seen some physics/chemistry, or more precisely some *quantum mechanics*, this is very much what happens with a hydrogen atom as well: there is a single proton—which we think of as analogous to the point (x)—and then an electron which exists in some kind of "cloud" around the proton but not in any one definite, specific location, similar to the yellow "point" (0) above. This point is just kind of "everywhere" all at once!

One can then move on to consider elements of $\mathbb{C}[[x]]$ as if they were "functions" on this space. What you actually get in this case is perhaps not so interesting (we'll leave it to you to work out the details), but the point is that elements of $\mathbb{C}[[x]]$ cannot really be considered to be functions on $\mathbb{C}$, because there are convergence issues to worry about when working with power series, so we are now providing a setting where viewing a formal power series as a "function" makes sense even without any analysis to consider. (Such is the power of algebraic geometry!) A commutative unital ring with a unique maximal ideal is called a *local ring*, with $F[[x]]$ where F is a field being an example. The term "local" is used here precisely because of this type of example: recall that a power series $\sum a_n(x-a)^n$ in calculus is used to analyse the behavior of a function *locally* near a point a (in the case of $F[[x]]$, this means the behavior near the point given by the unique maximal ideal), and local rings in general can be used in a similar way.

From integral domains to fields. After going off on this brief tangent to introduce some algebro-geometric concepts, we return now to something more concrete. An integral domain is, of course, not a field in general due to a potential lack of inverses. But, in some ways, an integral domain does have "field-like" properties, in particular because cancellation still holds, as we've seen: if $ac = bc$, then $a = b$. This would be true in a field using inverses, but it is also true in an integral domain by applying the lack of zero divisors to $(a-b)c = 0$. So, perhaps integral domains are actually not so far off from being fields. Indeed, the main example of an integral domain we have is $\mathbb{Z}$, and in this case we know that $\mathbb{Z}$ does lie inside of a field, namely $\mathbb{Q}$. So, we want to know whether an integral domain in general can be "enlarged" to a field as well, by somehow forcing every nonzero element to actually have an inverse.

The answer is yes, and we can get a sense for how to construct such a field by using the example of $\mathbb{Z} \subseteq \mathbb{Q}$. A rational number looks like $\frac{a}{b}$ with $a, b \in \mathbb{Z}$ and $b \neq 0$, so a first approach might be to consider such "fractions" with numerator and denominator coming from other integral domain. However, there is a conceptual problem with this, in that we do not know what it means to "divide" elements of an integral domain in general, so it is not clear what $\frac{a}{b}$ should actually mean in that context. (This is not an issue with $\mathbb{Q}$ because we do a priori know beforehand what "division" means.) The way around this is to not focus on using "fractions" at first, but instead on pairs (a, b) consisting of a "numerator" and a "denominator". In the case of $\mathbb{Q}$, this means that instead of $\frac{1}{2}$ we use the pair $(1, 2) \in \mathbb{Z} \times \mathbb{Z}_{\neq 0}$. (The second factor should be nonzero if it is going to be meant to serve as a "denominator".) Thinking of rational numbers as pairs of integers in this way *almost* works, except that there are different pairs which give the same rational: $(1, 2)$, $(2, 4)$, and $(-3, -6)$ for instance all give the same rational $\frac{1}{2} = \frac{2}{4} = \frac{-3}{-6}$. So, if we are literally trying to think of rationals as pairs of integers, we should really declare different pairs to be the "same" when they give the same rational—which can be done by using a certain equivalence relation! Namely, we define $(a, b) \sim (c, d)$ if $\frac{a}{b} = \frac{c}{d}$. Then, it is the *equivalence class* of (a, b) which should be viewed as characterizing the rational $\frac{a}{b}$.

A final wrinkle is that the equation $\frac{a}{b} = \frac{c}{d}$ we have used in order to define this equivalence relation still uses fractions, which we were trying to avoid. But this is easy to fix: we can simply rewrite this equation as $ad = bc$, with no division needed. The point is that this is an equation we can consider *solely* within the context of $\mathbb{Z}$ where multiplication does make sense, without needed to already have a concept for what $\mathbb{Q}$ and $\frac{a}{b}$ means first. Thus, in order to "construct" $\mathbb{Q}$, we define

an equivalence relation $\mathbb{Z} \times \mathbb{Z}_{\neq 0}$ by declaring

$$(a, b) \sim (c, d) \text{ if } ad = bc.$$

It is straightforward to check that this is an equivalence relation (check the details!), and that the set of equivalences classes is in bijection with $\mathbb{Q}$ via $[(a, b)] \leftrightarrow \frac{a}{b}$. In fact, we can take this to be a *definition* of the set $\mathbb{Q}$. This is exactly what we will do for other integral domains in a bit, but we make one final comment: we spent all of this time trying to find a way to make sense of $\frac{a}{b}$ in a way which avoided knowing what fractions were ahead of time, but at the end of the day we will still totally think about the result equivalence classes as if they indeed were fractions. That is, one never thinks of an element of $\mathbb{Q}$ as $[(a, b)]$, but rather as $\frac{a}{b}$, and we will do the same here and for the other rings we will consider in a bit. The point really is that we *only* use the equivalence class approach in order to give a definite meaning to $\frac{a}{b}$—it denotes the equivalence class corresponding to $\frac{a}{b}$, which can also be represented by any $\frac{c}{d}$ with $ad = bc$—but having done so we default to using the fraction notation to denote elements of the resulting set.

We can then define addition and multiplication on the set of equivalence classes via

$$[(a, b)][(c, d)] = [(ac, bd)] \quad \text{and} \quad [(a, b)] + [(c, d)] = [(ad + bc, bd)].$$

This are nothing but the usual addition and multiplication of fractions:

$$\frac{a}{b}\frac{c}{d} = \frac{ac}{bd} \quad \text{and} \quad \frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd},$$

only written in the (numerator,denominator) notation. There is one subtlety here to clarify, namely that these operations are well-defined: we are defining these operations using pairs (a, b), but different pairs might give equal equivalence classes and we have to know that these operations would give the same result even if we used an equivalent pair. In other words, if $(a, b) \sim (a', b')$, we have to know that

$$(a, b)(c, d) \sim (a', b')(c, d) \quad \text{and} \quad (a, b) + (c, d) \sim (a', b') + (c, d)$$

in order to guarantee that the equivalence classes we get the are same either way; otherwise, we could have $X = X'$ but $X + Y \neq X' + Y$ at the level fo equivalence classes, which would be nonsense. (This is the same issue we had to be aware of when considering quotient groups and quotient rings.) It turns out the operations defined above *are* well-defined in this say, but leave the details as to why to you. With these operations, the set of equivalence classes becomes a ring, and indeed a field, which in the end is isomorphic to the usual we think about $\mathbb{Q}$.

Multiplicative sets. The construction of $\mathbb{Q}$ from $\mathbb{Z}$ above via an equivalence relation works in the same way for any integral domain. But, we consider a smaller construction first, where we only seek to invert "some" elements (but not yet necessarily all nonzero) elements of a ring. Suppose R is commutative with identity. We want to construct a ring consisting of "fractions" $\frac{a}{b}$ with $a, b \in R$ (so, technically equivalence classes $[(a, b)]$) where the denominators are constrained to come from some specified subset $D \subseteq R$, which might not yet be all of $R - \{0\}$. But in order for this to work, we need D to have the following property:

> D should be nonempty, not contain 0, contain no zero divisors, and be closed under multiplication: if $b, d \in D$, then $bd \in D$.

We will call a subset $D \subseteq R$ with these properties a *multiplicative set*. (This is not a term the book uses. Most other sources, such as Wikipedia, require $1 \in D$ and otherwise only that D be closed under multiplication in order to be considered multiplicative. We will only need the version of "multiplicative" given here, so will stick with using this term for this specific setup.) The point is that if elements of D are to be the allowable denominators of fractions, these are the properties it should have: the non-empty condition guarantees there exist fractions to speak of in the first place; $0 \notin D$ guarantees that no fraction will have denominator zero; the "closed under multiplication" condition is needed so that the definition of products and sums of fractions makes sense (the denominator of both $\frac{a}{b}\frac{c}{d}$ and $\frac{a}{b} + \frac{c}{d}$ is bd, so if b and d are allowable denominators, bd should be as well); and the "no zero divisors" condition guarantees that the denominator of $\frac{a}{b}\frac{c}{d}$ and $\frac{a}{b} + \frac{c}{d}$ will never be zero.

Examples. Here are some key examples of multiplicative sets. First, if R is an integral domain, then the set $R - \{0\}$ of all nonzero elements is multiplicative. (This is the example which will eventually give us a field.) Next, if $P \subseteq R$ is a *prime* ideal, then its complement $D = R - P$ is multiplicative, which is essentially a way of phrasing the prime condition: "if $ab \in P$, then $a \in P$ or $b \in P$" is the same as saying "if $a \notin P$ and $b \notin P$, then $ab \notin P$", meaning that the complement is closed under multiplication. Note that the first example is just a special of the second by taking $P = (0)$ to be the zero ideal, which is prime in an integral domain.

Finally, for nonzero $b \in R$ which is not a zero divisor, the set of non-negative integer powers of b is multiplicative: $\{1, b, b^2, b^3, b^4, \ldots\}$. One example along these lines we will consider next time will come from taking $b = p \in \mathbb{Z}$ to be prime.

Rings of fractions. Thus, given a multiplicative $D \subseteq R$, we define the *ring of fractions* $D^{-1}R$ to be the set of equivalence classes of the equivalence relation on $R \times D$ defined by

$$(a, b) \sim (c, d) \text{ if } ad = bc.$$

As stated earlier, we will henceforth use the notation $\frac{a}{b}$ for the equivalence class determined by (a, b), and think of elements of $D^{-1}R$ as if they were just ordinary fractions $\frac{a}{b}$ with $a \in R$ and $d \in D$. The notation $D^{-1}R$ is meant to suggest that we have "inverted" every element of D.

Under the usual addition and multiplication of fractions (or their well-defined equivalence class versions), $D^{-1}R$ does form a commutative ring. This ring has an identity element, namely $\frac{d}{d}$ for $d \in D$. The original ring R can be viewed as a subring by considering those fractions of the form $\frac{rd}{d}$ for $r \in R, d \in D$. Finally, every element of D is now a unit, with $\frac{1}{d}$ being the inverse of $d = \frac{dd}{d}$. In fact, $D^{-1}R$ is the *smallest* ring with these properties, meaning that any ring with these properties (having an identity, containing a subring isomorphic to R, and having each element of D be a unit) will contain a subring isomorphic to $D^{-1}R$. This is straightforward but tedious to check, so we omit the details here.

For an example, take $\mathbb{Z}$ with multiplicative set given by the complement of the prime ideal $2\mathbb{Z}$. Then $(\mathbb{Z} - 2\mathbb{Z})^{-1}\mathbb{Z}$ consists of those fractions (i.e. rational numbers) with odd denominators, which is indeed the smallest subring of $\mathbb{Q}$ in which every odd integer has an inverse.

Fraction fields. And now we can complete our original task: enlarging any integral domain into a field. If R is an integral domain, the *fraction field* (or *field of fractions*) of R is the ring of fractions $(R - \{0\})^{-1}R$ corresponding to the multiplicative set $R - \{0\}$. Thus, elements of the fraction field are "fractions" $\frac{a}{b}$ (again, technically equivalence classes) with $a, b \in R$ and $b \neq 0$. R appears as the subring consisting of elements of form $\frac{a}{1}$, and the fraction field is the smallest field containing R, in the sense described before: any field F which contains R (or more precisely,

a subring isomorphic to R) will contain a subfield isomorphic to the field of fractions of R. The fraction $\frac{a}{b}$ then gets interpreted as ab^{-1} in F, where we already have a predetermined notion of inverses. In this language, $\mathbb{Q}$ is thus the fraction field of $\mathbb{Z}$.

Examples. We have seen examples before which can now be interpreted in terms of this new notion. For instance, if R is an integral domain, the fraction field of the polynomial ring $R[x]$ is the *field of rational functions* $R(x)$ over R, which we briefly introduced in the first week:

$$R(x) := \left\{ \frac{p(x)}{q(x)} \;\middle|\; p(x), q(x) \in R[x], q(x) \neq 0 \right\}.$$

The difference between using brackets vs parentheses in a notation is now clear: in general, brackets denote rings, and parentheses the corresponding fraction field. For instance, $\mathbb{Z}[\sqrt{2}]$ is the ring

$$\mathbb{Z}[\sqrt{2}] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Z}\}$$

generated by $\mathbb{Z}$ and $\sqrt{2}$ in $\mathbb{R}$, and $\mathbb{Z}(\sqrt{2})$ is its fraction field:

$$\mathbb{Z}(\sqrt{2}) = \left\{ \frac{a + b\sqrt{2}}{b + d\sqrt{2}} \;\middle|\; a, b, c, d \in \mathbb{Z} \right\}.$$

This fraction field is the smallest subfield of $\mathbb{R}$ which contains $\mathbb{Z}[\sqrt{2}]$.

We pointed out in a previous example that $\mathbb{Q}[\sqrt{2}] = \mathbb{Q}(\sqrt{2})$ are the same, since $\mathbb{Q}[\sqrt{2}]$ is already a field due to the fact that any quotient of numbers of the form $a + b\sqrt{2}$ with $a, b \in \mathbb{Q}$ can be rewritten to be of that same form after "rationalizing" the denominator to get rid of $\sqrt{2}$. In fact, we can also say that $\mathbb{Z}(\sqrt{2}) = \mathbb{Q}(\sqrt{2})$. Indeed, given anything of the form

$$\frac{a + b\sqrt{2}}{c + d\sqrt{2}} \text{ with } a, b, c, d \in Q,$$

we can "clear denoinators" by multiplying everything through by a common denominator m of the those of a, b, c, d:

$$\frac{a + b\sqrt{2}}{c + d\sqrt{2}} = \frac{ma + mb\sqrt{2}}{mc + md\sqrt{2}}.$$

What remains will have integer coefficients, and so belongs to $\mathbb{Z}(\sqrt{2})$. This is actually a general occurrence: if R has fraction field F, then $R[x]$ has fraction field $F(x)$ (whether x is a formal variable, or a more concrete element of some other ring containing R), and the proof is a similar "clearing of denominators". For this reason, the notation $R(x)$ when R is not a field is not commonly used. For instance, the fraction field of $\mathbb{Z}[\pi]$ is $\mathbb{Z}(\pi) = \mathbb{Q}(\pi)$, with $\mathbb{Q}(\pi)$ being the much more common notation. We will see more interesting examples of fraction fields next time.

Lecture 10: More on Fractions

Warm-up 1. Suppose R is commutative and that $P \subseteq R$ is a prime ideal. We show that the ring of fractions $(R - P)^{-1}R$, where we invert elements not in in P, has a unique maximal ideal. (So, it is a *local* ring. This ring of fractions is called the *localization* of R at P and is commonly denoted by R_P. It is used, in algebraic geometry for instance, to study the behavior of geometric objects "near" P, whatever that means.)

We claim that the maximal ideal PR_P of this ring of fractions is the one generated by P itself, only now considered as a subset of the ring of fractions. Multiplying a fraction $\frac{s}{t}$ by an element p of P will give a numerator ps which is in P since P is an ideal, so concretely we have:

$$PR_P = \left\{ \frac{a}{b} \;\middle|\; a \in P, b \notin P \right\} \subseteq R_P.$$

Suppose $I \subseteq R_P$ is an ideal which is *not* contained in PR_P. Then there is an element $\frac{c}{d} \in I$ which is not in PR_P, which means that $c \notin P$. But then $\frac{c}{d}$ is unit since $\frac{d}{c}$ is also in the ring of fractions R_P, so $\frac{c}{d}\frac{d}{c} = \frac{cd}{cd} \in I$ is the identity element of R_P, so $I = R_P$. Thus, the only ideal of R_P not contained in PR_P is all of R_P itself, which shows at once that PR_P is a maximal ideal of R_P, since the only ideal containing it (and not a subset of it) is R_P, and that PR_P is the only maximal ideal, since any other potential maximal ideal would have to actually be contained in PR_P.

To give a sense of how localizations show up in geometry, recall the example of the space of primes ideals of $\mathbb{C}[x, y]$. We previously briefly discussed the idea of viewing elements of $\mathbb{C}[x, y]$ as "functions" on this space, but what remained was how to do define the value of such a function on the type of "point" given by a non-maximal prime ideal P, since in this case $\mathbb{C}[x, y]/P$ is not a field. (Having a quotient as a field is how we were able to "define" the value of a function at the point given by a maximal ideal.) The answer is that instead of taking the quotient $\mathbb{C}[x, y]/P$ right away, we first pass to the localization $\mathbb{C}[x, y]_P$, and then take the quotient of *this* by its unique maximal ideal. In the end this gives an "evaluation at P" map which looks like:

$$\mathbb{C}[x, y] \to \mathbb{C}[x, y]_P \to (\mathbb{C}[x, y]_P)/(P\mathbb{C}[x, y]_P),$$

which does produce an element of the field on the right for any "function" $f \in \mathbb{C}[x, y]$. (Note something interesting here: the "field" in which functions take values actually *changes* as the input point P changes. This is in stark contrast to a usual functions like $f : \mathbb{R} \to \mathbb{R}$, but is needed in order to make this all work out. It turns out to actually be a pretty manageable issue!)

Warm-Up 2. If F is a field, we show that the fraction field of the ring $F[[x]]$ of formal power series is the field $F((x))$ of formal Laurent series. (This is actually half of a problem on the current homework.) The point is that elements of this fraction field are *quotients* of power series, and we claim that such quotients are the same as Laurent series. Indeed, given a formal Laurent series:

$$f(x) = \frac{a_{-N}}{x^N} + \cdots + \frac{a_{-1}}{x} + a_0 + a_1 x + \cdots,$$

factoring out $\frac{1}{x^N}$ expresses $f(x)$ as an element of the fraction field of $F[[x]]$:

$$f(x) = \frac{a_{-N} + a_{-N+1}x + \cdots + a_{-1}x^{N-1} + a_0 x^N + \cdots}{x^N}.$$

Conversely, take a quotient of power series: $\frac{p(x)}{q(x)}$ where

$$q(x) = b_0 + b_1 x + b_2 x^2 + \cdots.$$

If $b_0 \neq 0$, a problem from the first homework shows that $q(x)$ is invertible, so that in this case $\frac{p(x)}{q(x)} = p(x)q(x)^{-1}$ is an honest power series in $F[[x]]$. If instead $b_0 = 0$, take b_N to be the first nonzero coefficient in the expression for $q(x)$, so that

$$\frac{p(x)}{q(x)} = \frac{p(x)}{x^N(b_N + b_{N+1}x + b_{N+2}x^2 + \cdots)}.$$

Since b_N, $b_N + b_{N+1}x + b_{N+2}x^2 + \cdots$ is invertible, so $p(x)/(b_N + b_{N+1}x + b_{N+2}x^2 + \cdots)$ is a power series:

$$\frac{p(x)}{b_N + b_{N+1}x + b_{N+2}x^2 + \cdots} = c_0 + c_1 x + c_2 x^2 + \cdots .$$

Then multiplying through by $\frac{1}{x^N}$ gives a Laurent series:

$$\frac{p(x)}{q(x)} = \frac{1}{x^N}(c_0 + c_1 x + c_2 x^2 + \cdots) = \frac{c_0}{x^N} + \frac{c_1}{x^{N-1}} + \frac{c_2}{x^{N-2}} + \cdots + c_N + c_{N+1}x + \cdots ,$$

so Laurent series and quotients of power series over a field are the same.

Note that it is essentially here that we are taking coefficients in a *field* as opposed to just, say, an integral domain. In particular, the characterization of units in a power series as those for which the constant term is nonzero—an important step in the proof above—relies on having a field. For instance, the fraction field of $\mathbb{Z}[[x]]$ is not $\mathbb{Z}((x))$, the ring of formal Laurent series over $\mathbb{Z}$. (So, the distinction between notation using "brackets vs parentheses" in terms of passing to the fraction field does not hold in this regard, unless we are working over a field.)

p-adic numbers. Now we are able to provide another interesting example of a ring and a field, namely the ring of *p-adic integers* and the field of *p-adic numbers*, denoted by $\mathbb{Z}_p$ and $\mathbb{Q}_p$ respectively. (Take note: the use of $\mathbb{Z}_p$ to denote the ring of p-adic integers is one good reason why we avoided using $\mathbb{Z}_p$ to denote the cyclic group $\mathbb{Z}/p\mathbb{Z}$ last quarter, as many other sources do.) We introduce these as part of our side goal of giving examples of rings which show up in other areas, in order to give a sense of how ubiquitous ring theory actually is. The p-adics are most prominently found in number theory and algebraic geometry, but nowadays they have found uses in topology and even theoretical physics as well.

Fix $p \in \mathbb{N}$ prime. The *ring of p-adic integers* $\mathbb{Z}_p$ is defined to be the ring of "formal power series" $\mathbb{Z}[[p]]$ in the "variable" p. An element in this ring thus takes the form

$$\sum_{n=0}^{\infty} a_n p^n = a_0 + a_1 p + a_2 p^2 + a_3 p^3 + \cdots \text{ with } a_i \in \mathbb{Z}.$$

(We are not concerned with anything like "convergence" at all, and treat such an expression simply as a "formall" one. Addition and multiplication are defined as they usually are for formal series.) Moreover, we can restrict the coefficients to all lie in $\{0, 1, \ldots, p-1\}$, since any other integer coefficient can be "absorbed" into a higher power of p: for instance, for $p = 7$ we have

$$1 + 9 \cdot 7 + 2 \cdot 7^2 = 1 + (2 + 7) \cdot 7 + 2 \cdot 7^2 = 1 + 2 \cdot 7 + 3 \cdot 7^2.$$

This also applies to expressions with negative coefficients: for instance, $-1 = -1 + 0p + 0p^2 + 0p^3 + \cdots$ can be rewritten as

$$-1 = (p-1) + (p-1) \cdot p + (p-1) \cdot p^2 + \cdots = \sum_{n=0}^{\infty} (p-1)p^n.$$

with coefficients now in $\{0, 1, \ldots, p-1\}$. (Again, the expression on the right above is a formal one and not mean to be treated as saying that the series $\sum_{n=0}^{\infty}(p-1)p^n$ "converges" to -1. There *is* actually a notion of "convergence" for which this would in fact be true, but we will say more about that in a bit. To say that -1 is equal to the stated series just means that adding it to $1 = 1 + 0p + 0p^2 + \cdots$ as a formal power series should give zero $0 + 0p + 0p^2 + \cdots$, which you

can check is in fact true.) Any finite power series in p gives an ordinary positive integer (such as $162 = 1 + 2 \cdot 7 + 3 \cdot 7^2$ in the $p = 7$ example used above), and any positive integer can be written as such a finite series by taking its "base p" expansion. (For instance, $p = 2$ gives the binary expansion of a positive integer.) Thus, $\mathbb{Z}$ is contained in $\mathbb{Z}_p$ as a subring, only with negative integers being given by certain *infinite* expansions, as with the example of -1 above. Note that you can also characterize $\mathbb{Z}_p$ as the quotient $\mathbb{Z}[[x]]/(x - p)$, where we take ordinary power series and set $x = p$. The ring $\mathbb{Z}_p$ is in fact an integral domain, as you will show on the homework.

The *field of p-adic numbers* $\mathbb{Q}_p$ is then defined to be the fraction field of $\mathbb{Z}_p$. Concretely, elements of $\mathbb{Q}_p$ can indeed be characterized as "Laurent series" in p, meaning expressions of the form

$$\sum_{n=-N}^{\infty} a_n p^n = \frac{a_{-N}}{p^N} + \cdots + \frac{a_{-1}}{p} + a_0 + a_1 p + \cdots$$

with $a_i \in \{0, 1, \ldots, p-1\}$. The p-adic integers correspond to the case where $a_i = 0$ for $i < 0$. Ordinary positive rational numbers correspond to the case where the are only finitely many terms overall, in either direction. (Negative rationals require infinite expansions however.) To give a sense for how such things are used in geometry, say, recall (if you have had a course in complex analysis before) that ordinary Laurent series in complex analysis are used to study the local behavior of a function near a "singularity"—it turns out the same is roughly true for the p-adics, in that they are used to study the local behavior of objects "near" primes.

Decimal expansions. To give another perspective on the p-adics, and see how they are analogous to real numbers for instance, consider the notion of a "decimal expansion". When we say something like "the decimal expansion of π is $3.14159\ldots$ what we really mean is that π is given by the series:

$$\pi = 3 + 1 \cdot \frac{1}{10} + 4 \cdot \frac{1}{10^2} + 1 \cdot \frac{1}{10^3} + 5 \cdot \frac{1}{10^4} + 9 \cdot \frac{1}{10^5} + \cdots .$$

(Think of this as a "10-adic" expansion.) So, p-adics are exactly analogous, only that we use powers of p instead of 10 and allow such an expansion to have infinitely many terms in the *positive* exponent direction, but only finitely many in the negative exponent direction. Thus, the decimal expansion of a p-adic number

$$\frac{a_{-N}}{p^N} + \cdots + \frac{a_{-1}}{p} + a_0 + a_1 p + \cdots$$

would look like $\ldots a_2 a_1 a_0 . a_{-1} a_{-2} \ldots a_{-N}$, with infinitely digits to the *left* instead of the right. The p-adic decimal expansion of $-1 \in \mathbb{Z}_5 \subseteq \mathbb{Q}_5$ for instance is

$$\ldots 44444444.0$$

using $-1 = \sum_{n=0}^{\infty} (p-1) p^n$. (A positive rational would have a p-adic decimal expansion which is also finite in the leftward direction, in addition to the right.) Flip the (base 10) decimal expansion of an ordinary positive real number around and you get a valid decimal expansion of a p-adic number, apart from the restriction on digits. (So, for $p > 10$, you do literally get a p-adic.)

The upshot is that real numbers, in fact, are not the only way of "extending" $\mathbb{Q}$ to get more types of "numbers", and p-adics can play a similar role. We will say just a bit about this in a second, and see that literally $\mathbb{R}$ and $\mathbb{Q}_p$ can be obtained from $\mathbb{Q}$ via exactly the same type of construction.

$\sqrt{2}$ in the 7-adics. To give an example of the extent to which we can treat $\mathbb{Q}_p$ as a replacement for $\mathbb{R}$, we can consider the problem of determining whether "$\sqrt{2}$" is an element of $\mathbb{Q}_p$. Now, to

be clear, what we mean by this is whether or not there exists $a \in \mathbb{Q}_p$ such that $a^2 = 2$, since this equality should be taken as the defining property of what $\sqrt{2}$ usually means in $\mathbb{R}$. In the usual real number case, saying

$$\sqrt{2} = 1.4142\ldots = 1 + 4 \cdot \frac{1}{10} + 1 \cdot \frac{1}{10^2} + 4 \cdot \frac{1}{10^3} + 2 \cdot \frac{1}{10^4} + \cdots$$

means that multiplying the series on the right by itself should give 2, so, we look for the exact same thing in $\mathbb{Q}_p$, only with a Laurent series in terms of powers of p.

For instance, we claim that $\sqrt{2}$ is an element of $\mathbb{Q}_7$. (Again, forget about what $\sqrt{2}$ usually means in $\mathbb{R}$—here we mean something satisfing $a^2 = 2$.) In fact, we claim that this will actually be an element of $\mathbb{Z}_p \subseteq \mathbb{Q}_p$, so that $\sqrt{2} \in \mathbb{Q}_p$ is an "integer". The element we seek should be of the form

$$a = a_0 + a_1 \cdot 7 + a_2 \cdot 7^2 + \cdots$$

where $a_i \in \{0, 1, \ldots, 6\}$ and should satisfy

$$\begin{aligned} a^2 &= (a_0 + a_1 \cdot 7 + a_2 \cdot 7^2 + \cdots)(a_0 + a_1 \cdot 7 + a_2 \cdot 7^2 + \cdots) \\ &= a_0^2 + 2a_0a_1 \cdot 7 + (2a_0a_2 + a_1^2) \cdot 7^2 + \cdots \\ &= 2 + 0 \cdot 7 + 0 \cdot 7^2 + \cdots = 2. \end{aligned}$$

To begin constructing such an element, consider the "constant" term a_0^2 in the product a^2. Certainly, there is no integer a_0 for which this will equal 2 in the result we want, but the point to recall is that coefficients larger than 6 (in this $p = 7$ case) will be absorbed into higher powers of 7, so that we do not need a_0^2 to be literal equal to 2, only that it be so mod 7. Thus, we begin by taking $a_0 = 4$, since this does satisfy $a_0^2 \equiv 2 \bmod 7$. (Choosing $a_0 = 3$ would also work, and would give a different square root of 2 than the one we are constructing here.) Then, in the square of the series we are constructing, the a_0^2 term actually gives

$$16 = 2 + 2 \cdot 7,$$

where the initial 2 is precisely what we are aiming to have at the end. The second 2 gets pushed into 7 to the first power, and the idea is then to choose the remaining coefficients $a_1, a_2, \ldots$ to push this extra piece further and further to the right (and into larger and larger powers of 7), so that at the end of the day all we are left with is indeed

$$2 + 0 \cdot 7 + 0 \cdot 7^2 + \cdots = 2.$$

So, next we determine how to choose a_1 in:

$$\text{“}\sqrt{2}\text{”} = 4 + a_1 \cdot 7 + a_2 \cdot 7^2 + \cdots .$$

Squaring this gives the "first-order" term $2a_0a_1 \cdot 7$ in the expression for a^2 we computed above, but actually we now have to take into account the $4^2 = 2 + 2 \cdot 7$ from the first step, which modifies the first-order part by adding an additional 2. In other words, we now have:

$$(4 + a_1 \cdot 7 + a_2 \cdot 7^2 + \cdots)(4 + a_1 \cdot 7 + a_2 \cdot 7^2 + \cdots) = 2 + (2 + 2a_0a_1) \cdot 7 + \cdots$$

where $a_0 = 4$, so if we want this to equal $2 + 0 \cdot 7 + 0 \cdot 7^2 + \cdots$, we need $2 + 2a_0a_1$ to be zero, but in fact only 0 mod 7 since a larger value can be absorbed into a higher power. Thus, a_1 should be

chosen to satisfy $2 + 8a_1 \equiv 0 \bmod 7$, which is the same as $2 + a_1 \equiv 0 \bmod 7$ since $8 = 1$, so we can take $a_1 = 5$. Our desired square root of 2 now looks like:

$$4 + 5 \cdot 7 + a_2 \cdot 7^2 + \cdots .$$

Upon squaring this, we see that the second-order 7^2 term will be:

$$(4a_2 + 5^2 + 4a_2) \cdot 7^2,$$

so a_2 should be chosen to satisfy

$$8a_2 + 25 \equiv a_2 + 3 \equiv 0 \bmod 7,$$

so $a_2 = 4$ works. Thus we have so far

$$\sqrt{2} = 4 + 5 \cdot 7 + 4 \cdot 7^2 + a_3 \cdot 7^3 + \cdots .$$

As a check that this is working as intended, let us square this element:

$$\begin{aligned}(4 + 5 \cdot 7 + 4 \cdot 7^2 + a_3 \cdot 7^3 + \cdots)^2 &= 16 + 40 \cdot 7 + (32 + 25) \cdot 7 + \cdots \\ &= 2 + (2 + 40) \cdot 7 + (32 + 25) \cdot 7^2 + \cdots \\ &= 2 + 0 \cdot 7 + (6 + 32 + 25) \cdot 7^2 + \cdots \\ &= 2 + 0 \cdot 7 + 0 \cdot 7^2 + (9 + \text{stuff}) \cdot 7^3 + \cdots ,\end{aligned}$$

which is matching up with $2 = 2 + 0 \cdot 7 + 0 \cdot 7^2 + \cdots$ so far. We can keep going to find a_3, a_4, and so on in the same way and see that there will be an element $a \in \mathbb{Q}_7$ satisfying $a^2 = 7$ as claimd.

Now, actually *proving* that this will work is a different story, and takes some other results in number theory and algebra to actually do. The point is that the process of finding the required coefficients is not one we can actually complete in any real way since it would take an infinite amount of time, but it is in fact possible to show that such a 7-adic integer does actually exist, thereby showing that $\sqrt{2}$—interpreted appropriately—is an element $\mathbb{Q}_7$. You can then also ask questions such as: is π in $\mathbb{Q}_p$? Is e in $\mathbb{Q}_p$? Is i in $\mathbb{Q}_p$? Amazing stuff.

p-adic analysis. We finish our brief discussion of the p-adics by outling an alternate construction of $\mathbb{Q}_p$. We will use the language of *metric spaces*, *Cauchy sequences*, and *completions* from analysis, so there is definitely some background required here. But, in the end, we only do this for the fun of it and this material is not something we will come back to, so no worries if you have not seen the required analysis by this point in your mathematical careers.

Recall from the first homework that for p prime there was a *discrete valuation* on $\mathbb{Q}^\times$ defined by taking $\nu(\frac{a}{b})$ to be the largest power of p which "divides" the fraction $\frac{a}{b}$: factor p out of a and b as much as possible and write the fraction as $\frac{a}{b} = p^\alpha \frac{c}{d}$ with c, d not divisible by p, and take $\nu(\frac{a}{b}) = \alpha$. So, ν is positive when the numerator is more divisible by p than the denominator, and negative when the opposite is true. We then define the *p-adic metric* on $\mathbb{Q}$ by $d(r, r) = 0$ for $r \in \mathbb{Q}$ and

$$d(r, s) = p^{-\nu(r-s)} \text{ for } r \neq s \in \mathbb{Q}.$$

(Recall that a metric gives a way to measure "distance", so this gives a distance function on $\mathbb{Q}$ which is different the usual one given by taking an absolute value.) This is indeed a metric (for instance, it satisfies the triangle inequality), and the intuition is that things which are divisible by large positive powers of p become "small" and things which are divisible by small negative powers

of p become "large". (So, this metric essentially gives a way to interpret algebraic questions about divisibility in an analytic way instead.)

With respect to this metric, $\mathbb{Q}$ is not complete, in the sense that not all Cauchy sequences converge. We can thus consider the *completion* of $\mathbb{Q}$ with respect to the p-adic metric (the completion is a construction which gives the "smallest" complete metric space containing $\mathbb{Q}$ equipped with this metric), and the fact is that $\mathbb{Q}_p$ is precisely this completion. It thus makes sense to talk about the distance between p-adic numbers, and this in turn gives rise to notions like convergence, continuity, differentiability, and so on. From this point of view, $\mathbb{Q}_p$ truly is on the same footing as $\mathbb{R}$, since both are obtained by completing a certain metric on $\mathbb{Q}$ (the usual absolute value metric in the case of $\mathbb{R}$), giving rise to different forms of "analysis". In the $\mathbb{Q}_p$ case, the resulting subject is known as *p-adic analysis*, a now fundamental topic in number theory and numerous other areas. Good stuff!

Lecture 11: Chinese Remainder Theorem

Warm-Up. We determine the first few terms in the 5-adic expansion of $i \in \mathbb{Z}_5 \subseteq \mathbb{Q}_5$, where by "$i$" we mean a square root of -1, so an element $x \in \mathbb{Z}_5$ such that $x^2 = -1$. First, to be clear, -1 is given by the following power series in powers of 5:

$$-1 = 4 + 4 \cdot 5 + 4 \cdot 5^2 + 4 \cdot 5^3 + \cdots .$$

Indeed, as a check we have:

$$\begin{aligned} 1 + (4 + 4 \cdot 5 + 4 \cdot 5^2 + 4 \cdot 5^3 + \cdots) &= 5 + 4 \cdot 5 + 4 \cdot 5^2 + 4 \cdot 5^3 + \cdots \\ &= 0 + 5 \cdot 5 + 4 \cdot 5^2 + 4 \cdot 5^3 + \cdots \\ &= 0 + 0 \cdot 5 + 5 \cdot 5^2 + 4 \cdot 5^3 + \cdots \\ &= 0 + 0 \cdot 5 + 0 \cdot 5^2 + 5 \cdot 5^3 + \cdots \\ &\vdots \\ &= 0 \end{aligned}$$

where each coefficient of 5 gets pushed into the next larger power of 5, producing coefficients 0 all along the way. (The same happens with $-1 = \sum_{n=0}^{\infty}(p-1)p^n$ for general p.) This says that $4 + 4 \cdot 5 + 4 \cdot 5^2 + 4 \cdot 5^3 + \cdots$ is the additive inverse of 1 in $\mathbb{Z}_5$, so it equals -1.

Now we look for $x = a_0 + a_1 \cdot 5 + a_2 \cdot 5^3 + \cdots$ such that

$$(a_0 + a_1 \cdot 5 + a_2 \cdot 5^3 + \cdots)^2 = 4 + 4 \cdot 5 + 4 \cdot 5^2 + 4 \cdot 5^3 + \cdots .$$

The "constant" zeroeth-order term on the left is a_0^2, so this should match up (mod 5, since larger numbers get pushed into higher powers) with 4 on the right:

$$a_0^2 \equiv 4 \bmod 5.$$

We can thus take either $a_0 = 2$ or $a_0 = 3$, so let us use $a_0 = 2$. (Picking $a_0 = 3$ gives a different square root of -1: if we denote by i the one we construct with $a_0 = 2$, the one with $a_0 = 3$ would be $-i$.) So now we want:

$$(2 + a_1 \cdot 5 + a_2 \cdot 5^3 + \cdots)^2 = 4 + 4 \cdot 5 + 4 \cdot 5^2 + 4 \cdot 5^3 + \cdots .$$

By looking at first-order terms on both sides, we see that a_1 should satisfy

$$4a_1 \equiv 4 \bmod 5,$$

so $a_1 = 1$ works. Next we want:

$$(2 + 1 \cdot 5 + a_2 \cdot 5^3 + \cdots)^2 = 4 + 4 \cdot 5 + 4 \cdot 5^2 + 4 \cdot 5^3 + \cdots .$$

The second-order terms on both sides give the requirement

$$4a_2 + 1 \equiv 4 \bmod 5,$$

so $a_2 = 2$ works. Thus the 5-adic integer we are constructing looks like:

$$i = 2 + 1 \cdot 5 + 2 \cdot 5^2 + a_3 \cdot 5^3 + \cdots .$$

The condition needed on a_3 (from third-order terms in the square) will be $4a_3 + 4 \equiv 4 \bmod 5$, so we can take $a_3 = 0$. Continuing on this way will produce the element we want. (Again, we are not in a position to actually *prove* that this process will work at every step as intended, but in this case it will. In general, it is only for certain p that something like $\sqrt{-1}$ will exist in $\mathbb{Q}_p$.)

Integers mod mn. Now we work towards our final result about "general" rings: the *Chinese Remainder Theorem*. We call this our final result about "general" rings since next time we will begin to focus our attention more on special types of rings, so this is close to the last type of result we will consider for rings without a lot of extra properties. The name "Chinese Remainder Theorem" comes from a special case of the general result which was known to Chinese mathematicians in the 13^{th} century or so. The statement says something about solving systems of equations of congruences, or, more practically, finding an element which produces a prescribed list of "remainders" in quotients.

Before looking at the general statement, we first point out that the following isomorphism is a consequence:

$$\mathbb{Z}/mn\mathbb{Z} \cong \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z} \text{ if } gcd(m, n) = 1.$$

We actually mentioned this isomorphism previously when we first introduced the notion of a ring isomorphism, but we did not give a proof at the time. The fact that this isomorphism holds at the level of *additive* groups is something we already saw last quarter (the right side has an element of additive ;order mn, so it must generate the entire additive group), so the new part is that the multiplicative structures are isomorphic too. By taking groups of units of both sides, we also get the following group isomorphism as a consequence:

$$(\mathbb{Z}/mn\mathbb{Z})^\times \cong (\mathbb{Z}/m\mathbb{Z})^\times \times (\mathbb{Z}/n\mathbb{Z})^\times,$$

which, although it is a statement only about (multiplicative) groups, is actually not so straightforward to prove using only group-theoretic means, and is much simpler to approach ring-theoretically instead. Here we use the fact on the right side that the group of units of a product is the product of the groups of units of each factor:

$$(R \times S)^\times = R^\times \times S^\times \quad \text{(assuming } R, S \text{ both have identities)},$$

which should be straightforward to justify. (The point is that, since the ring operations in a product are defined component-wise, (a, b) will be invertible if and only if a and b are each individually invertible.) This isomorphic of multiplicative groups is one which perhaps you might have guessed by looking at some examples last quarter, but for which a proof was never given.

The map which induces the required isomorphism $\mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$ is simply the one which takes an integer and reduces it both mod m and mod n:

$$x \mapsto (x \bmod m, x \bmod n).$$

The key to it all is really that this map is surjective, so that given some candidate remainders $(r, s) \in \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$, we can find x which has those given remainders mod m and mod n:

$$\text{there exists } x \in \mathbb{Z} \text{ such that } x \equiv r \bmod m \text{ and } x \equiv s \bmod n.$$

The claim that the given map should be injective on the domain $\mathbb{Z}/mn\mathbb{Z}$ then translates into the property that x is unique as an element of $\mathbb{Z}/mn\mathbb{Z}$, so that any other integer with these given remainders will be congruent to x mod mn. Thus, the fact that we have the isomorphism above is *really* the statement that: if m, n are relatively prime, then given r and s there exists $x \in \mathbb{Z}$ (which is unique mod mn) satisfying the system of congruences

$$x \equiv r \bmod m, \;\; x \equiv s \bmod n.$$

This is the original "Chinese" form of the Chinese Remainder Theorem, and the modern version is just the formulation which applies to other rings.

Systems of congruences. Let us consider a concrete example of these ideas, which will helpful not only in understanding what the theorem states, but also in suggesting the key step in the proof. Suppose we want to find $x \in \mathbb{Z}$ satisfying the system

$$\begin{aligned} x &\equiv 1 \bmod 8 \\ x &\equiv 2 \bmod 25. \end{aligned}$$

The first requires that x be of the form $x = 8k + 1$ for some $k \in \mathbb{Z}$, and the second that $x = 25\ell + 2$ for some $\ell \in \mathbb{Z}$, so finding x boils down to finding $k, \ell \in \mathbb{Z}$ such that

$$8k + 1 = 25\ell + 2.$$

After rewriting, we see that we are looking for integer solutions of the *diophantine equation*

$$8k - 25\ell = 1.$$

(A *diophantine equation* is simply one for which we only seek integers solutions.)

But we have briefly said something about such equations previously (last quarter, when proving some properties of finite cyclic groups for instance), namely the relation between 8 and 25 which allows for 1 to be expressed this way: the fact that 8 and 25 are relatively prime. (In general, the smallest positive integer which can be written as $ax + by$ for some given integers a, b is precisely $gcd(a, b)$, a fact we will finally prove next time.) Thus, it is the fact that 8 and 25 are coprime (another term for "relatively prime") is what makes it possible to solve our congruence equations. It turns out that one possible solution for the diophantine equation $8k - 25\ell = 1$ is

$$8(-28) - 25(-9) = 1.$$

(The method for constructing such solutions would be covered more in a number theory course, and depends on the use of the *Euclidean algorithm*. We will touch on this a bit next time in the setting of a certain type of ring known as a *Euclidean domain*, but we will not go heavily into the method for finding such solutions explicitly.) Once we have a solution, we can construct the x we want, either as $8k + 1$ of $25\ell + 2$:

$$x = 8(-28) + 1 = 25(-9) + 2 = -223.$$

By viewing this as an element in $\mathbb{Z}/200\mathbb{Z}$ (note $200 = 8 \cdot 25$), we see that the unique solution in $\mathbb{Z}/200\mathbb{Z}$ to our system is $x = -223 \equiv 177 \bmod 200$.

You can double check on your own that $x = 177$ does satisfy $177 \equiv 1 \bmod 8$ and $177 \equiv 2 \bmod 25$ as intended. All other integer solutions of our system of congruence are then of the form $200t + 177$. In terms of the isomorphism $\mathbb{Z}/200\mathbb{Z} \to \mathbb{Z}/8\mathbb{Z} \times \mathbb{Z}/25\mathbb{Z}$ given by $x \mapsto (x \bmod 8, x \bmod 25)$, this means that 177 is the element which mapsto $(1, 2)$.

Coprimality. As stated above, the key point in solving the given system came in figuring out how to express 1 as $8k - 25\ell = 1$ with $k, \ell \in \mathbb{Z}$, which again is possible to do because 8 and 25 coprime. But we can easily rephrase this fact in terms of ideals of $\mathbb{Z}$ instead: $8k$ is an element of $8\mathbb{Z}$ and $25(-\ell)$ is in $25\mathbb{Z}$, so the property we need is the fact that the sum of these two ideals is all of $\mathbb{Z}$:

$$8\mathbb{Z} + 25\mathbb{Z} = \mathbb{Z}.$$

In other words, saying that 1 is expressible in the given way is the same as saying that it should be an element of the sum $8\mathbb{Z} + 25\mathbb{Z}$, which is the same as saying that this sum is everything. (In general, $a\mathbb{Z} + b\mathbb{Z}$ will equal $gcd(a, b)\mathbb{Z}$, which is just the statement we mentioned above—and will prove next time—that $gcd(a, b)$ is the smallest positive integer expressible as $ax + by$.)

Motivated by this, we make the following definition: given a commutative ring R with unity, we say that two ideals $A, B \subseteq R$ are *coprime* (or *comaximal*) if $A + B = R$. Concretely, this boils down to saying that $a + b = 1$ for some $a \in A$ and $b \in B$. (The term "comaximal" is the one our book uses, and comes from the fact that $A + B$ is as large as possible. But, I think "coprime" is a more common term, which matches the ordinary definition of "coprime" in the case of $\mathbb{Z}$, and so will tend to use that instead.) Note that then we explicitly have $ra + rb = r$ for any $r \in R$.

Chinese Remainder Theorem. Now we can state the general, modern version of the Chinese Remainder Theorem: given a commutative ring R with unity and ideals $A_1, \ldots, A_k \subseteq R$ which are pairwise coprime, the map $r \mapsto (r \bmod A_1, \ldots, r \bmod A_k)$ induces an isomorphism

$$R/(A_1 \cdots A_k) \cong R/A_1 \times \cdots R/A_k.$$

(So, not only can we solve a system of *two* congruence equations as in an earlier example, we can actually solve systems with *any* number of equations, as long as the coprimality conditions are satisifed.) Some clarifications are in order. First, saying that ideals are "pairwise coprime" means that A_i and A_j are coprime for any $i \neq j$, so $A_i + A_j = R$ for $i \neq j$. Second, the notation "$r \bmod A_i$" denotes r viewed as an element of the quotient R/A_i, or in other words it denotes the coset $r + A_i$. The *real* point of this result is the claim that the given map is surjective, which is the "system of congruences"-like claim that given any potential "remainders" $s_i \in R/A_i$, there exists $r \in R$ such that $r = s_i$ in R/A_i (i.e. $r \equiv s_i \bmod A_i$). The injectivity (which comes more easily) says that r is unique up to an element of $A_1 \cdots A_k$—i.e. as an element of $R/(A_1 \cdots A_K$—which, together with the existence of r, recovers the classical "Chinese Remainder Theorem" about integer congruences in the case $R = \mathbb{Z}$, with multiple equations.

The proof starts with the $k = 2$ case and then proceeds by induction. Indeed, if we know the result for two ideals, the case of more ideals follows inductively by considering $A_1 \cdots A_k$ to be the product of two ideals $(A_1 \ldots A_{k-1})A_k$: we get

$$R/[(A_1 \ldots A_{k-1})A_k] \cong R/(A_1 \cdots A_{k-1}) \times R/A_k \cong R/A_1 \times \cdots \times R/A_{k-1} \times R/A_k,$$

where the first isomorphism comes from the $k = 2$ case, and the second from the inductive step. One thing we have to verify here is that coprimality does survive the inductive step: $A_1 \cdots A_{k-1}$ is

coprime to A_k, since if we can write $a_i + a_{k_i} = 1$ for each $i < k$ and some $a_i \in A_i, a_{k_i} \in A_k$ (using the pairwise coprime assumption), then

$$1 = (a_1 + a_{k_1})(a_2 + a_{k_2}) \cdots (a_{k-1} + a_{k_{k-1}}) = a_1 \cdots a_{k-1} + (\text{element of } A_k) \in A_1 \cdots A_{k-1} + A_k$$

after expanding the product. The element of A_k referred to here is made up of all terms in the expansion which involve at least one a_{k_i}, since all of these terms will be in A_k by the ideal condition. (For instance, in the $k = 3$ case this looks like $(a_1 + a_{k_1})(a_2 + a_{k_2}) = a_1 a_2 + [a_1 a_{k_2} + a_2 a_{k_1} + a_{k_1} a_{k_2}]$ where $a_1 a_{k_2} + a_2 a_{k_1} + a_{k_1} a_{k_2} \in A_3$.)

Thus, take $A_1, A_2 \subseteq R$ to be coprime ideals. The map

$$R \to R/A_1 \times R/A_2, \ r \mapsto (r \bmod A_1, r \bmod A_2)$$

is a homomorphism since each component map $R \to R/A_i$ (the natural projection) is a homomorphism. The kernel of this map consists of all r which become zero in R/A_1 and R/A_2, which is true if and only if $r \in A_1$ and $r \in A_2$. Thus the kernel is $A_1 \cap A_2$, so the First Isomorphism Theorem gives

$$R/(A_1 \cap A_2) \cong \text{image}.$$

Now, since A_1 and A_2 are coprime, $A_1 \cap A_2 = A_1 A_2$, so the left side above is $R/A_1 A_2$. (The fact that coprime implies the intersection equals the product was on the set of problems for the second discussion section.) To show the image is all of $R/A_1 \times R/A_2$, pick $a_1 \in A_1$ and $a_2 \in A_2$ such that

$$a_1 + a_2 = 1.$$

Taking this as an equation in R/A_i (i.e. taking it "mod A_i") shows that $a_2 = 1 \bmod A_1$ and $a_1 = 1 \bmod A_2$, since the missing term becomes zero in the quotient. Thus, under our map we get:

$$a_1 \mapsto (a_1 \bmod A_1, a_1 \bmod A_2) = (0, 1) \quad \text{and} \quad a_2 \mapsto (a_2 \bmod A_1, a_2 \bmod A_2) = (1, 0),$$

meaning that $(1, 0)$ and $(0, 1)$ are in the image, and hence the image is everything since these two elements generated $R/A_1 \times R/A_2$ as a ring; to be explicit, for any $(r_1, r_2) \in (R/A_1, R/A_2)$, we have:

$$r_1 a_2 + r_2 a_1 \mapsto ([r_1 a_2 + r_2 a_1] \bmod A_1, [r_1 a_2 + r_2 a_1] \bmod A_2) = (r_1 \cdot 1, r_2 \cdot 1) = (r_1, r_2).$$

Thus we have $R/A_1 A_2 \cong R/A_1 \times R/A_2$ as desired.

Lecture 12: Euclidean Domains

Warm-Up. We prove the *Lagrange Interpolation Theorem*: if F is a field, then given $r_1, \ldots, r_n \in F$ and distinct $a_1, \ldots, a_n \in F$, there exists a unique polynomial $p(x) \in F[x]$ of degree at most $n - 1$ such that $p(a_i) = r_i$ for all $i = 1, \ldots, n$. (The standard statement of this theorem usually includes a bit more, namely what the polynomial $p(x)$ actually looks like. For us, the expression for $p(x)$ could be extracted from the proof of the Chinese Remainder Theorem, and requires knowing explicitly how to write $1 \in F[x]$ as an element in sums of coprime ideals. We do not need the explicit form, so you can check other sources if you'd like to see it.) One standard proof uses linear algebra: turn the conditions $p(a_i) = r_i$ into a system of n linear equations in the n coefficients of $p(x)$ (which has degree $n - 1$), and show this system has a unique solution. But here we seek a cleaner proof.

We will need one fact about polynomials over a field: $p(x) \in F[x]$ has $a \in F$ as a root if and only if $x - a$ divides $p(x)$, in the sense that $p(x) = q(x)(x - a)$ for some $q(x) \in F[x]$. We will take this for granted for the time being, but will give a proof shortly. The key observation in the Lagrange

Interpolation Theorem is then that $p(a_i) = r_i$ can be rephrased as saying $p(x) \equiv r_i \bmod (x - a_i)$. Indeed, $p(a_i) = r_i$ if and only if a_i is a root of $p(x) - r_i$, which by the fact stated above is true if and only if $p(x) - r_i \in (x - a_i)$ (where $(x - a_i)$ is the ideal of $F[x]$ generated by $x - a_i$), meaning that $p(x) = r_i$ in the quotient $F[x]/(x - a_i)$, or in other words $p(x) = r_i \bmod (x - a_i)$.

Thus, consider the ring $F[x]/(x - a_1) \cdots (x - a_n)$, where $(x - a_1) \cdots (x - a_n)$ is the product of the ideals $(x - a_i) \subseteq F[x]$. These are ideals are pairwise coprime: if $a_i \neq a_j$, then

$$a_j - a_i = [x - a_i] - [x - a_j]$$

is an element of the ideal $(x - a_i) + (x - a_j)$, and since $a_i \neq a_j$, this element is nonzero and hence a unit, which implies that $(x - a_i) + (x - a_j) = F[x]$. The Chinese Remainder Theorem thus gives an isomorphism:

$$F[x]/(x - a_1) \cdots (x - a_n) \to F[x]/(x - a_1) \times \cdots \times F[x]/(x - a_n)$$

defined by $p(x) \mapsto (p(x) \bmod (x - a_1), \ldots, p(x) \bmod (x - a_n))$. Hence, by taking the element $(r_1, \ldots, r_n)$ on the right, we find $p(x)$ on the left such that $p(x) = r_i \bmod (x - a_i)$ for each i. Moreover, we can take $p(x)$ to be of degree at most $n - 1$ since the relation

$$(x - a_1) \cdots (x - a_n) = 0$$

in the quotient on the left can be used to derive an expression for x^n in terms of $1, x, \ldots, x^{n-1}$ alone, which can be used to show that any polynomial is $F[x]$ is equivalent to one of degree at most $n - 1$ in $F[x]/(x - a_1) \cdots (x - a_n)$. The uniqueness comes from the fact that no two polynomials of degree at most $n - 1$ will be equivalent in the quotient: $p(x) = q(x)$ if and only if $p(x) - q(x) \in (x - a_1) \cdots (x - a_n)$, but the only polynomial of degree less than n in the product $(x - a_1) \cdots (x - a_n)$ is 0. (Note this product of ideals is the same as the ideal generated by the single element $[x - a_1] \cdots [x - a_n] = x^n + (\text{stuff})$, of which any nonzero multiple has degree $\geq n$.)

Later we will see another application of the same ideas used in this proof: if you consider a similar-in-spirit isomorphism to the one $F[x]/(x-a_1) \cdots (x-a_n) \to F[x]/(x-a_1) \times \cdots \times F[x]/(x-a_n)$ used here, only in the case where the a_i are the *eigenvalues* of a matrix (say, when $F = \mathbb{C}$), you get a step towards a proof of the existence of what's called its *Jordan normal form*. This is a standard topic covered in an abstract linear algebra course (for instance, MATH 334 here), and we will see a very clean proof of this in this course using the ring and module theory we develop.

Bézout's identity. We now justify the claim we've used multiple times, that $gcd(m, n)$ is the smallest positive integer of the form $mx + ny$ for $x, y \in \mathbb{Z}$; in other words, $m\mathbb{Z} + n\mathbb{Z} = gcd(m, n)\mathbb{Z}$. We do this not only to make complete various arguments we've given previously (even stretching back to last quarter!), but also to set the stage for some more specialized types of rings we will soon consider. The key property needed is the *division algorithm* for integers.

To start, here is a proper definition of $gcd(m, n)$: we say that $d \in \mathbb{N}$ is the *greatest common divisor* of m and n if it is a common divisor with the property that for any other divisor c of m and n, we have $c \mid d$. (So, the greatest common divisor should be divisible by all common divisors.) For Bézout's identity, set d to be the minimum of all possible positive expressions $mx + ny$:

$$d := \{mx + ny \mid x, y \in \mathbb{Z} \text{ and } mx + ny > 0\},$$

and pick $x, y \in \mathbb{Z}$ such that $d = mx + ny$. First, if c is any common divisor of m and n, then c divides $mx + ny$, so it divides $d = mx + ny$ as well. Now, to show that d is itself a divisor of both m and n, write

$$m = qd + r, \text{ where } 0 \leq r < d$$

according to the division algorithm. Then we have:

$$r = m - qd = m - q(mx + ny) = m(1 - qx) + n(-qy),$$

so that r is among the set of nonnegative numbers expressible as $mx + ny$. Since $0 \leq r < d$ and d is supposed to be the minimum such *positive* number, we must have $r = 0$, so that $m = qd$ and $d \mid m$ as claimed. The proof that $d \mid n$ is exactly the same, so we conclude that $d = mx + ny$ is indeed a common divisor of m and n which is divisible by all other common divisors, so it is the greatest common divisor $gcd(m, n)$ as claimed.

Euclidean algorithm. One computational point remaining is to determine how to write $gcd(m, n)$ in the actual form $mx + ny$. In particular, if m and n are relatively prime, how do we express 1 as $1 = mx + ny$ in a concrete way? (The need to do so showed up in in the "system of congruences" example we considered last time, and was also alluded to in the setting of polynomials in the comments we made about the Lagrange Interpolation Theorem above.) Now, to be clear, the *actual* way of doing this explicitly is not something we will focus on much at all, and belongs more to a course in number theory. But, it is useful to illustrate the general idea for how it can be done in order to, again, set the stage for some upcoming concepts.

The key idea comes from what's called the *Euclidean algorithm*. Take $m, n \in \mathbb{Z}$ nonzero with $m > n$. The division algorithm gives:

$$m = q_1 n + r_1$$

with $0 \leq r_1 < n$. Now, apply the division algorithm *again* to n and r_1 to get:

$$n = q_2 r_1 + r_2$$

with $0 \leq r_2 < r_1$. At this point note that the common divisors of m and n are precisely the same as the common divisors of n and r_1, and indeed of r_1 and r_2 as well! This comes from the fact that a common divisor of m, n will divide $r_1 = m - q_1 n$, and that a common divisor of n, r_1 will divide $m = q_1 n + r_1$, and similar for r_1, r_2 using the second equation above. Thus, in particular, the *greatest* common divisors of each of these pairs are the same:

$$gcd(m, n) = gcd(n, r_1) = gcd(r_1, r_2).$$

And so on, we continue repeatedly applying the division algorithm, obtaining smaller and smaller remainders r_i at each step, a process which eventually has to end since the decreasing non-negative remainders must eventually hit zero:

$$\begin{aligned}
m &= q_1 n + r_1, \ 0 \leq r_1 < n \\
n &= q_2 r_1 + r_2, \ 0 \leq r_2 < r_1 \\
r_1 &= q_3 r_2 + r_3, \ 0 \leq r_3 < r_2 \\
&\vdots \\
r_{k-2} &= q_k r_{k-1} + r_k, \ 0 \leq r_k < r_{k-1} \\
r_{k-1} &= q_{k+1} r_k.
\end{aligned}$$

At each step, we have $gcd(r_{i-1}, r_i) = gcd(r_i, r_{i+1})$ (which equals $gcd(m, n)$ as well) by the argument we gave above, so since $gcd(r_{k-1}, r_k) = r_k$ from the last equation, we have $r_k = gcd(m, n)$. Thus, the final nonzero remainder in the Euclidean algorithm *is* the greatest common divisor of m, n.

How does this help to express this *gcd* as $mx + ny$? Work the equations above backwards, substituting one into the next: substitute the third-to-last equation $r_{k-3} - q_{k-1}r_{k-2} = r_{k-1}$ into the second-to-last to get:

$$r_k = r_{k-2} - q_k r_{k-1} = r_{k-2} - q_k(r_{k-3} - q_{k-1}r_{k-2}) = (\text{someting})r_{k-3} + (\text{something})r_{k-2}.$$

Then use the fourth-to-last equation $r_{k-4} - q_{k-2}r_{k-3} = r_{k-2}$ to substitute into r_{k-2} here to get

$$r_k = (\text{someting})r_{k-4} + (\text{something})r_{k-3}.$$

And so on, we continue backwards-substituting remainders into the expression for r_k we are building up (using earlier pairs of remainders as we go), until we eventually subsitute $m - q_1 n = r_1$ in for r_1 to get something of the form

$$r_k = (\text{someting})m + (\text{something})n,$$

which is the $gcd(m, n) = mx + ny$ we want. Again, all we care about here is that this works, and we will not really do any explicit computations along these lines, *except* for possibly showing how to find inverses in quotients of $F[x]$ that are fields later on!

Euclidean domains. The point of the observations above is that a lot the structure which $\mathbb{Z}$ has comes from the existence of the division algorithm. So, one might expect that other rings with a similar "division algorithm" too will posses nice properties. This motivates what follows.

We will assume throughout that R is an integral domain. A *norm* on R is a function $N : R \to \mathbb{Z}_{\geq 0}$, mapping elements of R to nonnegative integers, such that $N(0) = 0$. That's it, no other conditions imposed, and we think of N as providing a way to measure how "large" an element of R is. We say R is a *Euclidean domain* if there is a norm N on R for which the following *division algorithm* holds: given any $a, b \in R$ with $b \neq 0$, there exist $q, r \in R$ such that

$$a = qb + r \text{ with } r = 0 \text{ or } N(r) < N(b).$$

So, we think of r as a sort of "remainder" obtained upon "dividing" a by b, which is either 0 or "smaller" than b as measured by the norm. For such a ring, whatever properties the division algorithm on $\mathbb{Z}$ gave us will still hold, such as the existence of a "Euclidean algorithm" (using the same equations as the ones we outlined above in $\mathbb{Z}$), and the fact that every ideal is principal, as we will see. (We never actually gave a formal proof that this was true even of $\mathbb{Z}$, but we will see give one next time for Euclidean domains in general.) Note in the case of $\mathbb{Z}$ that we get a uniqueness (of q and r) statement as well if we require that $0 \leq r < b$, but in general there is no uniqueness required. Moreover, a given might have more than one norm which turns it into a Euclidean domain, which is fine: all we need is one.

The most basic example, of course, is $\mathbb{Z}$ with norm given by the ordinary absolute value. Another example is $\mathbb{Z}[i]$, which is the subring of $\mathbb{C}$ generated by $\mathbb{Z}$ and i:

$$\mathbb{Z}[i] = \{a + ib \in \mathbb{C} \mid a, b \in \mathbb{Z}\}.$$

This is called the ring of *Gaussian integers*, and has some next structure we will see in various examples. The norm on $\mathbb{Z}[i]$ is given by $N(a + ib) = a^2 + b^2$. (We will say something about the existence of the division algorithm in this ring next time.) More generally, rings of the form $\mathbb{Z}[\sqrt{D}]$, with norms defined similarly to that above, also *often* give examples, but, as we'll see, not always.

Polynomials over fields. Apart from $\mathbb{Z}$, the most important example of a Euclidean domain is the polynomial ring $F[x]$, where F is a field. The norm function is given by the usual notion of the

degree of a polynomial, and the division algorithm comes simply from ordinary "long division" of polynomials. We will give an example of this next time, which should be enough to convince you that this will work over any field.

For now, we justify the one fact we used in the Warm-Up, that $a \in F$ is a root of $p(x) \in F[x]$ if and only if $x - a$ divides $p(x)$. Indeed, the division algorithm for polynomials gives

$$p(x) = q(x)(x - a) + r(x)$$

for some $q(x), r(x) \in F[x]$ with $\deg r(x) < \deg(x - a)$. But $x - a$ has degree 1, so $r(x)$ must have degree 0, meaning that it is a constant polynomial $r(x) = r \in F$. Then, we get

$$p(a) = q(a)(a - a) + r = r.$$

Thus, $p(a) = 0$ if and only if $r = 0$, which holds if and only if $p(x) = q(x)(x - a)$ in the original division algorithm. This final equality is precisely what it means for $x - a$ to divide $p(x)$. Hence, over a field at least, we can always "factor" $x-a$ out of a polynomial if a is a root of that polynomial. (We will see later the extent to which this might still hold if we only have coefficients coming from an integral domain instead of a field.)

Lecture 13: More on Euclidean

Warm-Up 1. We illustrate the division algorithm for $(\mathbb{Z}/5\mathbb{Z})[x]$ using x^6+2x+1 and $3x^4+x^2+1$. This looks like:

$$\begin{array}{r} 2x^2 + 1 \\ 3x^4 + x^2 + 1 \;\overline{)\; x^6 \qquad\qquad + 2x + 1} \\ \underline{x^6 + 2x^4 + 2x^2} \\ 3x^4 + 3x^2 + 2x + 1 \\ \underline{3x^4 + x^2 \qquad + 1} \\ 2x^2 + 2x \end{array}$$

First, we look for something to multiply $3x^4$ by in order to obtain the x^6 over $\mathbb{Z}/5\mathbb{Z}$, and $2x^2$ works since $(3x^4)(2x^2) = 6x^6 = x^6$. (Already here, it should be clear why the division algorithm will work over *any* field: whenever we need to multiply ax^k ($a \neq 0$) by something to get bx^ℓ, where $k \leq \ell$, we can always use $ba^{-1}x^{\ell-k}$ since a^{-1} exists in a field. There is thus nothing that will prevent the process we use here from working in general.) Then we compute

$$2x^2(3x^4 + x^2 + 1) = x^6 + 2x^4 + 2x^2.$$

Subtracting this from the original $x^6 + 2x + 1$ gives:

$$(x^6 + 2x + 1) - (x^6 + 2x^4 + 2x^2) = -2x^4 - 2x^2 + 2x + 1 = 3x^4 + 3x^2 + 2x + 1,$$

as it appears above. Next we need $(3x^4)(\text{something}) = 3x^4$, so 1 works. Then $1(3x^4 + x^2 + 1) = 3x^4 + x^2 + 1$, and subtracting from $3x^4 + 3x^2 + 2x + 1$ gives:

$$(3x^4 + 3x^2 + 2x + 1) - (3x^4 + x^2 + 1) = 2x^2 + 2x.$$

Since this has degree less than $3x^4 + x^2 + 1$, this is the desired remainder. Thus the division algorithm in this case gives:

$$x^6 + 2x + 1 = (2x^2 + 1)(3x^4 + x^2 + 1) + (2x^2 + 2x) \text{ in } (\mathbb{Z}/5\mathbb{Z})[x].$$

Principal ideal domains. Before the next Warm-Up, let us introduce the following convenient definition: a *principal ideal domain* is an integral domain in which every ideal is principal. Principal ideal domains are usually simply called "PIDs" for short, and, as we will see, form a very nice class of rings since they are the ones which behave the most like the ring of integers and over which straightforward generalizations of linear algebra are possible.

Warm-Up 2. We show that any Euclidean domain R is a PID. Indeed, suppose I is a nonzero ideal. (The zero ideal is generated by 0, so it is principal.) Pick $x \in I$ to be nonzero of smallest norm $N(x)$ among all nonzero elements of I. We claim that $I = (x)$. Let $y \in I$, and apply the division algorithm in R to get

$$y = qx + r \text{ where } r = 0 \text{ or } N(r) < N(x).$$

Since $r = y - qx$ is a difference of elements of I, it too is in I, so if it were nonzero the fact that $N(r) < N(x)$ would contradict the minimality of $N(x)$. Thus we have $r = 0$, so $y = qx \in (x)$ and hence $I = (x)$ as claimed.

As a consequence, $\mathbb{Z}$ and $F[x]$ are PIDs. The proof above is one we essentially already saw last quarter in the context of showing that subgroups of cyclic subgroups are cyclic: the proof back then used the division algorithm on exponents, and follows precisely the logic above.

Gaussian integers. Apart from $\mathbb{Z}$ and $F[x]$ for F a field, we mentioned last time that the ring of Gaussian integers $\mathbb{Z}[i]$ is an example of a Euclidean domain we will use from time to time. Recall that elements in this ring are $a + ib$ with $a, b \in \mathbb{Z}$, and the norm is $N(a + ib) = a^2 + b^2$. We now verify that this is indeed a Euclidean domain. One fact we use is that this norm is multiplicative, in the sense that $N(xy) = N(x)N(y)$ for $x, y \in \mathbb{Z}[i]$. This comes from the fact that $N(x) = x\overline{x}$ where $\overline{x}$ is the complex conjugate of x, and that complex conjugation is multiplicative: $\overline{xy} = \overline{x}\,\overline{y}$.

First we consider the case where $y = a + ib \in \mathbb{Z}[i]$ and $n \in \mathbb{N} \subseteq \mathbb{Z}[i]$. We apply the integer division algorithm to each of a, b and n to get:

$$a = q_1 n + r_1 \text{ and } b = q_2 n + r_2, \text{ where } |r_1|, |r_2| \leq \frac{n}{2}.$$

Now, some clarification is in order. Usually the integer division algorithm produces a remainder $0 \leq r < n$. But here, we can improve the restriction on the remainder to $|r| \leq \frac{n}{2}$ if we allow the remainder to be negative. Indeed, if $a = qn + r$ with $0 \leq r < n$ and $r > \frac{n}{2}$, then we can instead use $a = (q + 1)n + (r - n)$ where $|r - n| \leq \frac{n}{2}$. (If qn is more than $\frac{n}{2}$ away from a, then $(q + 1)n$ will be at most $\frac{n}{2}$ away from a. For a concrete example, $19 = 2 \cdot 7 + 5$ can be written instead as $19 = 3 \cdot 7 - 2$.) With these remainders, we then have:

$$y = a + ib = (q_1 + iq_2)n + (r_1 + ir_2) \text{ where } N(r_1 + ir_2) = r_1^2 + r_2^2 \leq \frac{n^2}{4} + \frac{n^2}{4} < n^2 = N(n),$$

so the division algorithm holds for $y \in \mathbb{Z}[i]$ and $n \in \mathbb{N}$.

For a general nonzero $x \in \mathbb{Z}[i]$, we apply the case above to $y\overline{x}$ and $x\overline{x} \in \mathbb{N}$. We get

$$y\overline{x} = qx\overline{x} + r \text{ where } N(r) < N(x\overline{x}).$$

By expressing r as $r = y\overline{x} - qx\overline{x}$, we thus have:

$$N(x)N(\overline{x}) = N(x\overline{x}) > N(r) = N(y\overline{x} - qx\overline{x}) = N(y - qx)N(\overline{x}).$$

Dividing by $N(\overline{x})$ gives $N(x) > N(y - qx)$, so

$$y = qx + (y - qx), \text{ with } N(y - qx) < N(x)$$

is of the form required in the division algorithm. Hence $\mathbb{Z}[i]$ is a Euclidean domain, and also a PID.

Greatest common divisors. In a PID R, for any nonzero $a, b \in R$ the ideal (a, b) should be principal, meaning generated by a single element $d \in R$. In the case $R = \mathbb{Z}$, this is simply the result of Bézout's identity:

$$m\mathbb{Z} + n\mathbb{Z} = gcd(m, n)\mathbb{Z},$$

and indeed we can now generalize this concept to other types of rings.

If R is an integral domain and $a, b \in R$ are nonzero, we say that $d \in R$ is a *greatest common divisor* of a and b if it has the following two properties:

- $d \mid a$ and $d \mid b$, where saying x divides y in R means there exists $r \in R$ such that $y = rx$, and
- if $d' \mid a$ and $d' \mid b$, then $d' \mid d$.

The first property says that d is a common divisor of a and b, and the second says it is the "largest" one in the sense that any other common divisor divides it. (Note that an inequality like $x < y$ has no meaning in a ring, which is why we rephrase "greatest" in the way above. In the case of $\mathbb{Z}$, of course, we get the usual notion of gcd.) Note also that we speak of "a" greatest common divisor instead of "the" since a greatest common divisor, if it exists, is not unique. This is even true in $\mathbb{Z}$, where we can take both 2 and -2 to be greatest common divisors of 4 and 6 for instance; by convention, in $\mathbb{Z}$ we usually use the positive one as the gcd, but again "positive" has no meaning in a general ring. What is true is that in a PID any two greatest common divisors can only "differ" by a unit, as we will clarify next time.

And so we claim that in a PID R, greatest common divisors always exist. As above, we have $(a, b) = (d)$ for some $d \in R$, and we claim that this d is a greatest common divisor of $a, b \in \mathbb{R}$. Since $a, b \in (a, b) = (d)$, we have $a = dk$ and $b = d\ell$ for some $k, \ell \in R$, so d is a common divisor of a and b. If d' is any other common divisor, then $a, b \in (d')$ so $(a, b) \subseteq (d')$, which means that $(d) \subseteq (d')$. But this says $d = d'r$ for some r, so $d' \mid d$ and hence d is a greatest common divisor of a and b.

Moreover, we also get that d can be written as $d = xa + yb$ for some $x, y \in R$, since this is what elements in $(d) = (a, b)$ looks like. Such an expression will be called an *R-linear combination* of a and b, and should be viewed as an analog of "linear combinations" from linear algebra, only with "coefficients" x, y coming from the ring R. We will come back to such linear combinations in the context of module theory later.

Euclidean vs PID. So, in any PID, greatest common divisors exist and can be expressed as R-linear combinations, generalizing Bézout's identity. Any Euclidean domain is a PID, so we can ask whether the extra Euclidean structure gives us anything a non-Euclidean PID would not. The answer is yes, in that in a Euclidean domain we have an explicit way of computing greatest common divisors *and* of expressing them as R-linear combinations. This is a consequence of the Euclidean algorithm, which we outlined last time for $R = \mathbb{Z}$, but works in the same way for any Euclidean domain. In this course we will not have much need to compute these things explicitly—apart from on the homework for practice and a Warm-Up next time—but being able to do so is important in more computational applications of ring theory.

Lecture 14: Principal Ideal Domains

Warm-Up. We determine a greatest common divisor of x^6+2x+1 and $3x^4+x^2+1$ in $(\mathbb{Z}/5\mathbb{Z})[x]$, and express it as a linear combination of these two. This is done via the Euclidean algorithm, whose first step we carried out last time:

$$x^6+2x+1=(2x^2+1)(3x^4+x^2+1)+(2x^2+2x).$$

Next, dividing $3x^4+x^2+1$ by $2x^2+2x$ gives:

$$\begin{array}{r|l} & 4x^2+x+2 \\ \hline 2x^2+2x & 3x^4\quad\quad +x^2+1 \\ & 3x^4+3x^3 \\ \hline & \quad 2x^3+x^2+1 \\ & \quad 2x^3+2x^2 \\ \hline & \quad\quad 4x^2+1 \\ & \quad\quad 4x^2+4x \\ \hline & \quad\quad\quad x+1 \end{array}$$

Thus:

$$3x^4+x^2+1=(4x^2+x+2)(2x^2+2x)+(x+1).$$

At the next step, $2x^2+2x$ is divisible by $x+1$, so the Euclidean algorithm ends:

$$\begin{aligned} x^6+2x+1&=(2x^2+1)(3x^4+x^2+1)+(2x^2+2x)\\ 3x^4+x^2+1&=(4x^2+x+2)(2x^2+2x)+(x+1)\\ 2x^2+2x&=2x(x+1). \end{aligned}$$

Hence $x+1$ is a greatest common divisor of x^6+2x+1 and $3x^4+x^2+1$.

To express it as a $(\mathbb{Z}/5\mathbb{Z})[x]$-linear combination of these, we write

$$x+1=3x^4+x^2+1-(4x^2+x+2)(2x^2+2x)$$

from the second equation above, and substitute in

$$2x^2+2x=x^6+2x+1-(2x^2+1)(3x^4+x^2+1)$$

from the first. We get:

$$\begin{aligned} x+1&=3x^4+x^2+1-(4x^2+x+2)(2x^2+2x)\\ &=3x^4+x^2+1-(4x^2+x+2)[x^6+2x+1-(2x^2+1)(3x^4+x^2+1)]\\ &=(-4x^2-x-2)(x^6+2x+1)+(1+[4x^2+x+2][2x^2+1])(3x^4+x^2+1)\\ &=(x^2+4x+3)(x^6+2x+1)+(3x^4+2x^3+4x^2+x+3)(3x^4+x^2+1) \end{aligned}$$

as the desired $(\mathbb{Z}/5\mathbb{Z})[x]$-linear combination.

Associate elements. We can rephrase the result above as the following ideal equality:

$$(x^6+2x+1,3x^4+x^2+1)=(x+1).$$

But $x+1$ is not the only generator of the ideal on the right, and any such generator also serves as a greatest common divisor of x^6+2x+1 and $3x^4+x^2+1$. In general, if $(d)=(d')$ in an integral domain, then d and d' are multiples of each other:

$$d = kd' \text{ and } d' = \ell d$$

for some k, ℓ. But this gives $d = k\ell d$, so $d(1-k\ell)=0$. If $d \neq 0$, this then forces $k\ell = 1$ (recall we are in an integral domain), so that k, ℓ are units. Thus, d and d' generate the same ideal if and only if one is the other times a unit. In the example above, $2(x+1), 3(x+1)$, and $4(x+1)$ are also greatest common divisors of x^6+2x+1 and $3x^4+x^2+1$ in $(\mathbb{Z}/5\mathbb{Z})[x]$.

Two elements such as d and d' above ("differing" by a unit) are said to be *associate*. For instance, n and $-n$ are associate in $\mathbb{Z}$.

Examples of PIDs. Recall that a PID (principal ideal domain) is an integral domain in which every ideal is principal. We saw last time that any Euclidean domain is a PID, so $\mathbb{Z}$, $\mathbb{Z}[i]$, and $F[x]$—with F a field—are examples of PIDs. Also, the ring $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-19}]$ can be shown to be a PID. (This is the smallest subring of $\mathbb{C}$ which contains $\mathbb{Z}$ and $\frac{1}{2}(1+\sqrt{-19})$, and, it turns out, explicitly consists of complex numbers of the form $(a+\frac{1}{2})+(b+\frac{1}{2})\sqrt{-19}$ with $a, b \in \mathbb{Z}$.) This is an interesting example since it is a PID which is *not* a Euclidean domain, as in there is no norm which will give a valid division algorithm in this ring. We omit the proofs of these facts here since they will not be important for us, but you can check the book for the details.

One question to ask here is why one would think to consider $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-19})]$ instead of simply $\mathbb{Z}[\sqrt{-19}]$? The quick answer is that $\mathbb{Z}[\sqrt{-19}]$ is not actually a PID, but this is not-so-satisfying since it doesn't shed any light on where $\frac{1}{2}(1+\sqrt{-19})$ comes from. The *real* reason why $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-19})]$ is more "natural" to consider than $\mathbb{Z}[\sqrt{-19}]$ has to do with some number theory, and in particular the notion of a "ring of integers" in an "algebraic number field". We will touch on this briefly next quarter as we develop field theory, but the point is that, for whatever reason, $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-19}]$ is the correct analog of $\mathbb{Z}$ in the field $\mathbb{Q}(\sqrt{-19})$ (an example of what's called an *imaginary quadratic number field*), and $\mathbb{Z}[\sqrt{-19}]$ is not. The book actually discusses this back when first introducing examples of rings, so you can check there for some info, or wait until next quarter.

Non-examples. The ring $\mathbb{Z}[x]$ of polynomials over $\mathbb{Z}$ is not a PID, since we showed a while back as a Warm-Up that the ideal $(4, x+3) \subseteq \mathbb{Z}[x]$ is not principal. (In fact, $\mathbb{R}[x]$ is a PID if and only if R is a field, as we will see shortly.) Also, $\mathbb{Z}[\sqrt{-5}]$ is not a PID. The book gives a direct brute-force proof that the ideal $(3, 2+\sqrt{-5}) \subseteq \mathbb{Z}[\sqrt{-5}]$ is not principal, whose basic idea we will essentially reproduce shortly in a slightly different context. Basically, if this ideal were principal, there would exist some $a+ib \in \mathbb{Z}[\sqrt{-5}]$ such that

$$3 = (a+ib)(c+id) \text{ and } 2+\sqrt{-5} = (a+ib)(s+it)$$

for some $c, d, s, t \in \mathbb{Z}$, and one can show that this is not possible.

Prime implies maximal in a PID. We know that in an integral domain in general, maximal ideals are prime, and in $\mathbb{Z}$ for instance, nonzero prime ideals are always maximal. In fact, this is true in any PID: if R is a PID and $P \subseteq R$ is a nonzero prime ideal, then P is maximal.

To see this, suppose $P=(p)$ for some $p \in R$. If $M=(x)$ is an ideal such that $(p) \subseteq (x) \subseteq R$, we aim to show that $(p)=(x)$ or $(x)=R$. (Recall R is PID, so any ideal M containing P is of this form.) Since $(p) \subseteq (x)$, we have

$$p = ax$$

for some $a \in R$, so $ax \in (p)$. Since (p) is prime, $a \in (p)$ or $x \in (p)$. If $x \in (p)$, then $(x) = (p)$, whereas if $a \in (p)$, we have $a = pk$ for some $k \in R$, so that

$$p = ax = pkx.$$

Since R is a PID, hence an integral domain, this implies $1 = kx$ ($p = pkx$ means $p(1 - kx) = 0$, so $1 - kx = 0$), so x is a unit and thus $(x) = R$. Thus $(p) = P$ is maximal as claimed.

As a consequence of this result, we can now see that $R[x]$ is a PID if and only if R is a field. The backwards direction we saw previously, and for the forwards direction we have: $(x) \subseteq R[x]$ is prime since $R[x]/(x) \cong R$ is an integral domain, so (x) is maximal since $R[x]$ is a PID, and thus $R[x]/(x) \cong R$ is a field.

Primes and irreducibles. We are working towards understanding rings which posses a type of "unique factorization" property, of which PIDs will be a key example. The key types of elements to consider are the following: a nonzero, nonunit element r of an integral domain R is said to be

- *prime* if whenever r divides ab for $a, b \in R$, then r divides a or r divides b, or
- *irreducible* if whenever $r = ab$ for $a, b \in R$, then a or b is a unit.

(Recall that to say r *divides* y in R means $y = rk$ for some $k \in R$.) The first property just says that (r) is a prime ideal, since $y \in (r)$ if and only if r divides y. The second property says that r cannot be factored in a nontrivial way: if a is a unit, certainly we can "factor" r as $r = a(a^{-1}r)$, and to be irreducible means that these are the only possible "factorizations". If r is not irreducible we simply say that it is *reducible*.

Now, in $\mathbb{Z}$ for instance, the two notions above agree: a nonzero nonunit $p \in \mathbb{Z}$ is prime if and only if it is irreducible. Indeed, as we mentioned a while back when introducing the notion of a prime ideal, the second definition above is what you normally think of when you say that p is a prime number, but in fact we proved earlier, using the characterization of $gcd(a, p)$ as the smallest positive integer of the form $ax+py$, that prime numbers have the first property as well. In a general integral domain, we will see next time that primes are always irreducible, but the converse is not true as we will show in an example shortly. Note that the definition above still allows for $2 \in \mathbb{Z}$ to be "factored" in a way other than $2 = 2$, namely as $2 = (-1)(-2)$ for instance, but this is ok since -1 here is a unit.

Irreducible but not prime. The standard example of an integral domain in which not all irreducibles are primes is $\mathbb{Z}[\sqrt{-5}]$. Indeed, we claim that 3 is irreducible in this ring but not prime. To see this will use the following norm: $N(a+b\sqrt{-5}) = a^2+5b^2$, which is simply the one inherited from $\mathbb{C}$ where we take a complex number times its conjugate. We used a similar thing in $\mathbb{Z}[i]$ previously, but here this norm does not turn $\mathbb{Z}[\sqrt{-5}]$ into a Euclidean domain, since the division algorithm will not hold. (Any Euclidean domain is a PID, and we have claimed that $\mathbb{Z}[\sqrt{-5}]$ is not a PID.) But, we can still work this norm (which is multiplicative: $N(zw) = N(z)N(w)$) anyway.

To see that 3 is irreducible, suppose

$$3 = (a + b\sqrt{-5})(c + d\sqrt{-5})$$

for some $a, b, c, d \in \mathbb{Z}$. We can try to multiply out the right side and compare it to the left, and get the following conditions: $3 = ad - 5bd$ and $ad + bc = 0$. The problem is that it is not so straightforward to derive information about a, b, c, d from these, since a lot of messy algebra is involved. Instead, we exploit the norm: taking norms of both sides of the equality above gives

$$N(3) = N(a + b\sqrt{-5})N(c + d\sqrt{-5}), \text{ so } 9 = (a^2 + 5b^2)(c^2 + 5d^2).$$

Now we are in business: since both $a^2 + 5b^2$ and $c^2 + 5d^2$ are nonnegative integers, the only possibilities in $9 = (a^2 + 5b^2)(c^2 + 5d^2)$ are:

$$a^2 + 5b^2 = 1, c^2 + 5b^2 = 9 \text{ or } a^2 + 5b^2 = 3, c^2 + 5b^2 = 3.$$

(The case where $a^2 + 5b^2 = 9, c^2 + 5b^2 = 1$ is simply the first case above after exchanging the roles of $a + b\sqrt{-5}$ and $c + d\sqrt{-5}$). The second case is not possible since $a^2 + 5b^2$ can never equal 3, and $a^2 + 5b^2 = 1$ in the first case forces $a = \pm 1, b = 0$. Thus, if $3 = (a + b\sqrt{-5})(c + d\sqrt{-5})$ we must have $a + b\sqrt{-5} = \pm 1$, which is a unit, so 3 is irreducible. The same type of argument shows that 2 is also irreducible in $\mathbb{Z}[\sqrt{-5}]$, and next time we will use the same idea to show that $1 + \sqrt{-5}$ is also irreducible.

To see that 3 is not prime in $\mathbb{Z}[\sqrt{-5}]$, note that

$$(1 + \sqrt{-5})(1 - \sqrt{-5}) = 6.$$

Thus 3 divides the product on the left, but we claim it does not divide either factor. Indeed, we cannot have

$$1 + \sqrt{-5} = 3(a + b\sqrt{-5})$$

for $a, b \in \mathbb{Z}$ since this would require that $1 = 3a$, so 3 does not divide $1 + \sqrt{-5}$, and 3 does not divide $1 - \sqrt{-5}$ for the same reason. Thus, 3 is irreducible but not prime in $\mathbb{Z}[\sqrt{-5}]$. (2 is another example of an irreducible which is not prime.)

Lecture 15: Unique Factorization

Warm-Up 1. We show that $1 + \sqrt{-5} \in \mathbb{Z}[\sqrt{-5}]$ is irreducible but not prime. Suppose

$$1 + \sqrt{-5} = (a + b\sqrt{-5})(c + d\sqrt{-5})$$

for some $a, b, c, d \in \mathbb{Z}$. Taking norms using $N(x + y\sqrt{-5}) = x^2 + 5y^2$ gives

$$6 = (a^2 + 5b^2)(c^2 + 5d^2).$$

Since both factors on the right are nonnegative integers, the only possibilities are

$$a^2 + 5b^2 = 1, c^2 + 5d^2 = 6 \text{ and } a^2 + 5b^2 = 2, c^2 + 5d^2 = 3.$$

The second is not possible since $a^2 + 5b^2$ can never equal 2, and in the first we must have $a = \pm 1$ and $b = 0$, so that $a + b\sqrt{-5} = \pm 1$ is a unit. Thus any factorization of $1 + \sqrt{-5}$ must involve a unit, so $1 + \sqrt{-5}$ is irreducible. (The same argument also shows that $1 - \sqrt{-5}$ is irreducible.)

To see that $1 + \sqrt{-5}$ is not prime, note, as last time, that

$$(1 + \sqrt{-5})(1 - \sqrt{-5}) = 6 = 2 \cdot 3.$$

Thus $1 + \sqrt{-5}$ divides $2 \cdot 3$, but it does not divide either 2 nor 3 alone precisely because these are each irreducible. For instance, in order to have $3 = (1 + \sqrt{-5})(a + b\sqrt{-5})$, it would necessarily follow that $a + b\sqrt{-5}$ must be a unit since 3 is irreducible, but the only units in $\mathbb{Z}[\sqrt{-5}]$ are ± 1, and neither satisfies $3 = (1 + \sqrt{-5})(\pm 1)$. Hence $1 + \sqrt{-5}$ (also $1 - \sqrt{-5}$ for the same reason) is not prime in $\mathbb{Z}[\sqrt{-5}]$.

Warm-Up 2. We show that in an integral domain R, any nonzero prime is always irreducible. Suppose p is nonzero and prime, and suppose

$$p = ab \text{ for some } a, b \in R.$$

Since p is prime and p divides itself, it must divide either a or b above—say it divides a. Then $a = kp$ for some $k \in R$, so that

$$p = ab = kpb.$$

Since R is an integral domain (cancellation), this implies $1 = kb$, so b is a unit and hence p is irreducible.

Irreducibles are primes in PIDs. We have seen that irreducibles are not always prime in a general integral domain, but now we show that indeed they are in a PID. Actually, what we really show is that irreducibles generate maximal ideals in a PID, which is the key takeaway.

Suppose R is a PID and $x \in R$ is irreducible. Take an ideal $M \subseteq R$ such that

$$(x) \subseteq M \subseteq R$$

and write it as $M = (y)$ using the PID condition. Then $x \in (y)$, so

$$x = y\ell \text{ for some } \ell \in R.$$

Since x is irreducible, we must have that either y or ℓ is a unit. If y is a unit, then $M = (y) = R$, whereas if ℓ is a unit, then x and y are associates so $(x) = (y)$. Thus $(x) \subseteq M \subseteq R$ implies $(x) = M$ or $M = R$, so (x) is maximal. Since maximal ideals are prime, we get that (x) is prime, so x is prime as claimed.

Again, the actual key point in this argument is that if x is irreducible in a PID, then (x) is maximal. We will use this in a few days to characterize maximal ideals in polynomial rings for instance. The converse is also true: if (x) is maximal, then x is irreducible since maximal implies prime implies irreducible in any integral domain.

Unique factorization domains. We can now state the definition we've been building up to, of a ring where a version of "unique factorization" holds. A *unique factorization domain* (UFD for short—who would have guessed!?) is an integral domain R such that:

- any nonzero nonunit $r \in R$ can be written as a product of irreducible elements: there exist irreducible $p_1, \dots, p_n \in R$ such that $r = p_1 \cdots p_n$, and
- such a factorization is unique up to associates and rearrangements: if $p_1 \cdots p_n = q_1 \cdots q_m$ where each p_i, q_j is irreducible, then $m = n$ and each p_i is associate to some q_j.

So, "prime" factorizations exist (we will see next time that primes and irreducibles are indeed the same in any UFD, so that the use of "prime factorization" in this context is appropriate), and are unique apart from the order in which the irreducibles are written and putting in some units here and there. For instance, $6 = 2 \cdot 3 = (-3)(-2)$ does not violate the uniqueness requirement in $\mathbb{Z}$ since -3 is associate to 3 and -2 is associate to 2.

Examples will be plentiful, since any Euclidean domain is a PID, and any PID (as we will show) is a UFD. So, for instance, $\mathbb{Z}, F[x]$ where F is a field, and $\mathbb{Z}[i]$ are all UFDs. Also, $\mathbb{Z}[x]$ is a UFD, so this is an example of a UFD which is not a PID. The standard example of an integral domain which is not a UFD is our favorite $\mathbb{Z}[\sqrt{-5}]$, since

$$6 = 2 \cdot 3 = (1 + \sqrt{-5})(1 - \sqrt{-5})$$

gives two distinct irreducible factorizations in this ring. (Neither factor in one is associate to either factor in the other, so these are truly "distinct" factorizations.)

Fundamental Theorem of Arithemtic. The Fundamental Theorem of Arithmetic is the claim that any positive integer has a unique prime factorization, or in other words that $\mathbb{Z}$ is a UFD. (Of course, we can factor $n \in \mathbb{N}$ and then use the unit -1 to get a factorization of $-n \in \mathbb{Z}$, so that $\mathbb{Z}$ being a UFD is indeed equivalent to the usual claim about positive integers alone.) This result will of course be a consequence of the general fact we'll prove that PIDs are always UFDs, but recalling the ideas behind the proof of the Fundamental Theorem of Arithmetic will help shed light on the proof of this more general fact.

The basic idea is to take $n \in \mathbb{N}$, and factor it as much as possible until we can't anymore. The uniqueness of the factorization will then come from the prime property $p \mid ab \implies p \mid$ or $p \mid b$, as we'll see. But, the key question is: how do we know that the process of "factoring" has to eventually terminate? If n is not already prime/irreducible, then we can write it as $n = ab$ for some nonunits $a, b \in \mathbb{N}$. Then, if a is a prime we leave it alone, and if not we can write it as $a = cd$ (similarly for b instead), so that we have $n = cdb$ so far. And so on and so on, we continue in this way. The point is that at each step the positive integer factors we find are strictly smaller than what we had before, so the process has to stop since there is a limitation on how small positive integers can be! The way to make this truly precise is to use induction: if n is prime, we're done, and if not we write it as $n = ab$ with $a, b < n$, and then by induction we may assume that each of a and b has a prime factorization, which then gives a prime factorization for n. (Induction is, after all, equivalent to the fact that any nonempty subset of $\mathbb{N}$ has a smallest element.)

The issue is that "induction" is not something which exists in arbitrary rings or integral domains, since there is no notion of "less than" in general. So, in order to make the type of argument we used above work in more generality, we need to rephrase the process involved in a more ring-theoretic way. But, we do have such a way, using containments between principal ideals: to say that $n = ab$ above is the same as saying $(n) \subseteq (a)$, or $(n) \subseteq (b)$. At the next stage, if a is not prime, we can factor it as $a = cd$, so that $(a) \subseteq (c)$, which altogether gives

$$(n) \subseteq (a) \subseteq (c).$$

Then we keep going: factoring each factor we have so far into more and more terms corresponds to extending this "chain" of principal ideals further and further:

$$(n) \subseteq (a) \subseteq (c) \subseteq (t) \subseteq \dots$$

to say that the process of "factoring" eventually has to stop is then to say that such an "ascending" chain of ideals must eventually stop as well. Thus, in a ring for which such a property holds, we can expect a similar "every nonzero nonunit has a factorization into irreducibles" to hold as well.

Noetherian rings. The resulting type of ring is so widely used that it is given its own name. A *Noetherian ring* is a commutative ring R which satisfies what's called the *ascending chain condition*:

> If $I_1 \subseteq I_2 \subseteq I_3 \subseteq \dots$ is an ascending chain of ideals in R, there exists $k \geq 1$ such that $I_k = I_\ell$ for all $\ell > k$.

In other words, every ascending chain of ideals *terminates*, or *stablizes*, in the sense that all ideals become the same after some point. In the case of $\mathbb{Z}$ above, once our chain stabilizes $(t) = (s) = \cdots$, we have that s and t are associates, so that we are no longer able to factor a positive integer like t into positive nonunits anymore, hence why our "keep factoring" process ends. (It is possible to

state the Noetherian condition for non-commutative rings as well, but in our course we will only consider it in the commutative case, so we include "commutative" as part of the definition.)

We will show shortly that any PID is Noetherian, which is the key observation needed in showing that any PID is a UFD. The Noetherian condition seems like a strange thing to pull out of thin air at first, but the overarching idea is that it should be viewed as an analog of "induction" for more general types of rings. Indeed, in this course we are introducing this notion now, not only so that we can discuss unique factorization, but in some sense more importantly to set the stage for later when such a Noetherian condition will play a crucial role in many arguments involving modules. From my point of view, one big reason why it is worth discussing unique factorization at all is precisely so that we can use it to motivate the ascending chain condition.

The name "Noetherian" comes from the name Emmy Noether, a German mathematician of the early 20th century. Noether did groundbreaking foundational work in algebra and number theory, and indeed was one of the first people who gave the modern definition of "ideal" as we know it today. Noether also did fundamental work in physics, where *Noether's Theorem* establishes an important link between "conserved quantities" and "symmetries", whatever that means. It is no exaggeration to say that Emmy Noether was quite possibly the important and influential female mathematician of all time. Check Wikipedia to learn more!

PIDs are Noetherian. And now we show that any PID is Noetherian, which, as discussed earlier, will be the key step in showing that any PID is a UFD. So, suppose R is a PID and that

$$I_1 \subseteq I_2 \subseteq I_3 \subseteq \dots$$

is an ascending chain of ideals in R. Then the union $\bigcup_i I_i$ is also an ideal of R. (We saw this in an earlier use of Zorn's Lemma, when showing that any proper ideal of a ring was contained in a maximal ideal.) Since R is a PID, we have $\bigcup_i I_i = (x)$ for some $x \in \bigcup_i I_i$. But this means that $x \in I_k$ for some k, which then implies that $\bigcup_i I_i = (x) \subseteq I_k$, so that $I_k = I_{k+1} = I_{k+2} = \dots$ etc. (In other words, everything in $\bigcup_i I_i$ is a multiple of $x \in I_k$, so every ideal beyond I_k is contained in I_k itself.) Thus the chain terminates, so R is Noetherian.

Lecture 16: More on UFDs

Warm-Up 1. We show that in any UFD R, irreducible elements are always prime. (We previously saw this for PIDs.) Suppose $x \in R$ is irreducible and suppose $x \mid ab$, so that $ab = xk$ for some $k \in R$. Since R is a UFD, a and b have unique factorizations into irreducibles:

$$a = p_1 \cdots p_n \text{ and } b = q_1 \cdots q_m$$

where the p_i, q_j are irreducible. Then

$$xk = ab = p_1 \cdots p_n q_1 \cdots q_m.$$

Expressing k as a product of irreducibles would give a factorization of xk into irreducibles, so by uniqueness of such factorizations the irreducibles showing up in $xk = p_1 \cdots p_n q_1 \cdots q_m$ must be the same up to associates. Thus x must be associate to one of the irreducibles appearing in $p_1 \cdots p_n q_1 \cdots q_m$; without loss of generality say that x is associate to p_1. Then $p_1 = ux$ for some unit $u \in R$, so

$$a = p_1 \cdots p_n = u x p_2 \cdots p_n,$$

which implies that $x \mid a$, so x is prime. (As a consequence, we can speak of uniqueness of "prime factorizations"—not just "irreducible"—in a UFD. The ability to compare factors on one side of

an expression to factors on the other as we did above is the practical benefit to having unique factorization. In a non-UFD, such comparisons are not possible, as we have seen for $2 \cdot 3 = (1+\sqrt{-5})(1-\sqrt{-5})$ in $\mathbb{Z}[\sqrt{-5}]$ for instance.

Warm-Up 2. We show that R (commutative) is Noetherian if and only if every ideal of R is finitely generated. The backwards direction is very much like the proof we gave last time that PIDs are Noetherian. Indeed, suppose every ideal of R is finitely generated, and let

$$I_1 \subseteq I_2 \subseteq I_3 \subseteq \dots$$

be an ascending chain of ideals. Then $\bigcup_k I_k$ is an ideal of R, so it is finitely generated, say $\bigcup_k I_k = (x_1, \dots, x_n)$. Each x_i is in some I_{k_i}, so by the chain property all x_i are in I_m where $m = \max\{k_1, \dots, k_n\}$. But this means that $I_m = (x_1, \dots, x_n) = \bigcup_k I_k$, so the chain terminates/stabilizes at I_m, and hence R is Noetherian.

Conversely, suppose there were an ideal I which was not finitely generated. Pick a nonzero $a_1 \in I$. Then a_1 does not generate I, so there exists $a_2 \in I - (a_1)$. But (a_1, a_2) also is not I, so there exists $a_3 \in I - (a_1, a_2)$. And so on, if we have found $(a_1, \dots, a_k)$ which is not I, we can pick a_{k+1} in I but not $(a_1, \dots, a_k)$. This gives an ascending chain of ideals

$$(a_1) \subseteq (a_1, a_2) \subseteq (a_1, a_2, a_3) \subseteq \dots$$

which does not terminate since each contains an element the prior one does not. Thus R cannot be Noetherian in this case, and we are done.

To highlight another way of thinking about the Noetherian condition, depending on what type of linear algebra course you've taken before, you might have seen a fact similar to that above there, namely the fact that a vector space V is infinite-dimensional if and only if there exists an infinite collection of linearly independent vectors:

$$v_1,\ v_2,\ v_3,\ \dots.$$

The proof works by picking, in an infinite-dimensional space, at each step a vector not contained in the span of the previous ones, which is similar to the construction of $a_1, a_2, a_3, \dots$ above. Indeed, the point is that the Noetherian property can be viewed as an analog of "finite-dimensional" in the setting of rings in some sense.

PIDs are UFDs. Now we finally prove that any PID is a UFD. So, let R be a PID, which is hence Noetherian. First, to show existence of factorizations into irreducibles, suppose $x \in R$ was a nonzero nonunit which could not be written as a product of irreducible elements. In particular then, x must be reducible itself, so $x = a_1 b_1$ with $a_1, b_1 \in R$ not units. At least one of a_1 or b_1 is then also reducible since otherwise $x = a_1 b_1$ would be a product of irreducibles, so without loss of generality suppose a_1 is reducible. Then $a_1 = a_2 b_2$ for some nonunits $a_2, b_2 \in R$. Again, at least one of these is reducible, so say it is a_2. Then $a_2 = a_3 b_3$ for nonunits $a_3, b_3 \in R$, and so on. Continuing in this way (which is possible to do since x is not a product of irreducibles) produces an ascending chain of ideals:

$$(x) \subseteq (a_1) \subseteq (a_2) \subseteq (a_3) \subseteq \dots.$$

Each containment here is strict since, by construction, no one generator is associate to the next: if a_{i+1} were associate to $a_i = a_{i+1} b_{i+1}$, we would have $a_{i+1} b_{i+1} = a_{i+1} u$ with u a unit, which would imply by cancellation that $b_{i+1} = u$ is a unit. Thus this chain never terminates, contradicting the Noetherian condition. Hence every nonzero nonunit of R is a product of irreducibles.

To verify uniqueness, suppose

$$p_1 \cdots p_k = q_1 \cdots q_\ell$$

where $k \leq \ell$ and the $p_i, q_j \in R$ are all irreducible. Since irreducibles in a PID are always prime and p_1 divides the left above, p_1 must divide some q_j, and after rearranging we may assume $p_1 \mid q_1$. Then $q_1 = u_1 p_1$ with $u_1 \in R$ a unit. Substituting this in for q_1 above and cancelling p_1 from both sides gives:

$$p_2 \cdots p_k = u_1 q_2 \cdots q_\ell.$$

Now we can repeat this process for p_2, then p_3, and so on until we get

$$1 = u_1 \cdots u_k q_{\ell-k} \cdots q_\ell.$$

But if $k < \ell$ this would imply that q_ℓ (leftover on the right side above) is a unit, so it would not have been irreducible. Thus we must have $k = \ell$ and each p_i is associate to some q_j as claimed. (More formally, we can induct on the number of p's showing up above: once we get $p_2 \cdots p_k = u_1 q_2 \ldots q_\ell$, by induction we may assume that the number of irreducibles here are the same, so $k = \ell$, and that each remaining p_i is associate to some remaining q_j as required.)

Primes in the Gaussian integers. We finish by giving one application of unique factorization, which recovers a classical theorem of Fermat on writing primes as sums of squares. This comes down to studying the behavior of primes in $\mathbb{Z}[i]$, and more concretely to determining when a prime integer p factors as an element of $\mathbb{Z}[i]$. We note first that $\pm 1, \pm i$ are the units of $\mathbb{Z}[i]$, and that these are the only elements of norm 1 with respect to $N(a+ib) = a^2 + b^2$.

Consider a prime/irreducible $\pi \in \mathbb{Z}[i]$. Then $(\pi) \cap \mathbb{Z}$ is a prime ideal of $\mathbb{Z}$, so $(\pi) \cap \mathbb{Z} = p\mathbb{Z}$ for some prime $p \in \mathbb{Z}$. (Note that this intersection is not the zero ideal, since it contains $\pi\overline{\pi} > 0$.) Then $p \in (\pi)$, so $p = \pi\pi'$ for some $\pi' \in \mathbb{Z}[i]$. Taking norms gives

$$p^2 = N(\pi)N(\pi').$$

Since all terms here are now positive integers, the only possibilities are

$$N(\pi) = p^2, N(\pi') = 1 \text{ and } N(\pi) = p, N(\pi') = p.$$

(Note this uses the fact that $\mathbb{Z}$ is a UFD.) In the first case, $N(\pi') = 1$ implies that π' is a unit, so $p = \pi\pi'$ is associate to π and hence p is irreducible in $\mathbb{Z}[i]$. In the second case, the fact that $N(\pi')$ is a prime integer implies that π' is irreducible: if $\pi' = zw$, then $N(\pi') = N(z)N(w)$ is plus/minus a prime, so one of $N(z), N(w)$ is 1, meaning that one of z, w is a unit. Thus in this second case, p is reducible and $p = \pi\pi'$ is its (unique) prime/irreducible factorization in $\mathbb{Z}[i]$. The conclusion is that a prime integer either remains prime in $\mathbb{Z}[i]$ or factors into two irreducibles, which moreover are necessarily of the form $\pi = a+ib$, $\pi' = \overline{\pi} = a-ib$: if $\pi\pi' = p$ is positive and real, then $\pi\pi' = \lambda(\pi\overline{\pi})$ for some $\lambda > 0$ since $\pi\overline{\pi}$ is a positive real number, which implies $\pi' = \lambda\pi$, so that $N(\lambda) = 1$ since π' and $\overline{\pi}$ both have norm p, and hence $\lambda = 1$ and $\pi' = \overline{\pi}$.

Thus when p is reducible in $\mathbb{Z}[i]$, we have $p = (a+ib)(a-ib) = a^2 + b^2$. Fermat's theorem (not his "last" theorem, just *a* theorem of Fermat) characterizes such primes which can be written as a sum of two squares:

An odd prime $p \in \mathbb{N}$ is a sum of two squares if and only if $p \equiv 1 \bmod 4$.

(Note $2 = 1^2 + 1^2$ is a sum of squares, so what happens for odd primes is the true interesting question.) One direction is easy without anything fancy: if $p = a^2 + b^2$ holds, then by considering

the different cases where a, b are even or odd, we see that either $p \equiv 0 \bmod 4$ (when a and b are both even), or $p \equiv 1 \bmod 4$ (otherwise). The $p \equiv 0 \bmod 4$ case requires that p be even, so if p is odd we must have $p \equiv 1 \bmod 4$. The backwards direction, that $p \equiv 1 \bmod 4$ is indeed enough to guarantee that p is a sum of two squares (or equivalently, reducible in $\mathbb{Z}[i]$) is the one which requires using the structure of $\mathbb{Z}[i]$ as described above. We will finish this up next time. One thing to note is that for odd primes p which are congruent to 3 mod 4, this all shows that it is not possible to express p as a sum of two squares, so such primes remain prime/irreducible in $\mathbb{Z}[i]$.

Lecture 17: Polynomials Revisited

Warm-Up. We show that if $p \in \mathbb{N}$ is a prime which is congruent to 1 mod 4, then there exists $n \in (\mathbb{Z}/p\mathbb{Z})^\times$ such that $n^2 \equiv -1 \bmod p$, which is the crucial step needed to finish off the "sums of squares" result we stated last time. For the most part, the proof we give is mainly group-theoretic, although the integral domain structure of $\mathbb{Z}/p\mathbb{Z}$ does come into play. We will use the fact that $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic, which is something we stated at various points last quarter and will soon be able to finally prove using results about polynomials over fields.

If $p \equiv 1 \bmod 4$, then 4 divides $p-1$, which is the order of the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$. Hence since this group is cyclic, there exists $n \in (\mathbb{Z}/p\mathbb{Z})^\times$ of order 4: $n^4 \equiv 1 \bmod p$. But then $x = n^2$ satisfies $x^2 \equiv 1 \bmod p$, so either $x \equiv 1 \bmod p$ or $x \equiv -1 \bmod p$. (This is where the fact that $\mathbb{Z}/p\mathbb{Z}$ is an integral domain comes in: if $x^2 = 1$, then $0 = x^2 - 1 = (x-1)(x+1)$, so that either $x - 1 = 0$ or $x + 1 = 0$.) But $n^2 = x \equiv 1 \bmod p$ is not possible since n has order 4, so we must have $n^2 \equiv -1 \bmod p$ as claimed.

Back to sums of squares. Thus when $p \equiv 1 \bmod 4$, some $n^2 + 1$ is divisible by p. But in $\mathbb{Z}[i]$ this means that p divides

$$n^2 + 1 = (n+i)(n-i).$$

If p were irreducible, it would be prime since $\mathbb{Z}[i]$ is a PID/UFD, so it would divide one of $n+i$ or $n-i$. But in fact it would then have to divide both: if $n \pm i = pz$ forsome $z \in \mathbb{Z}[i]$, then taking conjugates gives $n \mp i = p\overline{z}$ since p is real. Thus p would also divide $(n+i) - (n-i) = 2i$, which is not true since no $c + id \in \mathbb{Z}[i]$ satisfies $2i = p(c+id)$. Hence p must actually be reducible, which, as explained last time, implies that $p = (a+ib)(a-ib) = a^2 + b^2$ for some $a, b \in \mathbb{Z}$.

The upshot is that being able (or not) to write a prime as a sum of squares comes down to understanding the structure of the ring $\mathbb{Z}[i]$. This same type of argument applies to other similar "quadratic integer rings" as well, at least the ones which are UFDs. For instance, $\mathbb{Z}[\sqrt{2}]$ is a Euclidean domain with respect to the norm $N(a + b\sqrt{2}) = |a^2 - 2b^2|$, so it is a UFD as well. Some analysis along the lines of that above then leads to a characterization of primes which can be written as $a^2 - 2b^2$, with the answer being that an odd prime $p \in \mathbb{Z}$ is of this form if and only if $p \equiv 1 \bmod 8$ or $p \equiv 7 \bmod 8$. (In other words, odd primes congruent to 3 or 5 mod p remain irreducible in $\mathbb{Z}[\sqrt{2}]$, and those congruent to 1 or 7 are reducible. Note $2 = 2^2 - 2(1)^2$ is of the desired form.) The key step needed, analogous to the Warm-up above, is that if p is congruent to 1 or 7 mod 8, then there exists n such that $n^2 = 2 \bmod p$, which is slightly beyond our ability to prove easily in this course. A course in number theory would be able to say more.

Fermat's Last Theorem. Before we leave the world of UFDs and number-theoretic results, let us revisit a motivation we provided back on the first day of class, namely coming from attempts to prove Fermat's Last Theorem in the mid to late 1800's. The basic idea was to take the desired equality in Fermat's Last Theorem:

$$x^n + y^n = z^n$$

and rewrite it as

$$x^n = z^n - y^n = (z-y)(\text{product of some other factors}).$$

For instance, in the $n = 3$ case, the desired factorization is:

$$x^3 = z^3 - y^3 = (z-y)(z^2+zy+y^2) = (z-y)(z-\zeta y)(z-\zeta^2 y)$$

where $\zeta := e^{2\pi i/3}$ is a *primitive* cube root of unity. Indeed, we have $\zeta^3 = 1$ and so

$$1 + \zeta + \zeta^2 = \frac{1-\zeta^3}{1-\zeta} = \frac{0}{1-\zeta} = 0,$$

which implies that $1 = -\zeta - \zeta^2$. Hence the product $(z-\zeta y)(z-\zeta^2 y)$ does equal $z^2 + zy + y^2$:

$$(z-\zeta y)(z-\zeta^2 y) = z^2 + (-\zeta - \zeta^2)zy + \zeta^3 y^2 = z^2 + zy + y^2.$$

The factorization above takes place in the ring $\mathbb{Z}[e^{2\pi i/3}]$, which is called the ring of *Eisenstein integers*. (The Gaussian integers are obtained by adjoining to $\mathbb{Z}$ a primitive fourth root of unity, namely i. We will discuss these and other roots of unity next quarter in the context of *cyclotomic fields*, although we probably won't say much more about Fermat's Last Theorem then.)

Then, by comparing the primes showing up in the prime factorization of x^n with those showing up on the right for $z^n - y^n$, an argument could be made that no nontrivial x, y, z could in fact exist. Essentially, for each prime p dividing x, p^n divides x^n, so p^n (or the n-th power of irreducible factors of p if p is reducible in $\mathbb{Z}[\zeta]$) should show up in the prime factorizations of the factors making up $z^n - y^n$, and one can find a way to argue that this is not possible except for in the trivial cases. Such attempts at proving Fermat's Last Theorem, however, depend on the ability to compare prime factors on one side with those on the other, and hence are only valid when working over a UFD! The ring of Eisenstein integers $\mathbb{Z}[e^{2\pi i/3}]$ is in fact a UFD (it is actually a Euclidean domain), so carrying out this analysis does lead to a valid proof of Fermat's Last Theorem in the $n = 3$ case. Similarly, a proof in the $n = 4$ case can be found using $\mathbb{Z}[i]$, but there are many n for which this approach is not valid. Understanding when unique factorization does and doesn't hold necessitated the development of more ring theory.

Dedekind domains. So, not all rings showing up in number-theoretic contexts are UFDs, which makes certain questions more difficult than others. Still, it turns out that a type of "unique factorization" *can* be recovered in such rings, if we shift our focus sway from *elements* to *ideals* instead. That is, in such number-theoretic rings (more precisely, in the *rings of integers* of *number fields*), it is the case that any ideal can be written as a product of *prime* ideals in a unique way. For instance, we have seen that $\mathbb{Z}[\sqrt{-5}]$ is not a UFD, since

$$6 = 2 \cdot 3 = (1+\sqrt{-5})(1-\sqrt{-5})$$

gives two different irreducible factorizations of 6. But, if we interpret these equalities as equalities of the ideals generated by these elements instead:

$$(6) = (2)(3) = (1+\sqrt{-5})(1-\sqrt{-5}),$$

we do end up getting the same prime factorization of ideals: there exist prime ideals I_2, I_3, and I_3' (these are described explicitly on the homwework) such that both $(2)(3)$ and $(1+\sqrt{-5})(1-\sqrt{-5})$ lead the same factorization of the ideal (6):

$$(6) = I_2^2 I_3 I_3',$$

and this factorization is unique.

So, all is not lost, and it turns out that much of what we expect about "numbers" and "primes" holds at the level of ideals instead, which is often good enough for many applications in number theory. These types of rings, namely ones where ideals can be factored uniquely into product of prime ideals, are called *Dedekind domains*, and are of great important in advanced number theory. (To be sure, the modern definition of "ideal" we are using in this course was not developed until the 1920's or so, so back in the 1800's people spoke of "ideal numbers" instead and phrased this unique prime ideal factorization property in terms of these. But, it essentially amounts to the same thing.) Dedekind domains are also intimately related to various other types of rings we've seen—such as Noetherian and discrete valuation rings—but alas we will say no more about them in this course.

Back to polynomials. Now we move to studying another special type of ring—polynomial rings—in more detail. We have already seen many properties of these so far, so let us recap some things we know:

- if R is an integral domain, then $R[x]$ is an integral domain;
- if F is a field, then $F[x]$ is a Euclidean domain (so the division algorithm exists), hence a PID, hence a UFD (note $\mathbb{Z}[x]$ is a UFD which is not a PID); and
- if F is a field and $p(x) \in F[x]$, then $a \in F$ is a root of $p(x)$ if and only if $x - a$ divides $p(x)$.

The final property is one we proved after discussing the division algorithm for polynomials, and comes from writing $p(x) = (x-a)q(x)+r(x)$ with $\deg r(x) < \deg (x-a)$ and showing that $p(a) = 0$ if and only if $r(x) = 0$.

Let us now see one consequence of this third property, which justifies a property of polynomials you might expect going by what you know about $\mathbb{R}$ or $\mathbb{C}$. If F is a field and $p(x) \in F[x]$ is of degree $n \geq 1$, we claim that $p(x)$ has at most n roots, so that the degree restricts the number of roots. (We are making no claim here that roots exist at all, just that if they do, their number is controlled by the degree. We will talk about the types of fields were roots are guaranteed to exist—ones which are *algebraically closed*—next quarter.) Note that this fact would not be true if we working with polynomials over a non-integral domain, where there can be more roots than the degree. For instance, over $\mathbb{Z}/6\mathbb{Z}$, $2x$ has degree 1 but two roots, namely 0 and 3, and as another example $3x$ also degree 1 but now three roots: $0, 2$, and 4.

So, suppose $r_1, \ldots, r_m$ are the roots of $p(x) \in F[x]$, where F is a field. (Note these do not have to be distinct, so that this list takes into account the "multiplicity" of roots.) Then since r_1 is a root, $x - r_1$ divides $p(x)$ so we have

$$p(x) = (x - r_1)q_1(x)$$

for some $q_1(x) \in F[x]$. But then r_2 is a root of $p(x)$ so it must be a root of $q_1(x)$ (this depends on the lack of zero divisors, and works even if $r_2 = r_1$), and hence $q_1(x) = (x - r_2)q_2(x)$ for some $q_2(x)$, which gives

$$p(x) = (x - r_1)(x - r_2)q_2(x).$$

And so on, by continuing to factor the remaining $q_i(x)$ at each step we find that $p(x)$ is of the form

$$p(x) = (x - r_1)(x - r_2) \cdots (x - r_m)q_m(x).$$

Thus, $(x - r_1)(x - r_2) \cdots (x - r_m)$ divides $p(x)$, so it must have degree at most that of $p(x)$, which gives $m \leq \deg p(x)$ as claimed. (We will see one application of this result in a few days to justify the fact that $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic when p is prime.)

Since any integral domain R is contained in a field (say its fraction field F), the same result is true for polynomials over R as well: $p(x) \in R[x]$ is also in $F[x]$, so the number of roots in F is at most $\deg p(x)$, and hence the number of roots in $R \subseteq F$ is bounded by the same. We will see similar uses of fraction fields to relate polynomials over integral domain to those over fields soon.

Maximal polynomial ideals. If F is a field, then $F[x]$ is a PID, so not only is every ideal generated by a single polynomial, but we also know which elements generate the maximal ideals: $(p(x))$ is maximal if and only if $p(x)$ is irreducible. Said another way, $F[x]/(p(x))$ is a field if and only if $p(x)$ is irreducible. Constructing fields in this way will be a useful tool for us, in particular next quarter, so it will be useful to spend some time discussing irreducibility of polynomials.

We will develop various tests for irreducibility in the coming days, but here are some quick observations. First, since the only units in $F[x]$ are the nonzero constants, any polynomial of degree 1 is irreducible and polynomials of degree larger than 1 will be reducible if and only if they can be written as a product of polynomials of strictly smaller degree. As a consequence, we have an easy way to determine if a polynomial of degree 2 or 3 is reducible: if $\deg p(x) = 2$ or 3, then being reducible $p(x) = q(x)h(x)$ would require having a factor of degree 1 (since $\deg q(x) + \deg h(x)$ would have to be 2 or 3), but a factor of degree 1 looks like $x - a$ (if instead you have something like $cx - d$, you can always factor out c and incorporate it into the other factor instead), and having such a factor is equivalent to having a root $a \in F$. Thus, a polynomial of degree 2 or 3 over a field is reducible if and only if it has a root in that field. For polynomials of degree larger 3, however, we'll have to work harder; for instance, $x^4 + 2x^2 + 1 = (x^2 + 1)(x^2 + 1)$ is reducible over $\mathbb{R}$ even though it has no root in $\mathbb{R}$.

Lecture 18: Gauss's Lemma

Warm-Up. We show that for any prime p there exists a field of order p^2. The fields we want will be constructed as quotients of $(\mathbb{Z}/p\mathbb{Z})[x]$, so what we need is an irreducible polynomial of degree 2 in $(\mathbb{Z}/p\mathbb{Z})[x]$ for each prime. Indeed, if $p(x) = ax^2 + bx + c \in (\mathbb{Z}/p\mathbb{Z})[x]$ is such a polynomial, then $(p(x))$ is maximal so that $(\mathbb{Z}/p\mathbb{Z})[x]/(p(x))$ is a field, and the relation

$$x^2 = a^{-1}(-bx - c)$$

in the quotient will give a way to rewrite any element in the quotient as a single degree 1 expression $b_0 + b_1 x$; since there are p choices for each of b_0 and b_1, this will give p^2 many elements in $(\mathbb{Z}/p\mathbb{Z})[x]/(p(x))$ as desired.

In order for the degree 2 polynomial $p(x)$ we're looking for to be irreducible, it is equivalent that it not have a root in $\mathbb{Z}/p\mathbb{Z}$. For $p = 2$ for instance, $x^2 + x + 1$ has no root in $\mathbb{Z}/2\mathbb{Z}$, so it is irreducible and hence $(\mathbb{Z}/2\mathbb{Z})[x]/(x^2 + x + 1)$ is a field of order $2^2 = 4$. For $p > 2$, note that $x^2 = (-x)^2$ for each nonzero $x \in \mathbb{Z}/p\mathbb{Z}$, so that the nonzero elements get "paired off" when forming squares. This implies that there exists elements in $\mathbb{Z}/p\mathbb{Z}$ which are not squares, and in fact there are $\frac{p-1}{2}$ elements which are not squares and $\frac{p-1}{2}$ which are. (Another way to say this is the squaring map $x \mapsto x^2$ is not injective as a map $\mathbb{Z}/p\mathbb{Z} \to \mathbb{Z}/p\mathbb{Z}$, so it cannot be surjective since both the domain and codomain have the same size.) Thus, if $a \in \mathbb{Z}/p\mathbb{Z}$ is an element which is not a square (i.e. does not arise as any x^2 for $x \in \mathbb{Z}/p\mathbb{Z}$)), then $x^2 - a \in (\mathbb{Z}/p\mathbb{Z})[x]$ has no root in $\mathbb{Z}/p\mathbb{Z}$, so it is irreducible and thus $(\mathbb{Z}/p\mathbb{Z})[x]/(x^2 - a)$ is a field of order $2^2 = 4$. (Note the $p = 2$ case had to be handled separately since $x = -x$ in this case and every element of $\mathbb{Z}/2\mathbb{Z}$ *is* a square. Indeed, $x^2 + 1 = (x + 1)^2$ is reducible over $\mathbb{Z}/2\mathbb{Z}$.)

Finite fields. In fact, we will show next quarter that *any* finite field must have prime-power order: if F is a finite field, then $|F| = p^n$ for some prime p and $n \geq 1$. Moreover, such fields indeed exist and are unique: for any prime p and $n \geq 1$, there exists a field of order p^n which is unique up to isomorphism. Thus, we will essentially know all there is no know about finite fields. The unique field of order p^n is usually denoted by $\mathbb{F}_{p^n}$, and is often called the *Galois field* of order p^n. For instance, $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$, $\mathbb{F}_4 = \mathbb{F}_2[x]/(x^2 + x + 1)$, and $\mathbb{F}_{p^2} = \mathbb{F}_p[x]/(x^2 - a)$ where $a \in \mathbb{F}_p$ has no root, as constructed above.

The fact that any finite field has prime-power order will make use of linear algebra over fields of finite characteristic (the notion of characteristic was defined on the first homework, but we will review and consider it more carefully next quarter), and the construction of such fields will proceed in the same way as in the Warm-Up, by obtaining them as quotients of polynomial rings. The construction for $n \geq 3$ is not as straightforward as above since it is more challenging to prove that there exists an irreducible polynomial of degree n in $\mathbb{F}_p[x]$ for any $n \geq 3$. We will do it next quarter using the concepts of *algebraic closures*, *simple extensions*, and *minimal polynomials*. Good stuff!

Finite subgroups of $F^\times$. While we're still in the "finite" frame of mind, let us finally justify the fact that, for p prime, the multiplicative group $(\mathbb{Z}/p\mathbb{Z})^\times$ is cyclic. In fact, more generally, any finite subgroup of the groups of units of *any* field is cyclic: if F is a field and $G \leq F^\times$ is a multiplicative subgroup of $\mathbb{F}^\times$, then G is cyclic. For instance, any finite subgroup of $\mathbb{Q}^\times$, or of $\mathbb{R}^\times$, is cyclic. The $(\mathbb{Z}/p\mathbb{Z})^\times$ case follows by taking this to be a finite subgroup of itself. The proof will make use of properties of polynomials over fields, which is why we are presenting it here.

We will make use of the final result we proved last quarter: any finite abelian group is a product of cyclic groups $\mathbb{Z}/d_1\mathbb{Z} \times \cdots \times \mathbb{Z}/d_k\mathbb{Z}$, which can be chosen so that $d_1 \mid d_2, d_2 \mid d_3, \ldots,$ and $d_{k-1} \mid d_k$. (We will prove the more general structure theorem of finitely generated abelian groups soon enough, but we did provide a proof in the finite case on the final day of class last quarter. Check the notes from then to see it.) In our case, G is finite and abelian, so

$$G \cong \mathbb{Z}/d_1\mathbb{Z} \times \cdots \times \mathbb{Z}/d_k\mathbb{Z}$$

with each d_i dividing d_{i+1}. (Note that elements on the right are written additively, but they correspond to elements of G on the left, which is a multiplicative group. In particular, the identity $1 \in G$ corresponds to $(0, 0, \ldots, 0)$.) The goal is to show that in fact there is only one cyclic group appearing on the right. Note that since the first factor has order d_1, any element inside of it, viewed as an element of G of the form

$$(x, 0, 0, \ldots, 0)$$

under the isomorphism above, has order dividing d_1 and is thus a root of the polynomial $x^{d_1} - 1$ in $F[x]$. Now, if the second factor $\mathbb{Z}/d_2\mathbb{Z}$ where in fact present, then since $d_1 \mid d_2$ and $\mathbb{Z}/d_2\mathbb{Z}$ is cyclic, this second factor has a subgroup of order d_1, giving elements of G of the form

$$(0, x, 0, \ldots, 0)$$

(under the isomorphism above) which *also* have order dividing d_1, and hence which are also roots of $x^{d_1} - 1$ in $F[x]$. Between these two factors we thus get $2d_1 - 1$ (there is one overlapping element: the identity of G which corresponds to $(0, 0, 0, \ldots, 0)$ in the product of cyclic groups) roots of $x^{d_1} - 1$ in $F[x]$, which is not possible since $x^{d_1} - 1$ can only have at most d_1 roots. (Note $2d_1 - 1 > d_1$ because $d_1 > 1$, since otherwise $\mathbb{Z}/d_1\mathbb{Z} = \mathbb{Z}/\mathbb{Z} = \{0\}$.) Thus, the isomorphism on the above only has one factor on the right, so $G \cong \mathbb{Z}/d_1\mathbb{Z}$ is cyclic as claimed.

Gauss's Lemma. Let us now return to studying irreducibility more carefully, which we will continue with next time as well. The first basic step is to relate reducibility in an integral domain

to reducibility in its fraction field, with $\mathbb{Z}[x]$ vs $\mathbb{Q}[x]$ being the main example. Actually, to make this work as nicely as possible, we will assume that R is not only an integral domain, but actually a UFD. If F denotes its field of fractions, then *Gauss's Lemma* states that if $p(x) \in R[x]$ is reducible in $F[x]$, then it is reducible in $R[x]$. The point is that factoring in $F[x]$ allows for coefficients that might not come strictly from R, but that nevertheless there is a way to modify the factorization over F to obtain one valid over R.

So, suppose $p(x) \in R[x]$ can be written as

$$p(x) = f(x)h(x) \text{ for some } f(x), h(x) \in F[x] \text{ of degree smaller than that of } p(x).$$

All the coefficients of $f(x)$ and $h(x)$ are fractions of elements of R, so by multiplying by common denominators we can turn each of $f(x)$ and $h(x)$ into polynomials in $R[x]$ instead, to get:

$$dp(x) = \tilde{f}(x)\tilde{h}(x) \text{ for some } d \in R, \text{ where } \tilde{f}(x), \tilde{h}(x) \in R[x].$$

Now, the idea is that we can "cancel" out d from both sides to be left with a factorization in $R[x]$. This is where unique factorization comes in: since R is a UFD, we can write $d = q_1 q_2 \dots q_k$ as a product of irreducible/prime elements q_i. (Actually, this is only true if $d \in R$ is not a unit, but it is a unit then we can immediately cancel it out by multiplying by its inverse, and then there is nothing more to show.) We claim that each q_i is then also a divisor of $\tilde{f}(x)$ or $\tilde{h}(x)$, which is what will allow us to cancel out each q_i from both sides of $dp(x) = \tilde{f}(x), \tilde{h}(x)$. (In other words, we need to know that each q_i remains prime, not only in R, but in $R[x]$ as well. This is not obvious, since dividing products of elements of R is not the same as dividing products of polynomials over R, where the latter in the end involves expressions with *sums* of products of elements of R. Note it *is* true that q_i being irreducible in R immediately implies it is irreducible in $R[x]$, but this does not help us here since "irreducible implies prime" would only hold if we already knew $R[x]$ was a UFD, which we do not—yet. In fact, the proof that "R UFD $\Rightarrow$ $R[x]$ UFD" *depends* on Gauss's Lemma, so we cannot rely on it when proving Gauss' Lemma.)

In general, if I is an ideal of R, then it generates the ideal $I[x]$ of $R[x]$. The homomorphism $R[x] \to (R/I)[x]$ given by reducing all coefficients mod I has $I[x]$ as its kernel, so the First Isomorphism Theorem gives $R[x]/I[x] \cong (R/I)[x]$, meaning that reducing polynomials mod $I[x]$ is the same as reducing the coefficients alone mod I. In particular, $R[x]/I[x]$ is an integral domain if and only if $(R/I)[x]$ is an integral domain, which is true if and only if R/I is an integral domain. Hence, an ideal in R is prime if and only if the ideal it generates in $R[x]$ is prime as well. The upshot is that, in our case, a prime element $q \in R$ does in fact remain prime in $R[x]$. So, take q_1 in the prime factorization of d, which is then prime in $R[x]$ too, so it must divide a factor on the right of

$$dp(x) = \tilde{f}(x)\tilde{h}(x).$$

Say that d_1 divides $\tilde{f}(x)$, so that $\tilde{f}(x) = d_1 g(x)$ for some $g(x) \in R[x]$. Then the equality above becomes

$$q_1 q_2 \cdots q_k(x) = q_1 g(x)\tilde{h}(x), \text{ so } q_2 \cdots q_k p(x) = g(x)\tilde{h}(x).$$

Repeating this process for each remaining q_i will then lead to a factorization

$$p(x) = a(x)b(x)$$

with $a(x), b(x) \in R[x]$, so that $p(x)$ is reducible over R as well. (Note that $a(x), b(x)$ have the same degrees as the original $f(x), h(x)$ respectively since "cancelling out" elements of R at each step will never affect the degrees.)

Converse to Gauss's Lemma. So, being reducible in $F[x]$ implies being reducible in $R[x]$, when R is a UFD at least. The converse seems at first like it should be immediate, but actually some care is needed: being reducible in $R[x]$ does not always imply being reducible in $R[x]$. The reason has to do with the fact that the there are nonunit polynomials of degree 0 in $R[x]$ (namely the nonunit constants) which become units in $F[x]$. For instance, the polynomial $3x + 3$ is actually *reducible* in $\mathbb{Z}[x]$ since $3x + 3 = 3(x + 1)$ is a product of non-units of $\mathbb{Z}[x]$, but of course $3x + 3$ is irreducible in $\mathbb{Q}[x]$, where 3 now being a unit avoids the issue we had in $\mathbb{Z}[x]$.

But, this is the only way in which reducibility in $R[x]$ vs $F[x]$ could differ. More concretely, if $p(x) = q(x)h(x) \in R[x]$ is reducible with neither $q(x)$ nor $h(x)$ being nonunit constants, then $p(x)$ remains reducible in $F[x]$. When $p(x) = dq(x)$ *is* a multiple of a nonunit constant $d \in R$, then it is reducible in $R[x]$ but we cannot say anything definitive yet about reducibility over F because $p(x) = dq(x)$ is not a valid factorization in $F[x]$ into nonunits. This latter case occurs precisely when the greatest common divisor of the coefficients of $p(x) \in R[x]$ is a nonunit $d \in R$, so the conclusion is that a *primitive* polynomial—one where the greatest common divisors of the coefficients is 1—is reducible in $R[x]$ if and only if it is reducible in $F[x]$. As a special case, for a polynomial which is *monic*—meaning it has leading (i.e. highest degree) coefficient 1—irreducibility in $R[x]$ is equivalent to irreducibility in $F[x]$, since if 1 is a coefficient then the greatest common divisor of the coefficients is certainly 1 as well.

Lecture 19: Eisenstein's Criterion

Warm-Up 1. We prove the *rational root theorem*: if R is a UFD with fraction field F and $\frac{a}{b} \in F$ is a root of $p(x) = c_m x^m + \cdots + c_1 x + c_0 \in R[x]$ with a, b relatively prime (meaning 1 is a greatest common divisor of a and b), then a divides c_0 and b divides c_m. Thus, the "rational roots" of a polynomial over R can be found among those fractions with numerator and denominator constrained by a divisibility requirement. Note this does *not* mean that all such $\frac{a}{b}$, with a dividing the constant term and b the leading coefficient, *will* be roots, only that they are the only *potential* rational roots. (Truth be told, I have only heard the name "rational root theorem" used in the special case where $R = \mathbb{Z}$ and $F = \mathbb{Q}$, but I'll go ahead and use this name for the general version.)

If $\frac{a}{b}$ is a root of $p(x)$, then:

$$c_m \frac{a^m}{b^m} + c_{m-1} \frac{a^{m-1}}{b^{m-1}} + \cdots + c_1 \frac{a}{b} + c_0 = 0.$$

Multiplying through by b^m to clear denominators gives:

$$c_m a^m + c_{m-1} a^{m-1} b + \cdots + c_1 a b^{m-1} + c_0 b^m = 0.$$

Now, since a divides each $c_i a^i b^{m-i}$ $(i > 0)$ on the left, it must divide $c_0 b^m$ as well. Since a and b are relatively prime, this implies that a divides c_0. (Since a and b are relatively prime, none of the prime divisors in the prime factorization of a show in up in the prime factorizations of b nor b^m, so if a divides $c_0 b^m$ it must be that all of these prime divisors show up instead in the prime factorization of c_0, which is why a divides c_0.) Similarly, since b divides each $c_i a^i b^{m-i}$ $(i < m)$ on the left of

$$c_m a^m + c_{m-1} a^{m-1} b + \cdots + c_1 a b^{m-1} + c_0 b^m = 0,$$

b must divide $c_m a^m$ as well, so b divides c_m since a and b are relatively prime.

Note there was a problem in the Discussion 4 Problems handout, dealing with the notion of an "integrally closed" ring, which had a very similar proof. Check the solutions if you haven't done so

already. This notion of "integrally closed" is related to why something like $\mathbb{Z}[\frac{1}{2}(1+\sqrt{-19}]$ instead of $\mathbb{Z}[\sqrt{-19}]$ is the "correct" analog of $\mathbb{Z}$ in $\mathbb{Q}(\sqrt{-19})$, which we'll say a bit about next quarter.

Warm-Up 2. We show that the monic polynomial x^4+5x+5 is irreducible over $\mathbb{Z}[i]$, which by Gauss's Lemma is equivalent to irreducibility over the field $\mathbb{Q}(i)$ of complex numbers with rational real and imaginary parts. If this polynomial were to be reducible over $\mathbb{Q}(i)$, it would have to factor into either a product of linear and cubic polynomials, or a product of two quadratic polynomials.

We can rule out the first possibility using the rational root theorem, since any linear factor would have to correspond to a root in $\mathbb{Q}(i)$. The only possible rational roots $\frac{a}{b} \in \mathbb{Q}(i)$ (where a and b are both Gaussian integers) come from those numerators which divide 5 and denominators which divide 1, by the rational root theorem. The divisors of 5 and 1 in $\mathbb{Z}[i]$ are

$$5 : \text{units}, (\text{unit})5, (\text{unit})(2+i), (\text{unit})(2-i) \quad \text{and} \quad 1 : \text{units}.$$

(Note that something like $-1+2i$ is included in the "(unit)$(2+i)$" case: $i(2+i) = -1+2i$. Also, this makes use of the a^2+b^2 theorem from before: 5 is not prime in $\mathbb{Z}[i]$, so reduces as $5 = (a+ib)(a-ib)$ for some primes $a+ib, a-ib$.) This gives the following potential rational roots:

$$\text{unit}, \ (\text{unit})5, \ (\text{unit})(2+i), \ (\text{unit})(2-i)$$

since a unit divided by a unit is just a unit. The units of $\mathbb{Z}[i]$ are $\pm 1, \pm i$, so now we can perform a brute-force check that none of the above are actually roots of $p(x) = x^4+5x+5$: for instance,

$$p(5) = 5^4+5(5)+5 = 655 \neq 0, \ p(2+i) = (2+i)^4+5(2+i)+5 = 8+29i \neq 0,$$

and similarly for the others. Since x^4+5x+5 has no root in $\mathbb{Q}(i)$, it cannot be factored into a product linear and cubic polynomials.

If x^4+5x+5 could be written as a product of quadratics in $\mathbb{Z}[i]$, we would have

$$x^4+5x+5 = (x^2+ax+b)(x^2+cx+d) \text{ for some } a,b,c,d \in \mathbb{Z}[i].$$

(The factors can be taken to be monic since if the first had leading coefficient a unit $s \in \mathbb{Z}[i]$ and the second s^{-1}, we can factor s out of the first and s^{-1} out of the second to leave leading coefficients 1, without changing the product of the factors since $ss^{-1} = 1$.) Expanding on the right and comparing coefficients with the left gives the following requirements:

$$a+c = 0, \ b+ac+d = 0, \ ad+bc = 5, \ bd = 5.$$

From the first we have $c = -a$, and then the third becomes

$$a(d-b) = 5.$$

Thus b, d, and $d-b$ all must be divisors of 5 in $\mathbb{Z}[i]$. But by checking the possibilities, we can see that none will work: $b = \pm 1, d = \pm 5$ gives $d-b = \pm 4, \pm 6$, which cannot satisfy $a(d-b) = 5$ for any $a \in \mathbb{Z}[i]$; similarly if $b = \pm i, d = \mp 5i$; if $b = 2 \pm i, d = 2 \mp i$, then $d-b = \mp 2i$, which also does not divide 5 as in $a(d-b) = 5$; and so on. Thus x^4+5x+5 cannot be written as a product of quadratic polynomials in $(\mathbb{Z}[i])[x]$, so we conclude that it is irreducible over $\mathbb{Z}[i]$. (We are not checking all the distinct possibilities here by hand since we will soon see a more efficient way of showing that this polynomial is irreducible.)

Unique factorization of polynomials. We can now argue that if R is a UFD, then $R[x]$ is a UFD, and thus by induction that $R[x_1, \ldots, x_n]$ is a UFD no matter the number of variables.

The proof is essentially Gauss's Lemma: if F denotes the fraction field of R, then $F[x]$ is a UFD, so any $p(x) \in R[x]$ can be factored into irreducibles in $F[x]$, which can then be turned into a factorization in $R[x]$ as a result of Gauss's Lemma. The resulting factors can be shown to be irreducible in $R[x]$, and unique using the uniqueness in $F[x]$. Gauss's Lemma is truly the key step. (Indeed, the fact that "R UFD implies $R[x]$ UFD" is what many other sources refer to as "Gauss's Lemma". Yet another use of the term "Gauss's Lemma" refer to the fact that products of primitive polynomials are primitive, but all of these uses are essentially saying the same thing in the end.) As a consequence, we now know that $\mathbb{Z}[x]$ is indeed a UFD which is not a PID.

Reduction mod I. We finish with two final irreducibility tests, which are both dependent on taking quotients modulo an ideal. For the first, suppose R is an integral domain and $I \subseteq R$ a proper ideal. If $p(x) \in R[x]$ is monic, the claim is that if the reduction $p(x) \bmod I$ as an element of $(R/I)[x]$ cannot be factored into polynomials of smaller degree, then $p(x)$ is irreducible over R. The point is that $p(x) \bmod I$ will usually be a simpler polynomial than $p(x) \in R[x]$, for which being able to factor or not might be simpler to determine. Note we use "cannot be factored into polynomials of smaller degree" instead of simply "irreducible" only because we technically defined the notion of irreducible elements only in integral domains, and R/I is not an integral domain unless I is prime; if I is prime, we can just ask that $p(x) \bmod I$ be irreducible over R/I.

For example, consider $x^2 + 3x + 7$ over $\mathbb{Z}$. Reducing by the prime ideal (3) gives $x^2 + 1$ over $\mathbb{Z}/3\mathbb{Z}$, which cannot be factored into polynomials of smaller degree since it has no root in $\mathbb{Z}/3\mathbb{Z}$. Thus $x^2 + 3x + 7$ must have been irreducible over $\mathbb{Z}$ to begin with. Now, this type of example (over $\mathbb{Z}$) is perhaps not so enlightening, since irreducibility can be determined by other means, such as the rational root theorem. (Still, often a check via reduction will involve less work than other tests.) For a more interesting example, consider $x^3 + x^2y + y - 2$ in $\mathbb{Z}[x, y]$. Reducing by the ideal (y) gives $x^3 - 2$ in $\mathbb{Z}[x, y]/(y) \cong \mathbb{Z}[x]$, which cannot be factored by Gauss's Lemma since it is irreducible over $\mathbb{Q}$ by the rational root test. Thus $x^3 + x^2y + y - 2$ is irreducible in $\mathbb{Z}[x, y]$.

The proof of the "reduction mod I" test above is almost immediate: if $p(x) = a(x)b(x)$ was reducible in $R[x]$ with $a(x), b(x)$ nonunits, then both $a(x)$ and $b(x)$ have smaller degree than $p(x)$ (neither can be a constant nonunit since the leading coefficient of $p(x)$ should be 1), taking coefficients mod I would give $\overline{p(x)} = \overline{a(x)}\,\overline{b(x)}$ in $(R/I)[x]$ (bars indicate reduction), so that $\overline{p(x)}$ would be a product of polynomials of smaller degree. The contrapositive is then our given claim. Now, there is one subtely here: we should actually take $a(x)$ and $b(x)$ to be *monic* in order to guarantee that the reductions $\overline{a(x)}$, $\overline{b(x)}$ have the same degree as $a(x), b(x)$, so that $\overline{p(x)} = \overline{a(x)}\,\overline{b(x)}$ is indeed a product of polynomials of smaller degree. The fact that R is an integral domain guarantees that $p(x) = a(x)b(x)$ has both $a(x)$ and $b(x)$ of smaller degree than $p(x)$ (otherwise leading coefficients would have to multiply to zero), and then the leading coefficients can be taken to be one since if they were a_m and b_n respectively, we would have $a_m b_n = 1$ as the leading coefficient of $p(x)$, meaning that a_m and b_n are units, so that they can be "divided" out to get monic $a(x)$ and $b(x)$. (An example of what goes wrong in a non-integral domain is included on the homework.)

Eisenstein's Criterion. Finally, here is one more test for irreducibility—known as *Eisenstein's Criterion*—based on reduction by a *prime* ideal, although not usually stated in terms of reduction:

> Suppose R is an integral domain and $P \subseteq R$ is a prime ideal. Then if $p(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 \in R[x]$ is monic and $a_{n-1}, \ldots, a_1, a_0 \in P$ but $a_0 \notin P^2$, then $p(x)$ is irreducible. (Stated in terms of reductions instead, the assumption is that $p(x) = x^n \bmod P$ and $a_0 \neq 0 \bmod P^2$.)

This might seem like a strange set of assumptions at first, but the power lies in the ability to pick

any prime ideal P we need to make it work. Moreover, we will see that there are multiple ways of applying this result in cases where it might not seem to be applicable at first.

Here is a proof. Aiming for a contradiction, suppose that $p(x) = b(x)c(x)$ is reducible, with $b(x), c(x) \in R[x]$ monic of smaller degree than $p(x)$. (See above for why this always possible to achieve.) Then reducing mod P gives:

$$x^n = \overline{p(x)} = \overline{b(x)}\,\overline{c(x)} \text{ in } (R/P)[x]$$

since all $a_i \in P$. If b_0, c_0 are the constant terms of $b(x), c(x)$ respectively, this reduction says that $b_0 c_0 = 0$ in R/P, so that $b_0 = 0 \bmod P$ or $c_0 = 0 \bmod P$ since R/P is an integral domain. Say that $b_0 = 0$. Then consider the term $b_1 x$ in $b(x)$: we get $b_1 c_0$ as the coefficient of x in the reduced product, so again this must be zero, meaning that $b_1 = 0$ or $c_0 = 0$ in R/P. If it is b_1 which is zero, we can continue in this way until we come across some b_i for which $b_i c_0 = 0$ (coefficient of x^i as we work our way up) implies that c_0 is the one which is actually zero: more formally, once we have $b_0 = 0$, take $b_k x^k$ $(k > 0)$ to be the minimal degree term in $\overline{b(x)}$, so that $b_k \neq 0$ and then $b_k c_0 = 0$ (the potential minimal degree term in $\overline{b(x)}\,\overline{c(x)}$) implies $c_0 = 0$ in R/P. Since we now have $b_0, c_0 = 0$ in R/P, both b_0, c_0 are in P, so $b_0 c_0 = a_0$ is in P^2, a contradiction. Hence $p(x)$ was irreducible to begin with.

Example. Here is a standard type of example. Take $p(x) = x^4 + 30x^3 + 15x^2 + 6x + 33$ over $\mathbb{Z}$. Then 3 divides every non-leading coefficient (i.e. each of these coefficients is in the prime ideal (3)), and 3^2 does not divide the constant term 33 (i.e. $33 \notin (3)^2$), so $p(x)$ is irreducible in $\mathbb{Z}[x]$ by Eisenstein's Criterion. Note that since 3 is also prime in $\mathbb{Z}[i]$, this also shows that $p(x)$ is irreducible over $\mathbb{Z}[i]$, and hence over $\mathbb{Q}(i)$ by Gauss's Lemma.

It might seem at first that Eisenstein's Criterion is pretty limited in scope, since it is not very often the case that one encounters the type of polynomial to which it applies as written. However, as we will see next time, it does apply more broadly once we take into account the ability to make a "change of variables" in a polynomial. For instance, we will show that for p prime, the *p-th cyclotomic polynomial*

$$\phi_p(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$$

is irreducible over $\mathbb{Z}$ using Eisenstein's Criterion, even though as written the conditions in this criterion are not satisfied.

Lecture 20: Modules over Rings

Warm-Up 1. For p prime, we show that the *p-th cyclotomic polynomial*

$$\phi_p(x) = x^{p-1} + x^{p-2} + \cdots + x + 1$$

is irreducible over $\mathbb{Z}$, and hence $\mathbb{Q}$ by Gauss's Lemma. (We will study cyclotomic polynomials in detail next quarter in the context of cyclotomic fields, and the fact that these polynomials are irreducible will be important. Irreducibility is not obvious, and in fact $x^{n-1} + x^{n-2} + \cdots + x + 1$ is irreducible if and only if n is prime. For n composite, you should try to figure out how to factor this polynomial!) We will use Eisenstein's Criterion, although as written it does not seem to be applicable to $\phi_p(x)$. A first thing to note is that $\phi_p(x)$ can actually be written more compactly as

$$\phi_p(x) = \frac{x^p - 1}{x - 1},$$

which uses the standard formula $1 + x + x^2 + \cdots + x^k = \frac{x^{k+1}-1}{x-1}$, often proved using induction.

Here is the trick: instead of $\phi_p(x)$, consider the polynomial $\phi_p(x+1)$:

$$\phi_p(x+1) = (x+1)^{p-1} + (x+1)^{p-2} + \cdots + (x+1) + 1 = \frac{(x+1)^p - 1}{(x+1) - 1} = \frac{(x+1)^p - 1}{x}.$$

The point is that if $\phi_p(x+1)$ is irreducible, then so is $\phi_p(x)$ (!), since a factorization $\phi_p(x) = a(x)b(x)$ of $\phi_p(x)$ into polynomials of smaller degree would lead to such a factorization for $\phi_p(x+1) = a(x+1)b(x+1)$ as well. (Other types of "change of variables" are also valid in such an argument.)

To see that $\phi_p(x+1)$ is irreducible, we use the binomial formula:

$$(x+1)^p = \sum_{k=0}^{p} \binom{p}{k} x^{p-k} = x^p + px^{p-1} + \cdots + px + 1, \text{ where } \binom{p}{k} = \frac{p!}{k!(p-k)!}.$$

This gives

$$\phi_p(x+1) = \frac{(x+1)^p - 1}{(x+1) - 1} = \frac{(x+1)^p - 1}{x} = x^{p-1} + \sum_{k=1}^{p-2} \binom{p}{k} x^{p-k-1} + p$$

after simplifying. Now, for $0 < k < p$, p divides $\binom{p}{k} = \frac{p!}{k!(p-k)!}$: since $\binom{p}{k}$ is an integer, $k!(p-k)!$ divides $p! = p(p-1)!$, but $k!(p-k)!$ is relatively prime to p since p does not occur as a factor in either $k!$ or $(p-k)!$, so $k!(p-k)!$ must divide $(p-1)!$ and thus $\binom{p}{k} = p\frac{(p-1)!}{k!(p-k)!}$ is a multiple of p. Thus p divides each nonleading coefficient of $\phi_p(x+1)$, but p^2 does not divide the constant term p, so $\phi_p(x+1)$ is irreducible over $\mathbb{Z}$ by Eisenstein's Criterion. As explained before, this implies that $\phi_p(x)$ is irreducible as well.

Warm-Up 2. We show that $q(x) = x^4 + 8x^3 + 24x^2 + 37x + 31$ is irreducible over $\mathbb{Z}[i]$. (Hence it will also be irreducible over $\mathbb{Z}$, $\mathbb{Q}$, and $\mathbb{Q}(i)$.) As written Eisenstein's Criterion does not apply since the only prime dividing the constant term is 31, but this does not divide any other coefficient.

Consider instead $q(x-2)$:

$$q(x-2) = (x-2)^4 + 8(x-2)^3 + 24(x-2)^2 + 37(x-2) + 31 = x^4 + 5x + 5.$$

We already showed this polynomial was irreducible over $\mathbb{Z}[i]$ last time, by first ruling out linear factors and then quadratic ones by brute-force. But, we can more easily show this is irreducible using Eisenstein's Criterion. The prime $2 + i \in \mathbb{Z}[i]$ (prime since $(2+i)(2-i) = 5$ is an integer prime) divides 5 and 0 (coefficient of x^3 and x^2), but $(2+i)^2 = 3 + 4i$ does not divide the constant term 5, so $q(x-2)$ is irreducible over $\mathbb{Z}[i]$. (Of course, we cannot use 5—which divides the constant term and the coefficient of 1—as the prime in Eisenstein's Criterion for $\mathbb{Z}[i]$ since 5 is not prime in $\mathbb{Z}[i]$; 5 would work, however, if we were only checking for irreducibility over $\mathbb{Z}$.) Since $q(x-2)$ is irreducible, $q(x)$ is also irreducible over $\mathbb{Z}[i]$ because a factorization of $q(x)$ would necessarily give a factorization of $q(x-2)$ as well.

Minimal polynomials and fields. We've mentioned that one reason why we care about irreducibility is so that we can construct interesting examples of fields by taking quotients: $F[x]/(f(x))$ is a field if and only if $f(x)$ is irreducible over the field F. But, let us say a quick word about another way in which irreducible polynomials will show up next quarter.

Given a containment $E \subseteq F$ of fields, we call F an *extension* of E. One basic question we will then consider is whether a given element $a \in F$ is the root of a polynomial $p(x) \in E[x]$. For

instance, $\mathbb{Q}(\sqrt{2})$ is an extension of $\mathbb{Q}$ and $\sqrt{2} \in \mathbb{Q}(\sqrt{2})$ is a root of $x^2 - 2 \in \mathbb{Q}[x]$. If so, then we can ask for the *minimal polynomial* of a over E, which is the polynomial of smallest degree over E for which a is a root. In the example above, $x^2 - 2$ is the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$. It will not be hard to show that such a minimal polynomial is always irreducible over the base field E, and the point is that much of the structure of the field extension $E \subseteq F$ is in fact encoded within such minimal polynomials. Thus, having methods available to test for irreducibility will be useful for a wider study of fields in general. (We will also see the notion of a minimal polynomial in the context of modules at the end of this quarter.)

Modules. Now we come to our final topic for the quarter, the theory of *modules*. There are a few reasons why it makes sense to discuss this notion now:

- First, we have been mentioning "modules" as far back as last quarter, and indeed alluded to the fact that the structure theorem for finitely generated abelian groups—which says that any finitely generated abelian group is a product of cyclic groups—was really a statement about modules more generally. So, we will now be able to discuss this formally, and finally provide proofs of some results we stated at the end of last quarter. (We also briefly spoke about viewing the fact that $\mathrm{Aut}(\mathbb{Z}/p\mathbb{Z}\times\mathbb{Z}/p\mathbb{Z}) \cong GL_2(\mathbb{Z}/p\mathbb{Z})$—which we used when classifying semi-direct products last quarter—as being about a generalization of linear algebra, for which module theory will provide the correct context. In particular, using module theory we will be able to completely describe the conjugacy classes of $GL_n(\mathbb{Z}/p\mathbb{Z})$.)
- Second, given a field extension $E \subseteq F$, it will always be possible to view F as a "module over E", or really in this special case as a "vector space over E", and the theory of modules will give us the language we need (think "span", "linear independence", "basis", "dimension") to study such extensions in detail.
- Finally, modules are natural to consider after discussing rings since they provide the things on which rings "act", so that module theory is essentially the right way to study "ring actions".

Here, then, is the main definition: a *module* over a ring R (also called an *R-module*) is an abelian group M (written additively) equipped with a map $R \times M \to M,\ (r, m) \mapsto rm$ such that

(i) $r(x + y) = rx + ry$ for any $r \in R$ and $x, y \in M$;
(ii) $(r + s)x = rx + sx$ for any $r, s \in R$ and $x \in M$; and
(iii) $r(sx) = (rs)x$ for any $r, s \in R$ and $x \in M$.

Moreover, when R has unity, we usually also require the following:

(iv) $1x = x$ for all $x \in M$.

We think of the map $(r, m) \mapsto rm$ as a "scalar multiplication" on M by elements of R, akin to the notion of scalar multiplication in linear algebra. We should also view this as an *action* of R on M, and indeed properties (iii) and (iv) above directly analogous to properties we saw in the definition of a group action: (iii) says that "acting" by s and then by r is the same as acting by rs, and (iv) says that the identity "acts" as the identity. (Property (iv) is really there to exclude the "trivial" action $1x = 0$ from consideration.) The first two properties then express compatibilities between the action and the additive structure of M in terms of two distributive properties. All of these properties are also analogous to the ones we saw in the definition of "ring", and indeed any ring can be viewed as a module over itself. For the most part we will only care about modules over commutative rings with unity, but the notion of a module makes sense for any ring.

Technically, what we have defined above is a *left module* since elements of R act on the left. There is an analogous notion of *right module*, but we will strictly work with left modules (just as how we really only considered left actions of groups last quarter) in this course.

Examples. For any ring R, $R^n = R \times \cdots \times R$ (n times) is a module over R with componentwise addition and scalar multiplication. (Here we ignore the componentwise multiplication which turns R^n into a ring itself.) Also, the set of $n \times n$ matrices $M_n(R)$ is an R-module with usual matrix addition and scalar multiplication. For a third example of an R-module, take the set $R[x_1, \ldots, x_n]$ of a polynomials in n variables, equipped again with the usual addition and scalar multiplication.

Now, each of the examples above is a bit special, since each not only has a scalar multiplication defined on it, but indeed a *ring* multiplication as well: componentwise multiplication in R^n, multiplication of matrices in $M_n(R)$, and multiplication of polynomials in $R[x_1, \ldots, x_n]$. These additional structures (when R is commutative and has an identity) actually turn each of these into what are called *algebras* over R: an *R-algebra* is a unital ring A equipped with the structure of an R-module such that $r(ab) = (ra)b = a(rb)$ for all $r \in R$ and $a, b \in A$. So, an algebra over a ring is a module with a ring structure itself which is in a way (via the equation above) compatible with the module structure. We will not really say much if anything about algebras in this course since we will mainly only about their module structures, but they are widely used. (There are equivalent ways of phrasing the definition of an algebra, for instance as a ring homomorphism $R \to A$, but you can check the book or other sources for more information if interested.)

Vector spaces. In the case where the base ring is actually a field F, a module over F is called a *vector space* over F, or an *F-vector space*. This is exactly the same the notion of "vector space" you would have seen in a previous linear algebra which covered abstract vector spaces (such as MATH 291 or MATH 334 here), and we will see that pretty much everything you learned about real vector spaces (just think about $\mathbb{R}^n$ if you have not seen abstract vector spaces before) previously will hold for vector spaces over any field.

For example, $\mathbb{R}^n$ is a vector space over $\mathbb{R}$, $\mathbb{C}^n$ is a vector space of $\mathbb{C}$ which can also be viewed as a vector space over $\mathbb{R}$ by considering scalar multiplication only by real scalars, and so on. Perhaps most interestingly, we can work with vector spaces over *finite* fields, such as $(\mathbb{Z}/p\mathbb{Z})^n$, $M_n(\mathbb{Z}/p\mathbb{Z})$, and $(\mathbb{Z}/p\mathbb{Z})[x]$. And as stated earlier, if E is a field contained in a field F, then F becomes an E-vector space when equipped with the scalar multiplication which comes from the field multiplication F: define $e \in E$ acting on $f \in F$ just as the product ef in F.

Abelian groups. A key example of a module comes when the base ring is $\mathbb{Z}$, where the fact is that a $\mathbb{Z}$-module is nothing but the same thing as an ordinary abelian group. One direction is trivial, since by definition any module over a ring is an abelian group with extra structure. The interesting claim here is that *any* abelian group can be given the structure of a $\mathbb{Z}$-module. Indeed, suppose M is an abelian group and define the desired action of $\mathbb{Z}$ on M by "repeated addition":

- for positive $n \in \mathbb{Z}$, set $nx := \underbrace{x + \cdots + x}_{n \text{ times}}$ for any $x \in M$,
- for negative $n \in \mathbb{Z}$, set $nx := \underbrace{(-x) + \cdots + (-x)}_{|n| \text{ times}}$ for any $x \in M$ with $-x$ its additive inverse,
- set $0x = 0$ for all $x \in M$.

This integer "scalar multiplication" then turns M into a $\mathbb{Z}$-module. It is crucial here that M is an abelian group, since otherwise the module properties would not old. In particular, the distributive property:

$$n(x + y) = nx + ny,$$

say for n positive, becomes the requirement that

$$\underbrace{(x+y)+\cdots+(x+y)}_{n \text{ times}} = \underbrace{x+\cdots+x}_{n \text{ times}} + \underbrace{y+\cdots+y}_{n \text{ times}},$$

for which M needs to be abelian to hold.

So, the study of abelian groups can be recast as the study of $\mathbb{Z}$-modules, and in particular the structure theorem for finitely generated abelian groups will just become the structure theorem for $\mathbb{Z}$-modules. The key fact which makes this structure theorem work out nicely, as we'll see, is that $\mathbb{Z}$ is a PID, and indeed modules over PIDs in general are the nicest ones to work with, We will see that module theory over a PID behaves very much like ordinary linear algebra, even when the base ring is not a field, although for sure there will be some subtleties when working over PIDs which are not fields. (This will still be better than working over general rings however. The basic reason has to do with the notion of *Noetherian* we introduced previously.)

Other constructions. Many concepts from ring theory have easy analogues for modules. For instance, a *submodule* of an R-module M is an additive subgroup $N \leq M$ which is "closed under scalar multiplication", meaning that $rn \in N$ for any $r \in R$ and $n \in N$, or more compactly $rN \subseteq N$ for any $r \in R$. This guarantees that N itself is a module over R under the same operations as those present on M. As a special case, when we consider R to be a module over itself, the submodule condition $rI \subseteq I$ for an additive subgroup $I \leq R$ is precisely the ideal condition, so ideals are nothing but submodules of R viewed as a module over itself.

Next, given a submodule N of M, we can form the *quotient module* M/N (a module over R still) by taking the usual quotient of additive groups and using the scalar multiplication which exists on M to define one on M/N: for instance,

$$r(x+N) := rx + N$$

when we think of M/N in terms of cosets. As is usual when working with quotient constructions, we need to know this scalar multiplication is well-defined, but the reason for this is very similar to what happened with ideals: if $x = y$ in M/N, then $x - y \in N$, so $r(x-y) = rx - ry \in N$ since N is a submodule of M, and hence $rx = ry$ in M/N.

Finally, we can speak of *generating sets* for modules. We say that an R-module M is generated by $x_1, \ldots, x_n \in M$ if every element of M is of the form

$$r_1 x_1 + \cdots + r_n x_n \text{ where } r_i \in R.$$

We denote this by $M = \langle x_1, \ldots, x_n \rangle$, where the right side denotes the module obtained by taking $x_1, \ldots, x_n$ and including all things you need to actually get a module, which ends up being precisely the set of elements of the form above. (So, $\langle x_1, \ldots, x_n \rangle$ is the "smallest" module containing $x_1, \ldots, x_n$.) When $M = \langle x \rangle$ is generated by a single element, we call it a *cyclic* module. (Note that here we use the more group-theoretic term "cyclic" as opposed to the ring-theoretic term "principal" we used for the analogous ideal concept. When viewing R as a module over itself, its cyclic submodules are precisely its principal ideals.) The expression above should remind you of "linear combinations" from linear algebra, and indeed we will refer to

$$r_1 x_1 + \cdots + r_n x_n \text{ where } r_i \in R$$

as an *R-linear combination* of $x_1, \ldots, x_n \in M$. The module $\langle x_1, \ldots, x_n \rangle$ generated by $x_1, \ldots, x_n$ is then called the *R-linear span* of $x_1, \ldots, x_n$, borrowing further terminology from linear algebra.

Lecture 21: Module Homomorphisms

Warm-Up 1. Let M be a module over an integral domain R. A *torsion* element $x \in M$ is one for which there exists a nonzero $r \in R$ such that $rx = 0$. (We say here that r *annihilates* x. Yes, that is an actual, well-established mathematical term!) Denote by $\text{Tor}(M)$ the set of torsion elements of M. We show that $\text{Tor}(M)$ is a submodule of M, called the *torsion submodule*.

First, if $x, y \in \text{Tor}(M)$, there exist nonzero $r, s \in R$ such that $rx = 0$ and $sy = 0$. Then

$$(rs)(x+y) = (rs)x + (rs)y = s(rx) + r(sy) = s0 + r0 = 0 + 0 = 0,$$

where we use commutative of R to say that $rs = sr$, and the fact that $r0 = 0$ in any module, which comes from manipulating $r0 = r(0+0) = r0 + r0$. (We saw the same property for rings at the start of the quarter, with the same proof. It is also true, for instance, that $0 \in R$ annihilates any module element, which comes from $0x = (0+0)x = 0x + 0x$. This is why we require that $r \in R$ in the definition of a torsion element be nonzero, since if we allowed $r = 0$ then everything would be a torsion element.) Since R is an integral domain, $rs \neq 0$, so $(rs)(x+y) = 0$ shows that $x+y$ is a torsion element and hence $x + y \in M$. (If R is not an integral domain, but still commutative, the definition of a torsion element has to be modified to require that r not be a zero divisor.)

If r and x are as above, and $t \in R$, then

$$r(tx) = t(rx) = t0 = 0,$$

so $tx \in \text{Tor}(M)$, where again commutativity of R was necessary. This shows that $\text{Tor}(M)$ is closed under the action of R. Finally, using the fact that $r(-x) = -rx$ in any module (show $rx + r(-x) = 0$ using distributivity), we have

$$r(-x) = -rx = -0 = 0,$$

so $-x \in \text{Tor}(M)$ and $\text{Tor}(M)$ is closed under taking additive inverses. We conclude that $\text{Tor}(M)$ is a submodule of M as claimed.

In the case of a $\mathbb{Z}$-module, so an abelian group, this agrees with the notion of *torsion subgroup* we saw in some homework problems last quarter. Indeed, saying that $x \in G$ is a torsion element when G is an abelian group (written additive) says that $nx = 0$ for some nonzero $n \in \mathbb{N}$, which means precisely that x has finite order since nx means to repeatedly apply the group operation. So, the torsion submodule of G is the subgroup of elements of finite order, i.e. the torsion subgroup.

Warm-Up 2. We seek to determine the data on $\mathbb{R}^2$ encoded by the structure of an $\mathbb{R}[x]$-module. First, since $\mathbb{R}$ is a subring of $\mathbb{R}[x]$, an $\mathbb{R}[x]$-module structure on $\mathbb{R}^2$ also gives an $\mathbb{R}$-module structure by considering only scalar multiplication by constant polynomials. An $\mathbb{R}$-module structure is a vector space structure, so at the very least we have some scalar multiplication of $\mathbb{R}$ on $\mathbb{R}^2$ which turns it into a real vector space. Assume for now that this is just the usual componentwise scalar multiplication. (Of course, it could be some other one in general.)

We claim that the rest of the $\mathbb{R}[x]$-module structure on $\mathbb{R}^2$ is fully characterized solely by the action of $x \in \mathbb{R}[x]$ on $\mathbb{R}^2$. Indeed, if we know how x acts then we know how any x^k acts since:

$$x^k \mathbf{y} = \underbrace{x(\dots(x}_{k \text{ times}}\mathbf{y})\dots) \text{ for } \mathbf{y} \in \mathbb{R}^2,$$

i.e. acting by x^k is acting by x k times. Then we will know how any polynomial acts by distributivity; for instance

$$(2 + 3x - x^2)\mathbf{y} = 2\mathbf{y} + 3x\mathbf{y} - x^2\mathbf{y}.$$

So, the action of any $p(x) \in \mathbb{R}[x]$ is known once we know the map $x : \mathbb{R}^2 \to \mathbb{R}^2$ which x induces. But this map has some key properties:

$$x(\mathbf{y} + \mathbf{z}) = x\mathbf{y} + x\mathbf{z} \text{ for } \mathbf{y}, \mathbf{z} \in \mathbb{R}^2$$

and

$$x(a\mathbf{y}) = (xa)\mathbf{y} = (ax)\mathbf{y} = a(x\mathbf{y}) \text{ for } a \in \mathbb{R}.$$

This properties say precisely that the map $x : \mathbb{R}^2 \to \mathbb{R}^2$ induced by the action of x is a linear transformation in the usual linear algebraic sense, and is thus induced by a 2×2 matrix A. This then gives that the action by $p(x) \in \mathbb{R}[x]$ is the linear transformation $p(A)$ obtained by replacing every instance of x with the matrix A.

The upshot is that the data of an $\mathbb{R}[x]$-module structure on $\mathbb{R}^2$ is nothing but the data of an $\mathbb{R}$-vector space structure on $\mathbb{R}^2$ together with a linear map $\mathbb{R}^2 \to \mathbb{R}^2$, which in the special case where the scalar multiplication of $\mathbb{R} \subseteq \mathbb{R}[x]$ on $\mathbb{R}^2$ is the usual one, corresponds to a choice of a 2×2 matrix. More generally, the same is true for any field F: an $F[x]$-module structure on V is equivalent to an F-vector space structure on V (acting by constants) together with a linear transformation $V \to V$ describing the action of x. The point is that such objects are the standard thing one works with in linear algebra, and the theory of modules over polynomials rings (over fields) subsumes it all. An $F[x]$-submodule of V then corresponds to a subspace W of V on which the linear transformation induced by x should restrict to a map $W \to W$, which is usually called an *invariant* subspace in linear algebra. For instance, in the $\mathbb{R}^2$ case above, if we take the $\mathbb{R}[x]$-module structure given by the linear transformation which is reflection across the line $y = x$, then the line $y = x$ itself is a submodule of $\mathbb{R}^2$ since it is invariant under this reflection.

Another observation is that $\mathbb{R}^2$ is then a *torsion* module for such an $\mathbb{R}[x]$-module structure, meaning that it equals its own torsion submodule, or in other words every element of $\mathbb{R}^2$ is a torsion element. (The same is true of any *finite-dimensional* vector space over a field F equipped with a linear map $V \to V$ when viewed as an $F[x]$-module.) Indeed, take A to be the 2×2 matrix describing the action of $x \in F[x]$, and let $\mathbf{y} \in \mathbb{R}^2$. Then the vectors $\mathbf{y}, A\mathbf{y}, A^2\mathbf{y}$ are linearly dependent in $\mathbb{R}^2$ (since $\mathbb{R}^2$ is 2-dimensional), so there exists $a_0, a_1, a_2 \in \mathbb{R}$, at least one of which is nonzero, such that

$$a_0\mathbf{y} + a_1A\mathbf{y} + a_2A^2\mathbf{y} = \mathbf{0}.$$

But we can rewrite this as

$$(a_0I + a_1A + a_2A^2)\mathbf{y} = \mathbf{0}$$

where the left side is the action of the polynomial $a_0 + a_1x + a_2x^2 \in \mathbb{R}[x]$ on $\mathbf{y}$, so that this polynomial (which is nonzero) annihilates $\mathbf{y}$, and hence $\text{Tor}(\mathbb{R}^2) = \mathbb{R}^2$.

Let us make one final remark along these lines. In the discussion above, we obtain one annihilating polynomial $a_0 + a_1x + a_2x^2$ for each $\mathbf{y} \in \mathbb{R}^2$. But in fact, it is possible to find a single polynomial which annihilates all $\mathbf{y}$ at once: the *characteristic polynomial* of A. (We'll review what this is in the final week.) Indeed, the *Cayley-Hamilton Theorem* (covered in MATH 334)) says that if $p(x)$ is the characteristic polynomial of A, then $p(A)$— the matrix obtained by replacing x by A in $p(x)$—is the zero matrix, so that $p(a)\mathbf{y} = \mathbf{0}$ for all $\mathbf{y}$. In general, the *annihilator* $\text{Ann}(M)$ of an R-module M is the set of elements of R which annihilate *all* elements of M (which is in fact an ideal of R), so the Cayley-Hamilton Theorem from linear algebra can be interpreted as the claim that $p(x) \in \text{Ann}(\mathbb{R}^2)$ (or more generally for any finite-dimensional vector space over a field), and hence $\text{Ann}(\mathbb{R}^2) \neq \{0\}$. (In fact, in this setting, the annihilator is principal and generated by the *minimal polynomial* of the linear map induced by the action of x. We will discuss then notion of a

minimal polynomial shortly.) Again, the overall point here is that a whole bunch of linear algebra arises out of the study of modules over polynomial rings over fields!

Module homomorphisms . Given two R-module M and N, a *module homomorphism* $\phi : M \to N$ is a function which preserves the addition and action of R:

$$\phi(x+y) = \phi(x) + \phi(y) \text{ and } \phi(rx) = r\phi(x) \text{ for all } r \in R \text{ and } x, y \in M.$$

This should remind you of the definition of a linear transformation in linear algebra, and indeed module homomorphisms are also commonly called *R-linear maps*. A bijective module homomorphism is then a *module isomorphism* (or *R-linear isomorphism*), and we use the same $M \cong N$ as with other types of isomorphisms.

We can then give the usual meaning to terms like *kernel* and *image*, which will be submodules of M (the domain) and N (the codmain) respectively. It should also come as no surprise that the module versions of the *isomorphism theorems* hold, such as: $M/\ker\phi \cong \phi(M)$ as R-modules. The set of R-module homomorphisms $M \to N$ is denoted by $\mathrm{Hom}_R(M, N)$, and is an abelian (additive) group under addition of functions: $\phi + \psi$ defined by $(\phi+\psi)(x) = \phi(x) + \psi(x)$.

Example. For example, take R to be commutative and with unity, and consider the R-modules R^m and R^n. Then we claim that $\mathrm{Hom}_R(R^n, R^m)$ is isomorphic to the additive group $M_{m,n}(R)$ of $m \times n$ matrices with entries in R, which mimics the usual fact in linear algebra that linear transformations $\mathbb{R}^n \to \mathbb{R}^m$ (or you can replace $\mathbb{R}$ by any field you like) correspond to matrices. Indeed, the proof is the same: if we denote by e_i the element of R^n (think of elements here as column vectors) which has 1 in the i-th entry and 0 elsewhere, then for any R-linear map $\phi : R^n \to R^m$ we havec

$$\begin{aligned}
\phi \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} &= \phi(r_1 e_1 + \cdots + r_n e_n) \\
&= r_1\phi(e_1) + \cdots + r_n\phi(e_n) \\
&= \begin{bmatrix} \phi(e_1) & \cdots & \phi(e_n) \end{bmatrix} \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix}
\end{aligned}$$

where $\begin{bmatrix} \phi(e_1) & \cdots & \phi(e_m) \end{bmatrix}$ is the $m \times n$ matrix whose columns are the $\phi(e_i) \in \mathbb{R}^m$. This says that $\phi \mapsto \begin{bmatrix} \phi(e_1) & \cdots & \phi(e_n) \end{bmatrix}$ is then an isomorphism between $\mathrm{Hom}_R(R^n, R^m)$ and $M_{m,n}(R)$.

Endomorphism rings. In the special where $M = N$, so that we consider homomorphisms $M \to M$, $\mathrm{Hom}_R(M, M)$ is usually denoted by $\mathrm{End}_R(M)$ ("end" stands for *endomorphism*, which is a homomorphism from something to itself), and becomes a *ring* (the endomorphism ring of M) under composition. Thus, for instance, $\mathrm{End}_R(R^n) \cong M_n(R)$, the ring of $n \times n$ matrices over R. (Recall that composition corresponds to matrix multiplication.) The group of units of $\mathrm{End}_R(R^n)$ is then $GL_n(R)$, the group of invertible matrices over R. In general, the group of units of $\mathrm{End}_R(M)$ is $\mathrm{Aut}(M)$, the group of module isomorphisms from M to itself, i.e. automorphisms.

Invertible matrices mod p. And now we finish by saying a bit about something we used last quarter when classifying semi-direct products, namely the fact that $\mathrm{Aut}(\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z})$ is $GL_2(\mathbb{Z}/p\mathbb{Z})$. More generally, $\mathrm{Aut}((\mathbb{Z}/p\mathbb{Z})^n) \cong GL_n(\mathbb{Z}/p\mathbb{Z})$. By "Aut" here we mean group automorphisms, but the point is that $(\mathbb{Z}/p\mathbb{Z})^n$ is an abelian group, so it can be viewed as a $\mathbb{Z}$-module, in which case the

group automorphisms are the same as module automorphisms. Thus $\mathrm{Aut}((\mathbb{Z}/p\mathbb{Z})^n) \cong GL_n(\mathbb{Z}/p\mathbb{Z})$ follows directly from the discussion above.

Moreover, let us finally determine the order of this group. The key point is that a matrix over a field is invertible if and only if its columns are linear independent. No doubt you have seen this fact before for real matrices in linear algebra, and the proof there in fact works over any field. Thus, to count the number of elements of $GL_n(\mathbb{Z}/p\mathbb{Z})$, we must count the number of ways of forming linearly independent columns. There are p^n choices for the first column (p choices for each entry), except that we must exclude the zero vector if we want something invertible, which leavs us with $p^n - 1$ choices for the first column. Next, the second column should not be a multiple of the first column, so of the p^n possible column vectors we must exclude the p of which are multiples of the first column, which gives $p^n - p$ choices for the second column. The third column cannot be a linear combination of the first two, and there are p^2 such linear combinations, so there are $p^n - p^2$ choices for the third column, and so on. We end up with

$$(p^n - 1)(p^n - p)(p^n - p^2) \cdots (p^n - p^{n-1})$$

possible matrices, so $|GL_n(\mathbb{Z}/p\mathbb{Z})| = (p^n - 1)(p^n - p)(p^n - p^2) \cdots (p^n - p^{n-1})$. (The analogous formula works for invertible matrices over any *finite* field of order $q = p^n$.)

Lecture 22: Sums and Free Modules

Warm-Up. If R is a ring with unity, we show that any finitely generated R-module is a quotient of some R^n. Suppose $M = \langle x_1, \ldots, x_n \rangle$ is a finitely generated R-module. Denote by $e_i \in R^n$ the element which has 1 in the ith component and 0 elsewhere, and consider the map $\phi : R^n \to M$ defined by

$$\phi(e_i) = x_i \text{ and extended linearly.}$$

("Extended linearly" means to define it on other elements in the way it was to be in order to end up with an R-linear map: i.e. $\phi(\sum r_i e_i) = \sum r_i \phi(e_i)$.) This map is well-defined since there are no nontrivial relations satisfied by the e_i which would need to be preserved (i.e. these elements are "linearly independent", as we will soon define), and it is surjective since the $\phi(e_i) = x_i$ generate M. Thus by the First Isomorphism Theorem for modules we get

$$R^n / \ker \phi \cong M,$$

which exhibits M as a quotient of R^n as desired. (In fact, an analogous thing will be true of any module over R, not just those which are finitely generated, as soon as we figure out with what to replace R^n. The answer will be provided by the notion of a *free module*)

Linear independence. We now discuss the notion of linear independence in the setting of modules. The definition is the same as the (a?) one you would have seen in linear algebra: given an R-module M, we say $x_1, \ldots, x_n \in M$ are *linearly independent* over R (the base ring matters) if whenever $r_1, \ldots, r_n \in R$ satisfy

$$r_1 x_1 + \cdots + r_n x_n = 0,$$

we must have $r_1 = \cdots = r_n = 0$. (This is what it means to say that the only linear relation which holds among $x_1, \ldots, x_n$ is the trivial one.) If such a nontrivial relation holds (i.e. some r_i is nonzero), then $x_1, \ldots, x_n$ are *linearly dependent*.

Often in linear algebra the definition of linear dependence/independence is phrased in terms of whether one element can be expressed as a linear combination of the rest, which when over a

field is equivalent to the definition above. But this is not true when working over a general ring, in that for elements which are linearly dependent in the sense above it is not necessarily true that one is a linear combination of the others: for instance, 6 and 15 are linearly dependent in $\mathbb{Z}$ viewed as a module over itself since $(15)6 - (6)15 = 0$ is a nontrivial relation, but neither 6 nor 15 is a $\mathbb{Z}$-multiple of the other. The phrasing of "linear independence" given above in terms of linear relations is the one which works best over a general ring

Practically, a key consequence of the definition of linear independence—to which it is actually equivalent—is the following: $x_1, \dots, x_n \in M$ are linearly independent over R if and only if for any $x \in \langle x_1, \dots, x_n \rangle$, there are *unique* $r_1, \dots, r_n \in R$ such that

$$x = r_1 x_1 + \cdots + r_n x_n.$$

In other words, saying that $x \in \langle x_1, \dots, x_n \rangle$ by definition means x is an R-linear combination of the x_i, but independence says that there is only one possible set of coefficients we can use to actually express x in this way. This in turn is equivalent to saying that the module $\langle x_1, \dots, x_n \rangle$ generated by independent $x_1, \dots, x_n$ is isomorphic to R^n, with the isomorphism $\langle x_1, \dots, x_n \rangle \to R^n$ being the one that sends x to its "coefficient vector" $(r_1, \dots, r_n)$. (This map would not be well-defined if $x_1, \dots, x_n$ were linearly dependent.)

To see that this is true, suppose $x_1, \dots, x_n \in M$ are linearly independent and that

$$r_1 x_1 + \cdots + r_n x_n = s_1 x_1 + \cdots + s_n x_n$$

for some $r_i, s_i \in R$. Then

$$(r_1 - s_1)x_1 + \cdots + (r_n - s_n)x_n = 0,$$

so since $x_1, \dots, x_n$ are linearly independent we must have $r_i - s_i = 0$ for each i. Hence $r_i = s_i$ for each i, so the coefficients in $r_1 x_1 + \cdots + r_n x_n$ are unique. Conversely, suppose such coefficients are unique. Then if $r_1 x_1 + \cdots + r_n x_n = 0$, we have

$$r_1 x_1 + \cdots + r_n x_n = 0 x_1 + \cdots + 0 x_n.$$

Hence by uniqueness, we must have $r_i = 0$ for each i, so that $x_1, \dots, x_n$ are linearly independent. (So, definition of independence says that uniqueness of the coefficients for expressing 0 as a linear combination is enough to imply uniqueness for expressing any element in such a way as well.)

Direct sums. Using the notion of a sum of modules ($A + B$ is the module formed by sum $a + b$ with $a \in A$ and $b \in B$, just as we had before for ideals), note that

$$\langle x_1, \dots, x_n \rangle = \langle x_1 \rangle + \cdots + \langle x_n \rangle,$$

so that any finitely generated module is a sum of cyclic modules. We can ask what additional property this sum has when $x_1, \dots, x_n$ are linearly independent, with the answer being that this is actually a *direct sum*: a sum $N_1 + \cdots + N_k$ of submodules $N_i \leq M$ is a *direct sum* if the expression for an element $n_1 + \cdots + n_k$ in the sum is unique, meaning that if

$$n_1 + \cdots + n_k = n_1' + \cdots + n_k'$$

for $n_i, n_i' \in N_i$, then $n_i = n_i'$ for all i. We use the notation $N_1 \oplus \cdots \oplus N_k$ in this case to emphasize the direct sum nature. So, for instance, when $x_1, \dots, x_n \in M$ are linearly independent, then

$$\langle x_1, \dots, x_n \rangle = \langle x_1 \rangle \oplus \cdots \oplus \langle x_n \rangle.$$

We can rephrase the direct sum property in a few equivalent ways. First, as with linear independence, it is enough to check the uniqueness requirement for $0 \in N_1 + \cdots + N_k$: this sum is a direct sum if and only if $0 = n_1 + \cdots + n_k$ for $n_i \in N_i$ implies $n_i = 0$ for each i. The proof is this equivalence is just the same (essentially) as that for the equivalent characterization of linear independence we gave previously, so we omit it here. Second, $N_1 + \cdots + N_k$ is a direct sum if and only if the map

$$N_1 \times \cdots \times N_k \to N_1 + \cdots + N_k \text{ defined by } (n_1, \ldots, n_k) \mapsto n_1 + \cdots + n_k$$

is an isomorphism. (The direct product $N_1 \times \cdots \times N_k$ is a module under the componentwise operations.) The point here is that the direct sum property is equivalent to the fact that we can uniquely reconstruct $(n_1, \ldots, n_k)$ from the knowledge of $n_1 + \cdots + n_k$, or in other words that the map above is injective. (It is always surjective for any sum, not necessarily direct.) That the inverse $n_1 + \cdots + n_k \to (n_1, \ldots, n_k)$ is well-defined is also just a way of rephrasing the direct sum property. We should point out that, motivated by this isomorphism, a direct product $M_1 \times \cdots \times M_k$ of R-modules in general (not necessarily all submodules of a larger module) is also referred to as a direct sum $M_1 \oplus \cdots \oplus M_k$, at least when only a finite number of modules are being considered. (We'll elaborate on this in a bit.) In this case, an expression $x_1 + \cdots + x_k \in M_1 \oplus \cdots \oplus M_k$ has no meaning as a usual "sum" in the sense that we can "add" an element of M_1 to an element of M_2, since this is not possible if M_1 and M_2 are not sitting inside of a larger common module, and rather should be interpreted only as $(x_1, \ldots, x_k) \in M_1 \times \cdots \times M_k$, without any interaction among the different x_i.

Finally, we give one final equivalent way of phrasing the direct sum property, which is perhaps the most practically useful for actually verifying that a given sum is a direct sum. The claim is that $N_1 + \cdots + N_k$ is a direct sum if and only if $N_i \cap (N_1 + \cdots + \widehat{N_i} + \cdots + N_k) = \{0\}$ for each i, where $\widehat{N_i}$ indicates that the N_i term should be omitted from the expression. So, the direct sum property is equivalent to each summand having trivial intersection with the sum of the rest. (This is somehow the correct analog of the "no vector is a linear combination of the rest" characterization of independence in linear algebra.) If the sum is direct and $x \in N_i \cap (N_1 + \cdots + \widehat{N_i} + \cdots + N_k)$, then $x \in N_i$ and we can express x as

$$x = n_1 + \cdots + 0 + \cdots + n_k$$

for $n_t \in N_t$ with 0 occurring in the N_i location. This gives two expressions for $x \in N_1 \oplus \cdots \oplus N_k$:

$$0 + \cdots + \underbrace{x}_{\in N_i} + \cdots + 0 = n_1 + \cdots + \underbrace{0}_{\in N_i} + \cdots + n_k,$$

so since the sum is direct we must have $x = 0$, so $N_i \cap (N_1 + \cdots + \widehat{N_i} + \cdots + N_k) = \{0\}$. Conversely, if this intersection property holds and

$$n_1 + \cdots + n_k = n_1' + \cdots + n_k'$$

for $n_t, n_t' \in N_t$, then

$$n_i - n_i' = (n_1' - n_1) + \cdots + \underbrace{\widehat{(n_i' - n_i)}}_{\text{omit}} + \cdots + (n_k' - n_k).$$

The left side is in N_i, and the right side is in $N_1 + \cdots + \widehat{N_i} + \cdots + N_k$, so this common element must be zero since the intersection of these is trivial, so $n_i = n_i'$ for each i. Hence the sum is direct.

(Note in particular that for a sum of only two modules, the conclusion is that $A + B$ is a direct sum if and only if $A \cap B = \{0\}$.)

Bases and free modules. Now we can generalize the notion of a basis from linear algebra. We say that a subset $E \subseteq M$ is a *basis* of the R-module M if it generates $M = \langle E \rangle$ and is linearly independent. To be clear, to say that E generates M means that for any $x \in M$ there exist $x_i \in E$ and $r_i \in R$ such that $x = r_1 x_1 + \cdots + r_n x_n$ (i.e. anything in M is a linear combination of *finitely* many elements of E), and to say that E is linearly independent means that any *finite* collection of vectors in E is linearly independent. Thus, in the end, E is a basis of M if any $x \in M$ is a unique linear combination of unique elements of E: there exist unique $x_i \in E$ and unique $r_i \in R$ such that $x = r_1 x_1 + \cdots + r_n x_n$. A *free* module is then one that has a basis. (It is not true in general that every module must have a basis, as opposed to the analogous situation in linear algebra.)

When the basis $E = \{x_1, \ldots, x_n\}$ is finite, M is isomorphic to R^n via the map

$$M \to R^n \text{ defined by } x \mapsto (r_1, \ldots, r_n)$$

where the $r_i \in R$ are the unique coefficients for which $x = r_1 x_1 + \cdots + r_n x_n$. Conversely, if $M \cong R^n$ (assuming R has unity), then M is free with basis the elements which correspond to $e_i \in R^n$ (1 in one component, 0 elsewhere). So, in this setting M is free if and only if it is isomorphic to some R^n, which if we recall our discussion above we can think of as either a direct product of n-many copies of R or as a direct sum. (Again, in the direct sum case, $r_1 + \cdots + r_n$ by definition denotes $(r_1, \ldots, r_n)$ with no confusion since we never "add" terms from different components together.) Another way of saying this all is that each cyclic module $\langle x_i \rangle$ for x_i a basis element is isomorphic to R—since each element rx_i in this cyclic module is uniquely characterized by r—so we have

$$M = \langle x_1, \ldots, x_n \rangle = \langle x_1 \rangle \oplus \cdots \oplus \langle x_n \rangle \cong R \oplus \cdots \oplus R = R^n.$$

When E is infinite, the same is true, once we give the proper definition of a "direct sum" of infinitely many copies of R. In general, for a set E we define the *direct sum* of "E-many copies of R", denoted $\bigoplus_{y \in E} R$, as

$$\bigoplus_{y \in E} R := \{(r_y)_{y \in E} \mid r_y = 0 \text{ for all but finitely many } y\}.$$

Here, $(r_y)_{y \in E}$ is an "E-tuple" of elements of R, meaning elements r_y of R indexed by elements y of E. Such things make up the *direct product* $\prod_{y \in E} R$ of E-many copies of R, which is a module over R under the "componentwise" operations. (More formally, infinite direct products like this should be defined as functions $E \to R$ which assign $r_y \in R$ to $y \in E$, but the "tuple" notation better matches the intuition we expect when E is finite.) The direct sum $\bigoplus_{y \in E} R$ is thus a submodule of the direct product $\prod_{y \in E} R$. The "$r_y = 0$ for all but finitely many y" requirement is there to ensure that we only consider things which we can treat as "sums", meaning only expressions with a finite number of nonzero terms, since it does not make sense in the setting of general rings to talk about "infinite sums" of elements. For instance, when $E = \{y_1, y_2, y_3, \ldots\}$ is countably infinite, we have:

$$\bigoplus_{y_i \in E} R = R \oplus R \oplus R \oplus \cdots$$

with a countably infinite number of components on the right. The tuple

$$(1, 1, 1, \ldots)$$

with $1 \in R$ in all components is not element of this direct sum (it *is* an element of the direct product $R \times R \times R \times \cdots$), which reflects the fact that an "infinite sum of 1's"

$$1 + 1 + 1 + \cdots$$

is not a well-defined expression in an arbitrary ring with unity. The tuple which has, say, r_4 in the 4th component, r_{10} in the 10th component, and r_{100} in the 100th component with all other components being zero *is* in the direct sum and should be thought of as the "sum" $r_4 + r_{10} + r_{100}$, or $r_4 y_4 + r_{10} y_{10} + r_{100} y_{100}$ if we want to make the basis elements explicit.

With this definition, the result is that an R-module (R with unity) M is free if and only if it is isomorphic to a direct sum of some number (possibly infinite, or uncountable, or whatever) of copies of R. If E is a basis of M, the direct sum we need is $\bigoplus_{y \in E} R$, where the required isomorphism is given by

$$M \to \bigoplus_{y \in E} R \text{ defined by } x \mapsto (r_y)_{y \in E}$$

where $x_1, \ldots, x_n \in E$ are the unique elements and $r_{x_1}, \ldots, r_{x_n} \in R$ the unique scalars satisfying

$$x = r_{x_1} x_1 + \cdots + r_{x_n} x_n,$$

with all other r_y being taken to be zero for $y \notin \{x_1, \ldots, x_n\}$. It is the "finiteness" condition in the definition of a basis that guarantees we get tuples in the direct sum instead of more generally the direct product. The inverse map sends $(r_x)_{x \in E}$ to the element $x = r_{x_1} x_1 + \cdots + r_{x_n} x_n \in M$ where the x_i are the indices at which r_{x_i} is not zero (we only get finitely many such terms by the direct sum definition) and the r_{x_i} are then the actual entries in those specific components. The point is that the entirety of the definition of "basis" (linearly independent plus generating) is encoded by the direct sum definition alone. Conversely, if M is isomorphic to a direct sum of some number of copies of R, then M is free with basis E given by the elements of M that correspond to tuples in the direct sum with a 1 in one component and zeroes elsewhere under the isomorphism.

Example. As an example, we claim that $R[x]$ is free as an R-module, with basis

$$1, x, x^2, \ldots.$$

These certainly generate $R[x]$ over R, and are linearly independent since

$$a_{i_1} x^{i_1} + \cdots + a_{i_k} x^{i_k} = 0$$

for $i_1 \neq \ldots \neq i_k$ implies that each a_{i_ℓ} is zero simply by definition of the zero polynomial as the one which has all coefficients zero. Indeed, $R[x]$ is isomorphic to the direct sum of a countably infinite number of copies of R:

$$R[x] = \bigoplus_{i=0}^{\infty} R = R \oplus R \oplus R \oplus \cdots,$$

with the isomorphism given by sending a polynomial to its tuple of coefficients:

$$a_0 + a_1 x + \cdots + a_n x^n \mapsto (a_0, a_1, \ldots, a_n, 0, 0, 0, \ldots).$$

The direct sum property guarantees that we get actual polynomials (with finitely many terms) out of this. If we instead used the full direct product $R \times R \times R \times \cdots$ on the right, we would get the power series ring $R[[x]]$ viewed as a module over R:

$$R[[x]] \cong R \times R \times R \times \cdots.$$

(Note: this and the claim $R[x] = R \oplus R \oplus \cdots$ above are only statements about isomorphisms of *modules*, not of *rings*, in that only the addition and scalar multiplication are being considered. Indeed, these isomorphisms do not extend to rings since the ring multiplications on $R[x]$ and $R[[x]]$ are not the componentwise ones.)

Note that $R[[x]] \cong R \times R \times \cdots$ does not immediately say that $R[[x]]$ is not free, since it could be the case that $R[[x]]$ is still isomorphic to *some* direct sum of copies of R, just not the countably infinite direct sum. In fact, there are some cases where $R[[x]]$ is free over R, and some where it is not, with no easy way to tell in general. (For example, it turns out that $\mathbb{Z}[[x]]$ is *not* free over $\mathbb{Z}$, but $(\mathbb{Z}/4\mathbb{Z})[[x]]$ *is* free over $\mathbb{Z}/4\mathbb{Z}$. Both of these claims take quite a bit of effort to prove.)

Lecture 23: Modules over PIDs

Warm-Up 1. We show that $\mathbb{Q}(\sqrt[3]{2})$ is free as a $\mathbb{Q}$-module by finding a basis. (The module structure just comes from the usual multiplication of elements of $\mathbb{Q}(\sqrt[3]{2})$ by elements of $\mathbb{Q} \subseteq \mathbb{Q}(\sqrt[3]{2})$. Since $\mathbb{Q}$ is a field, $\mathbb{Q}(\sqrt[3]{(2)}$ is a *vector space* over $\mathbb{Q}$, also called a "rational vector space".) First, recall from an example in the first week of class that elements of $\mathbb{Q}(\sqrt[3]{2})$ look like

$$a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2 \text{ with } a, b, c \in \mathbb{Q},$$

which comes from forming all possible sums and products of rationals with $\sqrt[3]{2}$, then all sums and products of *those* things, and so on; the fact that $(\sqrt[3]{2})^3 = 2$ implies that we need not go to any higher power than $(\sqrt[3]{2})^2 = \sqrt[3]{4}$ in order to describe all elements. Thus, as written, we can immediately see that

$$1, \ \sqrt[3]{2}, \ (\sqrt[3]{2})^2$$

span/generate $\mathbb{Q}(\sqrt[3]{2})$ over $\mathbb{Q}$.

We claim that these three are actually linearly independent over $\mathbb{Q}$, so that they will form a basis and $\mathbb{Q}(\sqrt[3]{2}) \cong \mathbb{Q}^3$ (as modules only, not rings) will be free as claimed. Indeed, suppose $a, b, c \in \mathbb{Q}$ satisfy

$$a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2 = 0.$$

Then $\sqrt[3]{2}$ is a root of the polynomial $a + bx + cx^2 \in \mathbb{Q}[x]$. If this polynomial were nonzero, it would necessarily generate the ideal $I \subseteq \mathbb{Q}[x]$ consisting of all polynomials having $\sqrt[3]{2}$ as a root, since in general ideals in $F[x]$ (F a field) are generated by polynomials of minimal degree inside of them (go back to the proof that Euclidean domains are PIDs), and there is certainly no polynomial of degree 1 in I since $\sqrt[3]{2} \notin \mathbb{Q}$. But $x^3 - 2$ is in I, so it would have to be a multiple of $a + b\sqrt[3]{2} + c(\sqrt[3]{2})^2$, which is not possible since $x^3 - 2$ is irreducible over $\mathbb{Q}$. Thus $a + bx + cx^2$ must in fact be the zero polynomial, so $a = b = c = 0$ and hence $1, \sqrt[3]{2}, (\sqrt[3]{2})^2$ are linearly independent over $\mathbb{Q}$ as desired.

This proof is a concrete example of something we mentioned previously: irreducible polynomials will be useful in helping to understand the structure of field extensions, in particular when it comes to identifying *bases* of such extensions.

Warm-Up 2. We show that free modules are always torsion-free (i.e. there are no nonzero torsion elements), and then that $\mathbb{Q}$ is an example of a $\mathbb{Z}$-module which is torsion-free but not free. First, assume M is free over an integral domain R, and let $x \in \text{Tor}(M)$ be a torsion element. Then there exists $0 \neq r \in R$ such that $rx = 0$. If E is a basis for M, then we can write

$$x = r_1 x_1 + \cdots + r_n x_n$$

for some unique $x_i \in E$ and $r_i \in R$. This gives

$$0 = rx = rr_1x_1 + \cdots + rr_nx_n.$$

Since $x_1, \ldots, x_n$ are linearly independent over R, we must have $rr_i = 0$ for each i. But then since $r \neq 0$ and R is an integral domain, we must have $r_i = 0$ for each i, so that

$$x = r_1x_1 + \cdots + r_nx_n = 0x_1 + \cdots + 0x_n = 0.$$

Hence the only torsion element is $x = 0$, so $\mathrm{Tor}(M) = \{0\}$ and M is torsion-free. (As a consequence, a module with nonzero torsion can never be free; for instance, no nontrivial finite abelian group is free as a $\mathbb{Z}$-module, since in this case every element is a torsion element.)

Now, $\mathbb{Q}$ is torsion-free as a $\mathbb{Z}$-module since adding a nonzero rational to itself (or its additive inverse to itself) never produces 0: i.e. $nr = 0$ for $0 \neq n \in \mathbb{Z}$ and $r \in \mathbb{Q}$ implies $r = 0$. If $\frac{a}{b}, \frac{c}{d} \in \mathbb{Q}$ are nonzero, then since

$$-bc\frac{a}{b} + da\frac{c}{d} = -ca + ac = 0$$

with $bc, da \in \mathbb{Z}$ nonzero, we have that $\frac{a}{b}$ and $\frac{c}{d}$ are linearly dependent over $\mathbb{Z}$. But this means that a potential basis for $\mathbb{Q}$ over $\mathbb{Z}$ (if it were to be free) could only consist of a single element (more than one element must give a linearly dependent set), but there is no single rational that generates all of $\mathbb{Q}$ as a $\mathbb{Z}$-module: $n\frac{a}{b}$ will never equal a rational whose denominator is divisible by primes other than those which divide b, so $\langle \frac{a}{b} \rangle$ is not all of $\mathbb{Q}$. Thus $\mathbb{Q}$ is not free over $\mathbb{Z}$.

We will soon see that a finitely generated torsion-free module over a PID is always free, so as a consequence of the work above we can conclude that $\mathbb{Q}$ is not finitely generated over $\mathbb{Z}$. (There are simpler ways of seeing this—in fact, this was already shown on a homework problem last quarter, only without using the language of "modules".) The current homework asks for an example of a finitely generated torsion-free module over a non-PID which is not free, so the PID assumption in the claim above is important, and starts to hint at why modules over PIDs behave more nicely than modules over more general rings.

Modules vs vector spaces. Now that we have introduced the notion of a *basis* in the setting of modules, we clarify some things which are different than what you might expect from the setting of linear algebra. First, as we have seen, it is not true that every module has a basis (i.e. is free), not even finitely generated ones. This is in stark contrast to *vector spaces* (i.e. modules over fields), where every vector space *does* have a basis, even in the "infinite-dimensional" setting. The proof that every vector space over a field has a basis is an application of Zorn's Lemma, and proceeds along the same lines as in the first example of Zorn's Lemma we gave last quarter: showing that $\mathbb{R}^\infty$ (viewed as a vector space over $\mathbb{R}$) has a basis. (Note that, as we pointed out last quarter, the vectors e_i, which have a 1 in one entry and zeroes elsewhere, do not form a basis of $\mathbb{R}^\infty$ since it is not true that every element can be written as a *finite* linear combination of these. The basis which Zorn's Lemma guarantees exists cannot be written down explicitly.) You can check last quarter's notes for the details, but the strategy is to use Zorn's lemma to produce a *maximal* linearly independent set and then show that this must actually span the entire space and hence be a basis. This proof will work over any field, and for spaces without any kind of "finite-dimensional" assumption.

Second, and perhaps even worse, for modules which *are* free, it is not guaranteed that two bases must always the same size, at least over noncommutative rings. Again, this is different than in ordinary linear algebra, where any two bases of a given vector space must have the same cardinality. For example, take the direct product of countably infinite many copies of $\mathbb{Z}$ as a $\mathbb{Z}$-module and set $R = \mathrm{End}_{\mathbb{Z}}(\mathbb{Z} \times \mathbb{Z} \times \cdots)$ to be its endomorphism ring. If we view this ring as a module over

itself, it turns out to be isomorphic to R^2 as an R-module, but also to R^3, and in fact isomorphic to R^n in general. (This is of course highly non-obvious!) Thus, we can produce bases of different cardinalities: viewing this module as R gives a basis of size 1, viewing it as R^2 gives a basis of size 2, and so on. This means that the analog of the linear-algebraic term "dimension" as the cardinality of a basis is not a well-defined notion for modules over non-commutative rings.

However, it is in fact true that over a *commutative* ring, any two bases of a free module *do* have the same cardinality. The cardinality of such a basis is then called the *rank* of the free module, and is the proper analog of "dimension". When we are working over a field, we will usually just use the more familliar term *dimension* instead of rank. Thus for instance, the $\mathbb{Q}$-module $\mathbb{Q}(\sqrt[3]{2})$ of the first Warm-Up has rank 3 over $\mathbb{Q}$, or dimension 3 as a rational vector space. The fact that two bases of a free module over a commutative ring have the same cardinality can be derived from the analogous claim about vector spaces (which you may or may not have seen before) by taking a correctly chosen maximal ideal $M \subseteq R$ and reducing everything modulo M to obtain a free module (now a vector space) over the field R/M instead, and showing that the size of a basis remains unchanged under this reduction. Alternatively, one can show that any spanning set must have cardinality at least that of any linearly independent set, and then deduce equal cardinality of bases from this. Indeed, this latter approach is usually how one proves the claim for vector space over fields in the first place.

Modules over PIDs. For the rest of the quarter we will focus our attention on modules over PIDs, or more precisely *finitely generated* modules over PIDs. The ultimate goal is the *structure theorem* for such modules, which essentially states that any such module is isomorphic to a direct sum of finitely many cyclic modules. Actually, the structure theorem(s) give even more information, by dictating what form the cyclic modules should take—there are two structure theorems we will state, dependent on the form we want. The relation between the two versions comes from the Chinese Remainder Theorem.

At the end, we will finish we an application to modules over polynomials rings $F[x]$ over fields. We saw previously that the data of an $F[x]$-module structure on V was the data of an F-vector space structure on V together with the choice of a linear map $V \to V$. The structure theorems in this case (note $F[x]$ is a PID) produce what are called *canonical forms* for such linear maps, which amount to nice ways of expressing these maps as matrices with respect to well-chosen bases.

Noetherian modules. One approach to understanding the structures we'll consider is through the use Noetherian modules, which is the module analog of the Noetherian rings we briefly looked at previously. To be sure, an R-module M is *Noetherian* if it satisfies the *ascending chain condition* for submodules: every ascending chain of submodules

$$N_1 \leq N_2 \leq N_3 \leq \cdots$$

of M terminates in the sense that there exists i such that $N_i = N_k$ for all $k \geq i$. As with rings, this is in some sense a type of "finiteness" condition which places restrictions on how "large" submodules can become. Now, as we saw with rings, the Noetherian condition is equivalent to the claim that every submodule of M is finitely generated. (With rings this was claim that every ideal was finitely generated.) The proof in the module setting is the same as the one we gave for rings earlier, only where we phrase everything in terms of submodules instead of ideals.

But let us now introduce another equivalent characterization of "Noetherian": M is Noetherian if and only if every nonempty collection of submodules of M contains a maximal element, meaning an element which is not strictly contained in any other element. This looks like a "Zorn's Lemma" type of statement, but is actually stronger: to apply Zorn's Lemma we need to know that any

chain inside of our collection $\mathcal{S}$ as an upper bound in $\mathcal{S}$ before we can conclude that $\mathcal{S}$ contains a maximal element, but here when M is Noetherian there is no such upper bound requirement, and any nonempty collection $\mathcal{S}$ of submodules of M contains a maximal element without any further hypothesis needing to be verified.

To see that this equivalence holds, suppose first that M is Noetherian and let $\mathcal{S}$ be a nonempty collection of submodules of M. Pick some $N_1 \in \mathcal{S}$. If N_1 was not maximal among elements of $\mathcal{S}$, then there exists $N_2 \in \mathcal{S}$ which strictly contains N_1: $N_1 \subsetneq N_2$. If N_2 was not maximal, then there exists $N_3 \in \mathcal{S}$ which strictly contains N_2:

$$N_1 \subsetneq N_2 \subsetneq N_3.$$

And so on, if at each step we did not have a maximal element of $\mathcal{S}$, we can keep extending this to a larger and larger chain of ascending submodules which would not terminate, contradicting the Noetherian assumption. Thus some $N_i \in \mathcal{S}$ along the way must be maximal. Conversely suppose that every nonempty collection of submodules of M contains a maximal element. Then in particular given any chain $N_1 \leq N_2 \leq \cdots$, the collection containing the terms in that chain has a maximal element N_i, and this will then be where the chain terminates: since $N_i \leq N_k$ for all $k \geq i$ and N_i is maximal, $N_i = N_k$ for $k \geq i$.

Finitely generated free modules. A large class of examples of Noetherian modules are provided by finitely generated modules over Noetherian rings, where you will show on the final homework that such things are indeed Noetherian themselves. This in particular applies to finitely generated modules over a PID, so that if M is a finitely generated module over a PID, then any submodule $N \leq M$ of M is also finitely generated. This is a nontrivial result, since it is not true that a submodule of a finitely generated module over non-PIDs must be finitely generated. For instance, if $R = \mathbb{R}[x_1, x_2, x_3, \ldots]$ is the polynomial ring in *infinitely* many variables over $\mathbb{R}$ (R is not a PID), then R is finitely generated as a module over itself (it is generated by the constant polynomial 1), but the submodule $\langle x_1, x_2, \ldots \rangle \leq R$ is not finitely generated.

If R is a PID and M is a finitely generated *free* module over R, then we immediately get that any nonzero submodule $N \leq M$ is finitely generated from the discussion above, but actually, and more importantly, it is true that N itself is also *free* of rank at most that of M. (View this as an analog of the fact that subspaces of finite-dimensional vector spaces over a field are themselves finite-dimensional.) This is a crucial property of finitely generated free modules over PIDs, and forms part of the key set of ideas behind the structure theorems we will soon look at. (In particular, we will use this next time to show that finitely generated torsion-free modules over PIDs are always free.) The PID assumption is important since it is not true that submodules of finitely generated free modules over more general rings are necessarily free themselves.

The proof that $N \leq M$ is free in this case proceeds by induction on the rank of $M \cong R^n$. When M has rank 1 (the base case), M is isomorphic to R, and so submodules of M correspond to ideals of R, which are all principal, so either (0) or free on one generator, so that $\operatorname{rank}(N) \leq \operatorname{rank}(M) = 1$. (A generator $a \in R$ of $(a) \neq (0)$ forms a basis for (a) because it is a linearly independent set: if $ra = 0$ for some $r \in R$, then $r = 0$ since R is an integral domain and $a \neq 0$. By convention, the empty set is taken to be a basis for (0), so the zero ideal has rank 0.) Now consider $N \leq M \cong R^n$ for some n. Let $p : R^n \to R^{n-1}$ be the map which projects onto the first $n-1$ coordinates:

$$p(x_1, x_2, \ldots, x_n) = (x_1, x_2, \ldots, x_{n-1}).$$

The image $p(N)$ of $N \leq M \cong R^n$ under this map is then a submodule of R^{n-1}, so since R^{n-1} is a finitely generated free module of rank smaller than $n = \operatorname{rank}(M)$, by induction we may assume

that $p(N) \leq R^{n-1}$ is free of rank $k \leq n-1$. The map $N \to p(N)$ induced by p is surjective, so a problem on the homework (free modules are *projective*) shows that N contains a submodule isomorphic to $p(N)$ and that we have the following decomposition:

$$N = [(\ker p) \cap N] \oplus p(N).$$

(Here, $(\ker p) \cap N$ is the kernel of the restricted map $N \to p(N)$ induced by p.)

The kernel of $p : R^n \to R^{n-1}$ consists of elements of the form $(0, \dots, 0, r)$ for $r \in R$, and hence is isomorphic to R. Hence $(\ker p) \cap N \leq \ker p \cong R$ is a submodule of a free module of rank 1, so by the base case $(\ker p) \cap N$ is also free of rank $\ell \leq 1$. Thus, we have

$$(\ker p) \cap N \cong R^\ell \text{ and } p(N) \cong R^k,$$

so

$$N = [(\ker p) \cap N] \oplus p(N) \cong R^\ell \oplus R^k \cong R^{\ell+k},$$

meaning that $N \leq M \cong R^n$ is free of rank $\ell + k \leq 1 + (n-1) = n$ as desired.

Lecture 24: Structure Theorems

Warm-Up. We verify a clam made last time, that $\langle x_1, x_2, \dots \rangle$ is not finitely generated as a module over the non-PID $\mathbb{R}[x_1, x_2, \dots]$, which thus gives an example of a non-finitely generated module of a finitely-generated module. ($\mathbb{R}[x_1, x_2, \dots]$ is finitely generated—by 1—as a module over itself. Note that $\langle x_1, x_2, \dots \rangle$ contains no nonzero constant polynomials.) Take any finite collection of nonzero polynomials in $\langle x_1, x_2, \dots \rangle$:

$$p_1(\mathbf{x}),\ p_2(\mathbf{x}), \dots,\ p_n(\mathbf{x})$$

where $\mathbf{x}$ denotes the infinite vector $(x_1, x_2, \dots)$. Then each $p_i(\mathbf{x})$ is nonconstant and built up out of only finitely many x_i. (By definition, a polynomial, even in an infinite number of variables, is still a sum of only finitely many terms, each of which involve products of only finitely many variables.)

Thus there are only finitely many x_i showing up in all the $p_j(\mathbf{x})$, so we can pick some x_N with the property that all variables showing up in all the $p_j(\mathbf{x})$ are among $x_1, x_2, \dots, x_{N-1}$. Then we claim that x_N is not a linear combination of the $p_j(\mathbf{x})$, which would show that $p_1(\mathbf{x}), \dots, p_n(\mathbf{x})$ do not generate all of $\langle x_1, x_2, \dots \rangle$, and hence that no finite subset can. If we did have

$$x_N = c_1(\mathbf{x})p_1(\mathbf{x}) + \cdots + c_n(\mathbf{x})p_n(\mathbf{x})$$

for some $c_j(\mathbf{x}) \in \mathbb{R}[x_1, x_2, \dots]$, then applying the linear map $\mathbb{R}[x_1, x_2, \dots] \to \mathbb{R}$ which sends any real number to itself, x_i to 0 for $1 \leq i \leq N-1$, and x_i to 1 for $i \geq N$ would produce 1 on the left side and 0 on the right, where we get 0 on the right since all the variables making up the $p_j(\mathbf{x})$ involve only those which are sent to 0 under this map. (It is important here that the $p_j(\mathbf{x})$ are not constant, to guarantee that all terms making up the expression for it are indeed sent to 0.) Thus the equality above would imply $1 = 0$ in $\mathbb{R}$, which is not true, so no such linear combination expressions exists. Hence $x_N \notin \langle p_1(\mathbf{x}), \dots, p_n(\mathbf{x}) \rangle$, and we are done.

Torsion-free implies free. We now prove that a finitely generated torsion-free module M over a PID R is free. Say that M has generators $M = \langle x_1, \dots, x_n \rangle$, and pick among them the largest linearly independent set of generators you can, say $x_1, \dots, x_k$ after renumbering. (Such a largest set exists since there are only finitely many subsets of the generators, so at least one must be "maximal". If you want a more elaborate way of deriving this, take the collection of all free modules with bases

among the x_i, and use the fact that M is Noetherian—it is finitely generated over a PID—to say that a maximal such free module exists, so that its basis is the maximal linearly independent set we want.) Since $\{x_1, \dots, x_k\}$ is maximally linearly independent, for each $k+1 \leq i \leq n$ the elements $x_1, \dots, x_k, x_i$ must be linearly dependent, so there exist $0 \neq r_i \in R$ and $r_1, \dots, r_k \in R$ such that

$$r_i x_i + r_1 x_1 + \cdots + r_k x_k = 0.$$

(Note that we know r_i must be nonzero, since if it were zero then one of the $r_1, \dots, r_k$ would be nonzero—by linear dependence of $x_1, \dots, x_k, x_i$—but then $r_1 x_1 + \cdots + r_k x_k = 0$ would contradict independence of $x_1, \dots, x_k$.) If we denote by F the free module generated by $x_1, \dots, x_k$, the equation above then gives

$$r_i x_i = -r_1 x_1 - \cdots - r_k x_k \in F.$$

Thus, for the ring element $r_{k+1} \cdots r_n \in R$, we have $r_{k+1} \cdots r_n M \subseteq F$: we have $r_{k+1} \cdots r_n x_i \in F$ for $1 \leq i \leq k$ since these x_i belong to F by definition of F, whereas

$$r_{k+1} \cdots r_n x_i = (r_{k+1} \cdots \widehat{r_i} \cdots r_k) r_i x_i \in F$$

for $k+1 \leq i \leq n$ since $r_i x_i \in F$ for these, so $r_{k+1} \cdots r_n M$ is contained in F since $r_{k+1} \cdots r_n x_i$ is in F for any generator x_i of M. Hence $r_{k+1} \cdots r_n M$ is a submodule of the finitely generated free module F, so it is free itself by the final fact we proved last time. Since M is torsion-free, the surjective map

$$M \to r_{k+1} \cdots r_n M \text{ defined by } x \mapsto r_{k+1} \cdots r_n x$$

is in fact injective (any nonzero element in the kernel would be a torsion element of M), so M is isomorphic to the free module $r_{k+1} \cdots r_n M$, and is thus free itself. (We will emphasize again that both the PID assumption on R and the finitely generated assumption on M are important here, since this result fails if either one of these conditions does not hold.)

Baby structure theorem. Over any integral domain R, the quotient $M/\operatorname{Tor}(M)$ is torsion-free. Indeed, if $x \in M/\operatorname{Tor}(M)$ is a torsion element, then there exists nonzero $r \in R$ such that $rx = 0$ in the quotient, which means that $rx \in \operatorname{Tor}(M)$. But this then in turns mean there exists nonzero $s \in R$ such that $s(rx) = 0$, so that $(sr)x = 0$ and $x \in \operatorname{Tor}(M)$ because $sr \neq 0$ by the integral domain property. Thus $x = 0$ in $M/\operatorname{Tor}(M)$, so $M/\operatorname{Tor}(M)$ is torsion-free.

With this at hand, we can state what I'll call the *baby structure theorem*, which is a simplified version of the full structure theorems we'll soon give. The claim is that if M is a finitely generated module over a PID, then M is isomorphic to a direct sum of a free module and $\operatorname{Tor}(M)$:

$$M \cong R^n \oplus \operatorname{Tor}(M)$$

where F is free. The proof is simple: since $M/\operatorname{Tor}(M)$ is torsion-free and finitely generated, it is free by the previous result, and hence for the surjective map $p : M \to M/\operatorname{Tor}(M)$ we have

$$M \cong (M/\operatorname{Tor}(M)) \oplus \ker p$$

by the problem on the current homework which says that free modules are *projective*, in this case applied to the free module $M/\operatorname{Tor}(M)$. (We also used this result in the proof that submodules of finitely generated free modules are free last time, and again will leave the details pertaining to the notion of "projective" to the homework.) The kernel of $M \to M/\operatorname{Tor}(M)$ is just $\operatorname{Tor}(M)$, so for the free module $R^n \cong M/\operatorname{Tor}(M)$ ($n = \operatorname{rank}(M/\operatorname{Tor}(M))$ we have $M \cong R^n \oplus \operatorname{Tor}(M)$ as desired.

The upshot is that any finitely generated module over a PID can be decomposed into a *free* part and a *torsion* part. The free part represents the largest free submodule of M, and its rank is also called the *rank* (or sometimes the *free rank*) of M. (There is a uniqueness claim—which we will not prove—to the "baby structure theorem" as well, that guarantees the free parts of any two such decompositions are isomorphic, so that the rank of M is well-defined.) Free modules R^n are simple enough to understand, so the study of such M reduces to the study of their torsion parts. The full structure theorems are then essentially statements about the form which these torsion parts can take. (We should note that the book does not discuss any of this until *after* it proves the general structure theorems, but we are taking a different approach where we use the "baby structure theorem" as a first-step towards the more general ones. I feel that this approach is more enlightening and better highlights the ingredients—PID, finite generation, Noetherianness—that go into it than the book's approach.)

Structure theorem, invariant factors. Let us now state the first form of the structure theorem, the *invariant factors* version: if M is a finitely generated module over a PID R, then there exist nonzero $a_1, a_2, \ldots, a_m \in R$ with $a_1 \mid a_2 \mid \ldots \mid a_m$ and $n \geq 0$ such that

$$M \cong R^n \oplus R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_m).$$

Here, R^n is the free part of M and $R/(a_1) \oplus R/(a_2) \oplus \cdots \oplus R/(a_m) \cong \mathrm{Tor}(M)$ the torsion part. Each of the free factors $R = \langle 1 \rangle$ are cyclic, as are the torsion factors $R/(a_i) = \langle 1 \bmod a_i \rangle$, so this is indeed a direct sum of cyclic modules, but with more information encoded. Note that each torsion factor is indeed a torsion module: a_i annihilates $R/(a_i)$, so $\mathrm{Tor}(R/(a_i)) = R/(a_i)$. Moreover there is an accompanying uniqueness statement: if

$$M \cong R^k \oplus R/(b_1) \oplus R/(b_2) \oplus \cdots \oplus R/(b_\ell)$$

is another such decomposition (with b_i dividing b_{i+1}), then $n = k$, $m = \ell$, and the a_i's are associate to the b_t's in some order. The ring elements $a_1, \ldots, a_n$ are called the *invariant factors* of M, and are thus unique up to associates, so that knowledge of the (free) rank and the invariant factors completely determines M.

The book proves this by using the fact that M is finitely generated to pick a surjective map $R^s \to M$ for some s, applying the First Isomorphism Theorem, and then finding "good bases" to use to describe R^s and the kernel. The "good bases" construction is the difficult and tedious part, spanning a few pages in the book. (This is how the book actually proves that submodules of finitely generated free modules over PIDs are free, by constructing the basis for the submodule at the same time. The proof we gave of this fact last time did not produce a basis, but has the benefit of being cleaner.) We will say a bit about the book's approach next time, but our approach will instead derive the invariant factors version of the structure theorem from the "elementary divisors" version, which we will prove (except for a few details) directly.

Structure theorem, elementary divisors. Consider one of the torsion factors $R/(a)$ in the invariant factor decomposition of M. Since R is a PID, it is a UFD, so we can write a in terms of its prime factorization: $a = p_1^{k_1} \cdots p_\ell^{k_\ell}$ with p_i distinct primes and $k_i > 0$. Since two distinct primes have greatest common divisor 1, we can express 1 as an R-linear combination of $p_i^{k_i}$ and $p_j^{k_j}$ for $i \neq j$, which implies that the sum of ideals $(p_i^{k_i}) + (p_j^{k_j}) = R$ is everything. Hence the ideals $(p_i^{k_i})$ are pairwise coprime, so the Chinese Remainder Theorem applies to give:

$$R/(a) = R/(p_1^{k_1} \cdots p_\ell^{k_\ell}) \cong R/(p_1^{k_1}) \oplus \cdots \oplus R/(p_\ell^{k_\ell}).$$

Applying this fact to each torsion factor $R/(a_i)$ in the invariant factors decomposition of M results in the second form of the structure theorem, the *elementary divisors* version: if M is a finitely generated module over a PID R, then there exist primes $p_1, \ldots, p_m$ (not necessarily distinct), $k_i > 0$ and $n \geq 0$ such that

$$M \cong R^n \oplus R/(p_1^{k_1}) \oplus \cdots \oplus R/(p_m^{k_m}).$$

The torsion part in this case is thus $R/(p_1^{k_1}) \oplus \cdots \oplus R/(p_m^{k_m})$. Note that the primes are not necessarily unique since, if we go back to the Chinese Remainder Theorem application to the invariant factors form, we see that a single prime p might show in the prime decomposition of more than one invariant factor a_i, and so such a prime will show up more than once in the decomposition above. Again we have a uniqueness statement attached to this, so that the rank n is unique and the prime powers (and number of primes) are unique up to associates. The $p_i^{k_i}$ are called the *elementary divisors* of M, and the elementary divisor decomposition above is also called the *primary decomposition*.

The Chinese Remainder Theorem can be run in reverse to derive the invariant factors form from this elementary divisors form. We will see how to do so next time, and indeed, as mentioned before, in our approach it is the elementary divisors form that we will actually (mostly) prove. One final observation here is that the torsion factor $R/(p_i^{k_i})$ is annihilated by the prime power $p_i^{k_i}$: in general, a module which is annihilated by a power of a prime p is said to be a *p-torsion module*, and we will discuss such things in more detail next time as we seek to prove this version of the structure theorem.

Finitely generated abelian groups. Applying this all in the case $R = \mathbb{Z}$ thus recovers the structure theorems for finitely generated abelian groups we stated at the end of last quarter: a finitely generated abelian group G can be decomposed as

$$G \cong \mathbb{Z}^r \times \mathbb{Z}/d_1\mathbb{Z} \times \cdots \times \mathbb{Z}/d_k\mathbb{Z}$$

where $d_i \mid d_{i+1}$, and as

$$G \cong \mathbb{Z}^r \times \mathbb{Z}/p_1^{k_1}\mathbb{Z} \times \cdots \times \mathbb{Z}/p_m^{k_m}\mathbb{Z}$$

where the p_i are primes in $\mathbb{Z}$. For instance, in

$$\mathbb{Z}/4\mathbb{Z} \times \mathbb{Z}/24\mathbb{Z} \cong \mathbb{Z}/4\mathbb{Z} \times (\mathbb{Z}/8\mathbb{Z} \times \mathbb{Z}/3\mathbb{Z}),$$

the left side is the invariant factor form and the right side the elementary divisors form. (This group can also be written as $\mathbb{Z}/8\mathbb{Z} \times \mathbb{Z}/12\mathbb{Z}$, but this is neither the invariant factors nor elementary divisors form.)

Lecture 25: More on Structure Theorems

Warm-Up. Suppose M is a finitely generated module over a PID R with elementary divisors

$$p_1,\ p_1^3,\ p_2^2,\ p_2^4,\ p_2^5,\ p_3,\ p_4^2,\ p_4^3$$

where p_1, p_2, p_3, p_4 are distinct primes of R. We derive the invariant factors of M from these. Recall from the "invariant factors $\implies$ elementary divisors" argument we gave last time that the elementary divisors come from the prime factorizations of the invariant factors via the Chinese Remainder Theorem. Thus, in this case the invariant factors should be built out of the specific prime powers given above, and the only requirement is to group them into factorizaitons which

satisfy the divisiblity requirement $a_1 \mid \ldots \mid a_m$ of the invariant factors. We claim that there are three invariant factors in this case, which comes from the fact that p_2 occurs in three elementary divisors, which is more than the number of elementary divisors corresponding to p_1, p_3, or p_4.

Set a_3 to be the element whose prime facotorization uses the *largest* powers of each prime occuring among the elementary divisors: $a_3 = p_1^3 p_2^5 p_3 p_4^3$. Next, take a_2 to consist of the *next* largest prime powers: $a_2 = p_1 p_2^4 p_4^2$. (There is no p_3 factor in a_2 since the largest power of p_3 was already used up to construct a_3.) Finally, take a_1 to consist of the next largest prime powers which remain, which in this case is just $a_1 = p_2^2$. The fact that at each step we took *smaller* prime powers than before guarantees the divisibility $a_1 \mid a_2 \mid a_3$ we want (p^a divides p^b if and only if $a \leq b$) so the invariant factors of M are:

$$a_1 = p_2^2 \quad a_2 = p_1 p_2^4 p_4^2 \quad a_3 = p_1^3 p_2^5 p_3 p_4^3.$$

The invariant factor decomposition of the torsion part of M is thus

$$\mathrm{Tor}(M) \cong R/(p_2^2) \oplus R/(p_1 p_2^4 p_4^2) \oplus R/(p_1^3 p_2^5 p_3 p_4^3),$$

and applying the Chinese Remainder Theroem to each factor here reproduces the elementary divisor (or primary) decomposition:

$$\mathrm{Tor}(M) \cong R/(p_2^2) \oplus R/(p_1) \oplus R/(p_2^4) \oplus R/(p_4^2) \oplus R/(p_1^3) \oplus R/(p_2^5) \oplus R/(p_3) \oplus R/(p_4^3).$$

This same reasoning works in complete generality, and shows that the invariant factor decomposition of any M can always be derived from its elementary divisor decomposition: the "largest" invariant factor a_m (i.e. the one occurring at the end of the divisibility chain $a_1 \mid \ldots \mid a_m$) is built from the largest prime powers for each prime that occur among the elementary divisors; the next "largest" invariant factor a_{m-1} uses the next largest prime powers; and so on. Thus, if we prove that the elementary divisor decomposition exists, we will have also proven that the invariant factor decomposition exists.

Good basis theorem. Let us say a word about the book's approach to the structure theorems, where it derives the invariant factor decomposition first. The techincal fact needed is the following, which I will call the *good basis theorem*:

> if M is a finitely generated free module over a PID R and $N \leq M$ is a submodule, then there exists a basis $x_1, \ldots, x_n$ for M and nonzero $a_1, \ldots, a_m \in R$ with $m \leq n$ such that $a_1 x_1, \ldots, a_m x_m$ is a basis for N and $a_1 \mid \ldots \mid a_m$.

(Note that $m \leq n$ here is not necessarily equal to n, so that the basis for the submodule only uses the first m things among $x_1, \ldots, x_n$.) The fact that N is itself free is something we proved ourselves a few lectures ago, but our proof did not produce an explicit basis, so the point here is that we have some explicit information about bases for M and N, both built up out of the same x_i. It is this relation between these specific bases for M and N that allows for the invariant factor decomposition to exist.

Now, the proof of this "good basis theorem" in the book is quite technical, so we will not reproduce it. From my perspective, the proof is not very englightening and obscures much of the underlying *structure* which allows for everything to work out nicely. (The elementary divisor approach we'll take better highlights this structure in my opinion.) The proof is a brute-force construction of the basis elements, one-at-a-time, where much care is needed to keep track of all the details. The proof is by no means overly complex and hard to follow, it's just that there's not much *joy* in it for my tastes.

But, if we assume the "good basis" result, let us see how to derive the invariant factor decomposition from it. Suppose now that M is a finitely generated module over R. Then, as we have seen before, M can be obtained a quotient of some R^s: if M has s generators, applying the First Isomorphism Theorem to the surjective map $R^s \to M$ which sends $e_i \in R^s$ to the i-th generator results in the isomorphism

$$R^s/\ker \cong M.$$

Since R^s is free and finitely generated over R, there exists a "good basis" $x_1, \dots, x_s$ for R^s and nonzero $a_1, \dots, a_m \in R$ ($m \leq s$) satisfying $a_1 \mid \dots \mid a_m$ such that $a_1x_1, \dots, a_mx_m$ is a basis for the kernel. With these we have:

$$M \cong R^s/\ker = \langle x_1 \rangle \oplus \cdots \oplus \langle x_s \rangle / \langle a_1x_1 \rangle \oplus \cdots \oplus \langle a_mx_m \rangle,$$

where the direct sums come from independence of the basis vectors. The natural projection map

$$\langle x_1 \rangle \oplus \cdots \oplus \langle x_m \rangle \to \langle x_1 \rangle / \langle a_1x_1 \rangle \oplus \cdots \oplus \langle x_m \rangle / \langle a_mx_m \rangle$$

has kernel $\langle a_1x_1 \rangle \oplus \cdots \oplus \langle a_mx_m \rangle$, so we get that $R^s/\ker$ is isomorphic to

$$\begin{aligned} R^s/\ker &= \langle x_1 \rangle \oplus \cdots \oplus \langle x_s \rangle / \langle a_1x_1 \rangle \oplus \cdots \oplus \langle a_mx_m \rangle \\ &\cong \langle x_1 \rangle / \langle a_1x_1 \rangle \oplus \cdots \oplus \langle x_m \rangle / \langle a_mx_m \rangle \oplus \langle x_{m+1} \rangle \oplus \cdots \oplus \langle x_s \rangle \end{aligned}$$

where $x_{m+1}, \dots, x_s$ are the "good basis" vectors of R^s that do not contribute to the basis for the kernel. Since

$$\langle x_i \rangle / \langle a_ix_i \rangle \cong R/(a_i) \text{ for } i \leq m \quad \text{and} \quad \langle x_j \rangle \cong R \text{ for } m < j \leq s,$$

this gives the invariant factor decomopsition of M:

$$M \cong R^s/\ker \cong \underbrace{R/(a_1) \oplus \cdots \oplus R/(a_m)}_{\text{torsion}} \oplus \underbrace{R \oplus \cdots \oplus R}_{s-m \text{ times}},$$

only with free part R^{s-m} written after the torsion part.

Euclidean domain approach. The fact that the "good basis" approach produces the invariant factors a_i explicitly is a nice benefit (our derivation of the elementary divisors to come will not produce these divisors in such an explicit way), but in practice these "explicit" factors are still difficult to compute in concrete situations in general. But one instance where it *is* feasible to compute these factors is in the case where R is a Euclidean domain.

Indeed, here is an alternate approach to the "good basis" theorem in this case. (This is outlined in some of the exercises in the book, although not ones you'll see on the current homework.) Again we take M to be a finitely generated free module over R and $N \leq M$ a submodule. Suppose we already know that N has to be finintely generated, say as a consequence of the fact that M is Noetherian. Let $x_1, \dots, x_n$ be a basis for M and $y_1, \dots, y_k$ a set of generators for N. Since each y_i is in M, we can write each as an R-linear combination of $x_1 \dots, x_n$:

$$y_i = d_{i1}x_1 + \cdots + = d_{in}x_n \text{ for some } d_{i1}, \dots, d_{in} \in R.$$

Form an $k \times n$ matrix A by taking all of these d_{ij} as entries:

$$A = \begin{bmatrix} d_{11} & \cdots & d_{1n} \\ \vdots & \ddots & \vdots \\ d_{m1} & \cdots & d_{mn} \end{bmatrix}.$$

The fact is that then, through a series of row and column operations which make use of the division algorithm, this matrix can be reduced to one in the following form:

$$A \to \begin{bmatrix} a_1 & & & \\ & \ddots & & \\ & & a_m & \\ & & & \text{ZERO} \end{bmatrix}$$

where $a_i \mid a_{i+1}$, ZERO denotes an entire zero $(k-m) \times (n-m)$ matrix, and all missing entries are zero. These a_i are the invariant factors we want, where the point is that the effect of the row and column operations needed is to manipulate the basis $x_1, \ldots, x_n$ and generators $y_1, \ldots, y_k$ to produce a "good basis": row operations manipulate the y_j's, and column operations the x_i's. The upshot is that with the division algorithm at hand, an explicit good basis can be found using row and column operations, and thus the explicit invariant factors can be computed algorithmically. (This is the computational basis behind software programs which can actually compute these things. The reduced matrix above is called the *Smith Normal Form* of A, and has other applications as well.) Finding the invariant factors over PIDs which are not Euclidean domains is a tougher ordeal, hence the tougher proof of the "good basis" theorem in general.

p-torsion parts. We now move towards our "elementary divisors" approach to the structure theorems. As stated earlier, I feel that this approach is conceptually "nicer" and does a better job of highlighting the underlying structure involved. The drawback is that, in the end, we do not get an easy explicit way of computing the elementary divisors, as opposed to, say, the Euclidean domain approach to the invariant factors. Nevertheless, at least for the applications we will see in the final days, we will see that the elementary divisors *can* be described explicitly in terms of linear algebraic data, so all is not lost.

Our approach begins with the "baby structure theorem", where we have

$$M \cong R^n \oplus \text{Tor}(M).$$

So, we can focus our attention only on the torsion part, and thus reduce to the case where M itself is torsion. The goal is to decompose M using elementary divisors, and as a first step we show that we can decompose it into "p-torsion" parts: for a prime $p \in R$, the *p-torsion* submodule M_p of M (also called the *p-primary submodule*) is the submodule consisting of those elements which are annihilated by a power of p:

$$M_p := \{x \in M \mid p^k x = 0 \text{ for some } k\}.$$

We know that, since M is torsion, every element of M is annhilated by something, and so here we are "breaking off" those elements annhilated by some p^k. The first result is that M can be decomposed as a direct sum of its (finitely many) p-torsion parts.

But before we prove this, let us think about it from the perspective of the elementary divisors structure theorem. For instance, if we consider the example of the Warm-Up (let us now take M to be fully torsion, so that $M = \text{Tor}(M)$), we have

$$M \cong R/(p_2^2) \oplus R/(p_1) \oplus R/(p_2^4) \oplus R/(p_4^2) \oplus R/(p_1^3) \oplus R/(p_2^5) \oplus R/(p_3) \oplus R/(p_4^3)$$

as the elementary divisor decomposition of M. If we group here the terms which correspond to the same prime, we obtain the p-torsion submodule decomposition:

$$M \cong \underbrace{R/(p_1) \oplus R/(p_1^3)}_{M_{p_1}} \oplus \underbrace{R/(p_2^2) \oplus R/(p_2^4) \oplus R/(p_2^5)}_{M_{p_2}} \oplus \underbrace{R/(p_3)}_{M_{p_3}} \oplus \underbrace{R/(p_4^2) \oplus R/(p_4^3)}_{M_{p_4}}.$$

Indeed, the factors corresponding to p_i are all annihilated by some power of p_i. (In fact, the largest power of p_i is enough to annihilate them all: for instance, p_2^5 will annihilate the entire p_2-torsion submodule.) Thus the elementary divisor struture theorem produces the p-torsion decomposition, so we are essentially proving this structre theorem by going in the reverse direction: decompose M into its p-torsion parts, and then argue that each p-torsion part can be broken down in terms of elementary divisors by focusing only on the behavior of one prime at-a-time.

p-torsion decomposition. We now prove that any finitely generated torsion module M over a PID R can be written as a direct sum of its p-torsion submodules. First, we take it for granted that $\text{Ann}(M) \neq 0$, which you will prove on the final homework. (We know that each element of M is annihilated by some nonzero ring element since M is torsion, but the point here is that we can find a nonzero element of R which annihilates all things in M simultaneously.) Since $\text{Ann}(M)$ is then a nonzero ideal of R and R is PID, we have $\text{Ann}(M) = (a)$ for some $0 \neq a \in R$. Let $a = p_1^{k_1} \cdots p_m^{k_m}$ be the prime factorization of a. We claim specifically that

$$M = M_{p_1} \oplus \cdots \oplus M_{p_m}.$$

We proceed by induction on the number of prime factors of a. If $a = p_1^{k_1}$ only has one prime factor, then $M = M_{p_1}$ is already its own p_1-torsion submodule since $a = p_1^{k_1}$ annihilates everything in M. Otherwise, $a = p_1^{k_1} b$ where $b = p_2^{k_2} \cdots p_m^{k_m}$ is relatively prime to $p_1^{k_1}$. Let N be the submodule of M which is annihilated by b (i.e. $(b) = \text{Ann}(N)$). We claim that $M = M_{p_1} \oplus N$, in which case, since b has one prime factor less than a, we may assume that N decomposes as

$$N = M_{p_2} \oplus \cdots \oplus M_{p_m}$$

by induction, giving $M = M_{p_1} \oplus M_{p_2} \oplus \cdots \oplus M_{p_m}$ as desired. To see that $M = M_{p_1} \oplus N$, note that since $p_1^{k_1}$ and b are relatively prime we have $(p_1^{k_1}, b) = 1$, so there exist $s, t \in R$ such that

$$1 = tb + sp_1^{k_1}.$$

For $x \in M$, this gives

$$x = 1x = \underbrace{tbx}_{\in M_{p_1}} + \underbrace{sp_1^{k_1}x}_{\in N},$$

where $tbx \in M_{p_1}$ since it is annihilated by $p_1^{k_1}$ (recall that $p_1^{k_1} b = a$ generates the annihilator of M) and $sp_1^{k_1}x$ is in N since it is annihilated by b. Thus $M = M_{p_1} + N$. If $x \in M_{p_1} \cap N$, then it is annihilated by b and some p_1^{ℓ}, which are relatively prime, so for $1 = bf + p_1^{\ell} g \in R$ we have

$$x = 1x = (bf + p_1^{\ell} g)x = f(bx) + g(p_1^{\ell} x) = 0 + 0 = 0.$$

Thus $M_{p_1} \cap N = \{0\}$, so $M = M_{p_1} \oplus N$ is a direct sum, and as described above induction then gives $M = M_{p_1} \oplus \cdots \oplus M_{p_m}$ as claimed.

The study of finitely generated torsion modules over PIDs thus reduces to the case of a p-torsion module, whose structure we will determine at the start of next time and hence derive the elementary divisor structure theorem as a result. One final observation is that we actually saw a special case of this "p-primary decomposition" last quarter: the decomposition of a finite abelian group (which is a torsion $\mathbb{Z}$-module) as a product of its Sylow subgroups! Indeed, the p-torsion submodule in this case is the set of elements of order a power of p, which is precisely the (unique) Sylow p-subgroup. You can view the general p-torsion decomposition result as a generalization of this fact, and view p-torsion submodules in general as generalizations of Sylow p-subgroups.

Lecture 26: Jordan Normal Form

Warm-Up. (This Warm-Up is a bit disconnected from the immediate work we did last time, but forms a key step in finishing off the proof of the elementary divisors structure theorem.) Suppose M is an R-module, $x \in M$, and that $M/\langle x \rangle$ decomposes as a direct sum of cyclic modules:

$$M/\langle x \rangle = \langle y_1 \rangle \oplus \cdots \oplus \langle y_m \rangle$$

for some $y_i \in M/\langle x \rangle$. Suppose further that for each y_i there exists $z_i \in M$ with $z_i = y_i \bmod \langle x \rangle$ (so z_i gives the element y_i in the quotient $M/\langle x \rangle$) such that the annihilator of z_i in M is the same as the annihilator of y_i in $M/\langle x \rangle$. We claim that the map $\phi : M/\langle x \rangle \to M$ defined by $y_i \mapsto z_i$ and extending linearly is well-defined.

First, let us give an example of where this type of thing fails, so that we can better appreciate the actual meaning of this result. Consider the quaternion group Q_8 from last quarter. Then the quotient $Q_8/\langle i \rangle$ is isomorphic to $\mathbb{Z}/2\mathbb{Z}$:

$$Q_8 \to Q_8/\langle i \rangle \cong \mathbb{Z}/2\mathbb{Z},$$

so it does "decompose as a product of cyclic groups". Then $1 \in \mathbb{Z}/2\mathbb{Z}$ has order 2—think of this as saying that 1 is "annihilated" by 2, even though Q_8 is technically not a $\mathbb{Z}$-module since it is not abelian—but all of the elements $\pm j, \pm k$ in Q_8 which give $1 \in \mathbb{Z}/2\mathbb{Z}$ in the quotient have order 4 and are not "annihilated" by 2. Thus, in this case the "annihilator" of an element in Q_8 vs its "annihilator" in the quotient can be different. In the module setting above, the point is that if $rz_i = 0$ in M, so that r is in the annihilator of z_i in M, then certainly $ry_i = 0$ in $M/\langle x \rangle$ as well since equalities are maintained when passing to quotients. Hence r will also annihilate the element we get in the quotient, but the the point is that the converse need not be true: an element which annihilates y_i in the quotient does not necessarily annihilate z_i in M. That is, the annihilator might "enlarge" when passing to the quotient, as it what happens in the Q_8 example: the "annihilator" of j, say, in Q_8 is $4\mathbb{Z}$, but the annihilator of $j \in Q_8/\langle i \rangle$ (which corresponds to $1 \in \mathbb{Z}/2\mathbb{Z}$) is $2\mathbb{Z}$ instead. The assumption that the annihilator of z_i in M is the same as the annihilator of y_i in $M/\langle x \rangle$ in the setup of this Warm-Up says that this cannot happen, so that annihilators are preserved when passing to the quotient.

Now, the map $M/\langle x \rangle \to M$ we are attempting to construct is defined by its action on the generators y_i, but in order for "extending linearly" to make sense in this case requires that any linear relation among the y_i which exists in the quotient be preserved when moving into M. If we again consider the Q_8 example, the map we are attempting to define would be something like $\psi : \mathbb{Z}/2\mathbb{Z} \cong Q_8/\langle i \rangle \to Q_8$, where we might try to set

$$\psi(0) = 1 \quad \text{and} \quad \psi(1) = j,$$

with 1 and j in Q_8 being elements which give 0 and 1 respectively in $Q_8/\langle i \rangle \cong \mathbb{Z}/2\mathbb{Z}$. But we then run into the problem that $\psi(1)\psi(1)$ would have to equal both $\psi(1+1) = \psi(0) = 1$ and $jj = -1$ in Q_8, which is not true. The issue is the relation $1 + 1 = 0$ in the quotient is not preserved under ψ, so that ψ is in fact not well-defined. (In other words, the equal elements $1 + 1$ and 0 in $Q_8/\langle i \rangle$ are sent to the unequal things -1 and 1 respectively.) There is no way we can define ψ in a consistent way to get an actual homomorphism. In our module problem, what we thus need to know is that any linear relation $r_1 y_1 + \cdots + r_m y_m = 0$ among the y_i's is maintained when applying ϕ, since otherwise we would not be able to define the $\phi(y_i)$ independently of one another as we have done when setting $\phi(y_i) = z_i$. (In other words, we have to know that anything which is zero in the quotient remains zero in M.)

So, suppose $r_1y_1 + \cdots + r_my_m = 0$ in $M/\langle x\rangle$ for some $r_i \in R$. Since $M/\langle x\rangle = \langle y_1\rangle \oplus \cdots \oplus \langle y_m\rangle$ is a direct sum, we must have $r_iy_i = 0$ for each i. This means that r_i is in the annihilator of y_i, so by our assumption on z_i, we have that r_i is in the annihilator of z_i as well, so $r_iz_i = 0$ for each i. Thus:

$$\phi(r_1y_1 + \cdots + r_my_m) = r_1\phi(y_1) + \cdots + r_m\phi(y_m) = r_1z_1 + \cdots + r_mz_m = 0 + \cdots + 0 = 0,$$

so that $r_1y_1 + \cdots + r_my_m = 0$ is indeed maintained after applying ϕ, and hence ϕ as constructed is well-defined and gives an R-linear map $M/\langle x\rangle \to M$. (In the Q_8, the underlying issue is that $1 \in \mathbb{Z}/2\mathbb{Z}$ is annhilated by 2 but j, say, is not: the relation $2 \cdot 1 = 0$ in the quotient becomes $j^2 = -1$ in Q_9, so this relation is not preserved.)

The *true* point of this result is the following: by definition of $\phi : M/\langle x\rangle \to M$, if $p : M \to M/\langle x\rangle$ denotes the natural projection, then:

$$p(\phi(y_i)) = p(z_i) = y_i \text{ for each } i,$$

which implies that $p \circ \phi$ is the identity on the quotient since the y_i generate the quotient. Thus $M/\langle x\rangle$ satisfies a "projective module" type of property, and the same result about projective modules proved on the homework implies in this case that

$$M = \langle x\rangle \oplus M/\langle x\rangle,$$

where we think of $M/\langle x\rangle \cong \langle z_1, \ldots, z_m\rangle$ as a submodule of M. (Note that the kernel of the projection p is $\langle x\rangle$.) So, the upshot is that M decomposes as a direction sum of a quotient and the submodule by which we took the quotient. If we go back to the Q_8 case, the fact that no analogous results holds in this case is essentially the reason why Q_8 is not obtainable as a semidirect product of a subgroup of order 4 with a subgroup of order 2: expressing Q_8 as a such a semidirect product essentially requires realizing $Q_8/\langle i\rangle$ as being isomorphic to a subgroup of Q_8 in a way which is compatible with the projection $Q_8 \to Q_8/\langle i\rangle$. (This came up last quarter when classifying groups of order 8, which is the simplest case where semidirect products methods alone are not enough.)

Elementary divisors. Now we complete our proof of the elementary divisors form of the structure theorem. Recall what we know so far: if M is a finitely generated module over a PID R, then M decomposes into a free part and a torsion part as $M \cong R^n \oplus \mathrm{Tor}(M)$ for some $n \geq 0$, and the torsion part then decomposes as a direct sum of its p-torsion submodules. What remains is to understand the structure of these p-torsion modules.

So, suppose that M is p-torsion, which means that all of its elements are annihilated by a power of a prime $p \in R$. The claim is that M then decomposes as

$$M \cong R/(p^{\ell_1}) \oplus \cdots \oplus R/(p^{\ell_t})$$

for some ℓ_i. Since M is finitely generated, if we consider the smallest powers of p which annihilate each generator, we can take the maximal power which occurs as an annihilator of all the generators, and thus of all of M, so $\mathrm{Ann}(M) = (p^k)$ for some k. Take x_1 to be the generator for which p^k was the smallest power which annihilated it, so that $\mathrm{Ann}(M) = \mathrm{Ann}(x_1) = (p^k)$. If x_1 alone generates all of M, then we are done: the map $R \to M = \langle x_1\rangle$ defined by $r \mapsto rx_1$ is surjective with kernel $\mathrm{Ann}(x_1) = (p^k)$, so $R/(p^k) \cong \langle x_1\rangle = M$.

Otherwise M has more than generator, so say that $m > 1$ is the minimal number of generators needed. Then $M/\langle x_1\rangle$ is generated by $m - 1$ elements, so by induction we may assume that our result holds for this module:

$$M/\langle x\rangle \cong R/(p^{\ell_1}) \oplus \cdots \oplus R/(p^{\ell_m})$$

for some ℓ_i. (The base case for the induction was the $M = \langle x_1 \rangle$ case above.) Denote the generators of each of these cyclic factors by $y_i \in M/\langle x_1 \rangle$, so $\langle y_i \rangle \cong R/(p^{\ell_i})$. We claim that, as in the setup of the Warm-Up, we can find $z_i \in M$ mapping to $y_i \in M/\langle x \rangle$ under the natural projection, for which the annihilator of z_i in M is the same as the annihilator of y_i in $M/\langle x_1 \rangle$. If so, the result of the Warm-Up immediately gives

$$M = \langle x_1 \rangle \oplus M/\langle x_1 \rangle,$$

which when combined with the previous isomorphisms stated above gives

$$M = \langle x_1 \rangle \oplus M/\langle x_1 \rangle \cong R/(p^k) \oplus R/(p^{\ell_1}) \oplus \cdots \oplus R/(p^{\ell_t}),$$

which is our desired result. Applying this to each p-torsion part of $\mathrm{Tor}(M)$ (M now a general finitely generated R-module, not necessarily simply a p-torsion one) gives the elementary divisors decomposition of M, and we thus have our structure theorem.

To construct the desired z_i, suppose p^s is the smallest power of p that annihilates y_i in $M/\langle x_1 \rangle$, and hence generates the annihilator of y_i in the quotient. If we consider y_i as an element of M itself, then the fact that $p^s y_i = 0$ in the quotient $M/\langle x_1 \rangle$ means that $p^s y_i = a x_1$ in M for some $a \in R$. But y_i in M is also annihilated by p^k (the power which annihilates all of M), so $p^{k-s} a x_1 = p^{k-s} p^s y_i = p^k y_i = 0$. The ring element $p^{k-s}a$ thus annihilates x_1, so since the annihilator of x_1 is generated by p^k, it must be the case that $p^{k-s}a = p^k b$ for some $b \in R$, which by cancellation in an integral domain implies that $a = p^s b$. We claim that $z_i := y_i - b x_1 \in M$ is then element we need. First, since z_i differs from y_i by an element of $\langle x_1 \rangle$, z_i does give the same element as y_i in $M/\langle x_1 \rangle$. (This in particular implies that anything which annihilates z_i in M will annihilate y_i in the quotient.) But also, in M we have: $p^s z_i = p^s(y_i - b x_1) = p^s y_i - p^s b x_1 = a x_1 - a x_1 = 0$, so anything which annihilates y_i in the quotient also annihilates z_i in M. Hence the annihilators of z_i and y_i in M or the quotient respectively are the same as required. (This is a bit of tedious argument, which we skipped in class. But to make one final comment: if you go back to the proof we gave on the last day of class last quarter of the structure theorem for *finite* abelian groups, you will see that the proof there is *very* similar to this proof here, only phrased in the language of groups instead of modules. Of course, this is no accident, since the proof there is essentially a special case of this proof here via the fact that we can view finite abelian groups as torsion $\mathbb{Z}$-modules.)

Our structure theorem (elementary divisors) is now complete, at least as far as existence is concerned. (We will not prove uniqueness in full, but there is a problem on the final homework which works out a special case, namely the case where the module is torsion. If we accept that the free part is unique—i.e. the rank of M is well-defined—this gives uniqueness in general.) As stated earlier, running the Chinese Remainder Theorem in "reverse" then gives the invariant factors structure theorem, completing the circle. Before moving on, let us simply note all the machinery which went into deriving these results, where most everything (but not everything) we looked at this quarter showed up in one form or another: rings, ideals, quotients, isomorphism theorems, Chinese Remainder Theorem, PIDs, primes, UFDs, and all of module theory. The one main topic missing from this discussion so far is that of polynomial rings, but lo-and-behold that is where we will finish our journey this quarter, by applying the structure theorem to the case of modules over polynomial rings over fields.

The structure of polynomial modules. Our goal in the remaining time is to prove the existence of what's called the *Jordan normal form* (or *Jordan canonical form*) of a linear operator on a finite-dimensional vector space, at least under an assumption on the base field we will clarify later. This will come from interpreting the result of the elementary divisor structure theorem in the case of an $F[x]$-module, where F is a field. (The invariant factor structure theorem produces what's called the *rational canonical form* of a linear operator, which we will say a few words about as well.)

So, suppose V is a finite-dimensional vector space over a field F, with a linear *operator* $T : V \to V$. (A linear operator is just a linear map from the space to itself. In the case where $F = \mathbb{R}$ and $V = \mathbb{R}^n$, this is just an $n \times n$ matrix A. Indeed, if you have not seen abstract vector spaces before beyond thinking about them as modules over fields as we have described them in this course, the case of $\mathbb{R}^n$ with an $n \times n$ matrix acting on it will be good enough when it comes to understanding what all the hoopla is about.) As we described previously, this all gives the data of a finitely generated $F[x]$-module structure on V, where $x \in F[x]$ acts on V as T, and polynomials in x acts as polynomials in T. Since F is a field, $F[x]$ is a PID, so the structure theorem applies to V.

A first observation is that, as an $F[x]$-module, V is torsion and so has no free part. We saw this previously in a Warm-Up explicitly in the $V = \mathbb{R}^2$ case (Lecture 21), and the general argument runs as follows. Pick a nonzero $v \in V$. If $\dim V = n$ as a vector space over F, then

$$v,\ Tv,\ T^2v, \dots,\ T^nv$$

must be linearly dependent, as are any $n+1$ vectors in an n-dimensional vector space. Thus there exist $a_0, \dots, a_n \in F$, at least one of which is nonzero, such that

$$a_0v + a_1Tv + \cdots + a_nT^nv = 0.$$

But the left side is the same as $(a_0I + a_1T + \cdots + a_nT^n)v$, which is precisely the action of the polynomial $a_0 + a_1x + \cdots + a_nx^n \in F[x]$ on v. This nonzero polynomial (recall that some a_i is nonzero) thus annihilates v, so v is a torsion element.

The elementary divisor structure theorem (with no free part) thus produces:

$$V \cong F[x]/(p_1(x)^{k_1}) \oplus \cdots \oplus F[x]/(p_m(x)^{k_m})$$

for some irreducible (i.e. prime) polynomials $p_i \in F[x]$ and $k_i \geq 1$. Now, we will assume that the irreducible polynomials which appear here are all linear: $p_i(x) = x - \lambda_i$ for some $\lambda_i \in F$. This is guaranteed to be true over $\mathbb{C}$ for instance, where the fact that $\mathbb{C}$ is "algebraically closed"—a concept we will study next quarter—ensures that only linear polynomials are irreducible. It is not true over $\mathbb{R}$ for instance, where irreducible polynomials can also be quadratic like $x^2 + 1$. The Jordan normal form we will derive thus always exists over $\mathbb{C}$ (or any algebraically closed field), but existence over other fields (such as $\mathbb{R}$) depends on the $p_i(x)$ having this form; soon we will clarify exactly when this holds and when it does not in terms of the operator T itself. Under this assumption, the decomposition above thus looks like:

$$V \cong F[x]/(x - \lambda_1)^{k_1} \oplus \cdots \oplus F[x]/(x - \lambda_m)^{k_m}.$$

Here comes the eigenstuff. Let us focus on a single factor $F[x]/(x - \lambda)^k$ above, viewed as a vector space over F. We claim that the elements $1, x-\lambda, (x-\lambda)^2, \dots, (x-\lambda)^{k-1} \in F[x]$ form a basis for this space. First note that we can describe elements in this quotient using only polynomials of degree at most $k-1$ since the relation $(x - \lambda)^k = 0$ in the quotient allows us to reduce any higher degree term. Now, the elements

$$1, x - \lambda, (x - \lambda)^2, \dots, (x - \lambda)^{k-1}$$

are linearly independent in the quotient, since a linear dependence relation among them:

$$a_0 + a_1(x - \lambda) + \cdots + a_{k-1}(x - \lambda)^{k-1} = 0$$

would require that $x - \lambda$ be a root of some polynomial of degree at most $k-1$, which is not possible since $(x - \lambda)^k = 0$ is the smallest degree expression which becomes zero in the quotient. The

space $F[x]/(x-\lambda)^k$ is k-dimensional since $1, x, x^2, \ldots, x^{k-1}$ for sure form a basis, so the k linearly independent elements listed above also form a basis since any k linearly independent elements of a k-dimensional vector space automatically span that space. (You can also see that any polynomial of degree at most $k-1$ can be written as a linear combination of $1, x-\lambda, (x-\lambda)^2 k, \ldots, (x-\lambda)^{k-1}$ by taking its $(k-1)$st-order Taylor expansion centered at λ.)

We now determine the action of $x \in F[x]$ on this basis. In fact, let us write this basis in the opposite order:

$$(x-\lambda)^{k-1}, (x-\lambda)^{k-2}, \ldots x-\lambda, 1.$$

If we write x as $x = (x-\lambda)+\lambda$, we have the following:

$$\begin{aligned}
x \cdot (x-\lambda)^{k-1} &= [(x-\lambda)+\lambda](x-\lambda)^{k-1} = (x-\lambda)^k + \lambda(x-\lambda)^{k-1} = \lambda(x-\lambda)^{k-1} \\
x \cdot (x-\lambda)^{k-2} &= [(x-\lambda)+\lambda](x-\lambda)^{k-2} = (x-\lambda)^{k-1} + \lambda(x-\lambda)^{k-2} \\
&\vdots \\
x \cdot (x-\lambda) &= [(x-\lambda)+\lambda](x-\lambda) = (x-\lambda)^2 + \lambda(x-\lambda) \\
x \cdot 1 &= [(x-\lambda)+\lambda]1 = (x-\lambda) + \lambda \cdot 1,
\end{aligned}$$

where in the first line we use the fact that $(x-\lambda)^k = 0$ in the quotient. Here are the key observations: $(x-\lambda)^{k-1}$ is sent to λ times itself, meaning that this element is an *eigenvector* of the action of x with *eigenvalue* λ; and every other basis vector is sent to the sum of the previous basis vector and λ times itself. If we translate this via the isomorphism $V \cong F[x]/(x-\lambda_1)^{k_1} \oplus \cdots \oplus F[x]/(x-\lambda_m)^{k_m}$ into a statement about V, these basis elements for the one factor we are considering become linearly independent vectors $v_1, v_2, \ldots, v_k$ in V such that $Tv_1 = \lambda v_1$ (so v_1 is an eigenvector of T with eigenvalue λ) and $Tv_i = v_{i-1} + \lambda v_i$ for $i > 1$.

Before moving on to consider what this all says about V and T, note we now know that the scalar λ appearing in the elementary divisor $(x-\lambda)^k$ for the $F[x]$-module V must in fact be an eigenvalue of T. Going back to the assumption we made about the primes $p_i(x)$ showing up in decomposition

$$V \cong F[x]/(p_1(x)^{k_1}) \oplus \cdots \oplus F[x]/(p_m(x)^{k_m}),$$

namely that they were are all linear $p_i(x) = x - \lambda_i$, we now see that this is true precisely when all eigenvalues of T lie in the field F, so it is under this condition that the Jordan normal form of T will exist. This is always the case when F is "algebraically closed", but for other fields it depends on the particulars of the linear operator T. For instance, if $A = \left[\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right]$ is the operator which describes counterclockwise rotation by $\pi/2$ around the origin in $\mathbb{R}^2$, the eigenvalues $\pm i$ of A do not exist in $\mathbb{R}$, so A does not have a Jordan normal form over $\mathbb{R}$—it only has one over $\mathbb{C}$.

Jordan normal form. So, one factor $F[x]/(x-\lambda)^k$ appearing the primary decomposition of V as an $F[x]$-module gives a subspace of V which has a basis $v_1, \ldots, v_k$ satisfying

$$Tv_1 = \lambda v_1 \quad \text{and} \quad Tv_i = v_{i-1} + \lambda v_i \text{ for } i > 1.$$

We now consider the matrix of the restriction of T to this subspace with respect to this basis. As a reminder (or introduction if you have not seen this notion before), the idea is that when we express an element of this space in terms of this given basis:

$$v = c_1 v_1 + \cdots + c_k v_k,$$

the *matrix of* (the restriction of) T *relative to this basis* is the matrix which describes the effect which T has on the coefficients $c_1, \ldots, c_k$:

$$(\text{matrix of } T) \begin{bmatrix} \text{coefficient} \\ \text{vector} \\ \text{of } v \end{bmatrix} = \begin{bmatrix} \text{coefficient} \\ \text{vector} \\ \text{of } Tv \end{bmatrix}.$$

Another way of saying this is that using the isomorphism $\phi : \text{span}(v_1, \ldots, v_k) \to F^k$ which the basis gives us (send a vector to its coefficient vector), the composition $\phi T \phi^{-1}$ is an isomorphism $F^k \to F^k$, which is thus described by a $k \times k$ matrix, and this is the matrix of (the restriction of) T. All information about T can be derived from this one matrix alone, so describing this matrix is equivalent to describing (the restriction of) T.

For v expressed as above, we have:

$$\begin{aligned} Tv &= c_1 Tv_1 + c_2 Tv_2 + c_3 Tv_3 + \cdots + c_k Tv_k \\ &= c_1 \lambda v_1 + c_2(v_1 + \lambda v_2) + c_3(v_2 + \lambda v_3) + \cdots + c_k(v_{k-1} + \lambda v_k) \\ &= (\lambda c_1 + c_2)v_1 + (\lambda c_2 + c_3)v_1 + \cdots + \lambda c_k v_k. \end{aligned}$$

The coefficients of Tv are thus related to those of v via the following equality:

$$\begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_{k-1} \\ c_k \end{bmatrix} = \begin{bmatrix} \lambda c_1 + c_2 \\ \lambda c_2 + c_3 \\ \vdots \\ \lambda c_{k-1} + c_k \\ \lambda c_k \end{bmatrix},$$

so the matrix on the left is the matrix of T with respect to $v_1, \ldots, v_k$. (This matrix is *almost* diagonal with λ down the diagonal, with an additional diagonal of 1's directly above the main diagonal. The missing entries should be interpreted as zeroes.) The idea is that the first column alone describes the effect T has on v_1, where the single λ in the first column in the first location reflects the fact that $Tv_1 = \lambda v_1$ (i.e. v_1 is an eigenvector with eigenvalue λ), the second column describes the effect of T on v_2, so that the 1 in the first location and λ in the second means "take $1 \cdot v_1$ plus λv_2 to get $Tv_2 = v_1 + \lambda v_2$", and so on for the effect of T on the remaining basis vectors: the entries, and the locations where they occur, in the ith column describe the effect of T on v_i.

A matrix of the form above is said to be a *Jordan block* of size $k \times k$. We can find such a matrix for the restriction of T to each of the factors in

$$V \cong F[x]/(x - \lambda_1)^{k_1} \oplus \cdots \oplus F[x]/(x - \lambda_m)^{k_m},$$

so that we get m Jordan blocks of sizes $k_1 \times k_1, k_2 \times k_2$, and so on. Putting together the bases for all elementary divisor factors above gives a basis for all of V, and the matrix of T on V relative to this basis is then of the form

$$\begin{bmatrix} J_1 & & & \\ & J_2 & & \\ & & \ddots & \\ & & & J_m \end{bmatrix}$$

where each J_i is a Jordan block. (Again, treat omitted entries as zeroes.) This matrix—whose existence we have now proven via the elementary divisors structure theorem, under the assumption

that the eigenvalues of T are all in F—is called the *Jordan normal form* (or *Jordan canonical form* of T. By the uniqueness of the elementary divisors, this Jordan normal form is unique up to a rearrangement of the Jordan blocks. We will say more about this construction, as well as look at one or two concrete examples and some applications, next time.

Lecture 27: More on Normal Forms

Warm-Up 1. Suppose V is a finitely generated $F[x]$-module, F a field, where the action of x is given by $T : V \to V$, and that F contains all eigenvalues of T. We determine the elementary divisor decomposition of V in the case where T is *diagonalizable*. To be diagonalizable means that there exists a basis (over F) of V consisting of eigenvectors of T, or equivalently a basis for V with respect to which the matrix of T is diagonal:

$$\begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}.$$

This is then the Jordan normal form of T, which in this case is made up of 1×1 Jordan blocks $\begin{bmatrix} \lambda_i \end{bmatrix}$ whose entries are the eigenvalues of T. From last time, each such Jordan block corresponds to one of the cyclic factors $F[x]/(x-\lambda)^k$ in the elementary divisors decomposition, so the decomposition in this case looks like

$$V \cong F[x]/(x-\lambda_1) \oplus \cdots \oplus F[x]/(x-\lambda_n)$$

with elementary divisors all of degree 1. Each of these factors has a single basis vector, which is necessarily an eigenvector of T, and putting them all together gives the basis for V consisting of eigenvectors of T.

Note that the factor $F[x]/(x-\lambda)$ here is annihilated by $x-\lambda$. In terms of T, this corresponds to elements of V such that $(T-\lambda I)v = 0$, which is just a rewritten form of the eigenvector equation $Tv = \lambda v$. Thus, factors in the elementary divisors decomposition of any V which correspond to *linear* elementary divisors give eigenvectors of T in the normal sense. For a more general elementary divisor factor $F[x]/(x-\lambda)^k$ with $k > 1$, the first power of $x-\lambda$ is not enough to annihilate everything, so not all elements in this factor are ordinary eigenvectors—only those which *are* annihilated by $x-\lambda$ are. But, all elements do satisfy $(x-\lambda)^k v = 0$, which in terms of T says $(T-\lambda I)^k v = 0$. Such vectors, for which possibly a higher power of $T-\lambda I$ is needed to annihilate them, are called *generalized eigenvectors* of T with eigenvalue λ. The basis of V which gives rise to the Jordan normal form of T thus consists of generalized eigenvectors of T. The subspace of V consisting of all generalized eigenvectors of T for the same eigenvalue is called the *generalized eigenspace* of T corresponding to that eigenvalue. In terms of the elementary divisor decomposition, the generalized eigenspaces are precisely the p-torsion parts of V where $p = x-\lambda$ is one of our primes. That is, grouping together all the elementary divisors given by the *same* prime $x-\lambda$, or equivalently the same eigenvalue λ, gives the subspace

$$F[x]/(x-\lambda)^{k_1} \oplus \cdots \oplus F[x]/(x-\lambda)^{k_s}$$

of all generalized eigenvectors for this λ, and the resulting expression for V:

$$V \cong \bigoplus_{\lambda} (\text{generalized eigenspace for } \lambda)$$

is the decomposition $M_{p_1} \oplus \cdots \oplus M_{p_t}$ we described before into p-torsion submodules.

Warm-Up 2. With the same notation as before, suppose concretely that V decomposes as

$$V \cong F[x]/(x-\lambda_1) \oplus F[x]/(x-\lambda_1)^2 \oplus F[x]/(x-\lambda_2)^3 \oplus F[x]/(x-\lambda_2)^3.$$

We determine the (monic) polynomial in $F[x]$ of *smallest* degree which annihilates all of V. Since nonzero ideals in $F[x]$ are generated by the smallest degree polynomials within them (note $F[x]$ is a Euclidean domain), this polynomial—called the *minimal polynomial of* T—then generates the annihilator $\mathrm{Ann}(V)$ of V. In this case, T only has two eigenvalues λ_1 and λ_2, the first of *multiplicity* 3 (i.e. the number of times λ_1 appears in the elementary divisors overall is 3: once in $x-\lambda_1$ and twice in $(x-\lambda_1)(x-\lambda_1)$) and the second of multiplicity 6. Picking basis for each of these factors gives a basis for V with 9 elements, so V is a 9-dimensional vector space over F. The Jordan norm form of T consists of four Jordan blocks: one of size 1 corresponding λ_1, another of size 2 also corresponding to λ_1, and two blocks of size 3 each corresponding to λ_2.

Now, the first factor in the decomposition above is annihilated by $x-\lambda_1$ and the second by $(x-\lambda_1)^2$. Thus in order to annihilate *both* of these factors our minimal polynomial would include a $(x-\lambda_1)^2$ term. The remaining two factors are each annihilated by $(x-\lambda_2)^3$ and nothing of smaller degree, so this must be a part of the minimal polynomial too. The minimal polynomial of T is thus

$$(x-\lambda_1)^2(x-\lambda_2)^3$$

and the annihilator of V (as an $F[x]$-module) is $\mathrm{Ann}(V) = ((x-\lambda_1)^2(x-\lambda_2)^3)$. This means that $(T-\lambda_1 I)^2(T-\lambda_2 I)^3 = 0$ as a linear transformation and any other polynomial which when "evaluated" at T gives 0 must be a multiple of the minimal polynomial.

In particular, the product of all the elementary divisors, so

$$(x-\lambda_1)(x-\lambda_1)^2(x-\lambda_2)^3(x-\lambda_2)^2 = (x-\lambda_1)^3(x-\lambda_2)^6$$

in this case, also annihilates everything (each factor in the construction of this polynomial annihilates the corresponding $F[x]/(x-\lambda)^k$ term in the elementary divisors decomposition), so it is itself a multiple of the minimal polynomial. This polynomial is called the *characteristic polynomial* of T, and is precisely the one you obtain via the standard $\det(A-\lambda I)$ computation when computing eigenvalues in a previous linear algebra course, only with variable x instead of λ. The fact that the characteristic polynomial annihilates everything (equivalently the fact that it is divisible by the minimal polynomial) is called the *Cayley-Hamilton Theorem*: i.e. $p(T) = 0$ where $p(x) \in F[x]$ is the characteristic polynomial of T.

If we group together the factors in the decomposition of V as:

$$V \cong \underbrace{\left(F[x]/(x-\lambda_1) \oplus F[x]/(x-\lambda_1)^2\right)}_{\text{generalized eigenspace for } \lambda_1} \oplus \underbrace{\left(F[x]/(x-\lambda_2)^3 \oplus F[x]/(x-\lambda_2)^3\right)}_{\text{generalized eigenspace for } \lambda_2}$$

we get the generalized eigenspace decomposition of V mentioned before, or equivalently the p-torsion decomposition with $(x-\lambda_1)$-torsion and $(x-\lambda_2)$-torsion parts in this case.

Examples. Let us take the explicit example of $V = \mathbb{R}^4$ with linear operator

$$A = \begin{bmatrix} 1 & -1 & -2 & 3 \\ 0 & 0 & -2 & 3 \\ 0 & 1 & 1 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

viewed as an $\mathbb{R}[x]$-module. The characteristic polynomial of A is $\det(A - xI) = (x-1)^4$, and one can check that in the fact the third power of this polynomial is enough to satisfy $(A-I)^3 = 0$ (the second power doesn't work), so the minimal polynomial of A is $(x-1)^3$. Thus the elementary divisor decomposition of $\mathbb{R}^4$ must be:

$$\mathbb{R}^4 \cong \mathbb{R}[x]/(x-1) \oplus \mathbb{R}[x]/(x-1)^3$$

since the largest elementary divisor which can occur is of degree 3 given that the minimal polynomial has degree 3; then, to get a 4-dimensional space overall, we need one more 1-dimensional factor, which corresponds to the elementary divisor $x-1$ of degree 1. The Jordan form of A thus consists of two blocks corresponding to the eigenvalue 1, of sizes 1 and 3, and so is:

$$\begin{bmatrix} 1 & & & \\ & 1 & 1 & \\ & & 1 & 1 \\ & & & 1 \end{bmatrix}.$$

Now suppose we had a 6×6 real matrix B acting on $\mathbb{R}^6$ with characteristic polynomial $(x-1)^5(x-2)$ and minimal polynomial $(x-1)^3(x-2)$. The elementary divisor corresponding to 2 must then be $x-2$, and the elementary divisor corresponding to 1 of largest degree must be $(x-1)^3$ going by what's given as the minimal polynomial. But there are other elementary divisors corresponding to 1 if we want to get to multiplicity 5 in the characteristic polynomial, and there are two possibilities: either there is one more quadratic elementary divisor $(x-1)^2$, or two more $x-1$ and $x-1$ each linear. In the first case, the $\mathbb{R}^6$ decomposes as

$$\mathbb{R}^6 \cong \mathbb{R}[x]/(x-1)^2 \oplus \mathbb{R}[x]/(x-1)^3 \oplus \mathbb{R}[x]/(x-2)$$

and in the second as

$$\mathbb{R}^6 \cong \mathbb{R}[x]/(x-1) \oplus \mathbb{R}[x]/(x-1) \oplus \mathbb{R}[x]/(x-1)^3 \oplus \mathbb{R}[x]/(x-2).$$

Each of these give different Jordan forms for B: blocks of sizes 2 and 3 for eigenvalue 1 and a block of size 1 for eigenvalue 2 in the first case, and blocks of sizes $1, 1, 3$ for 1 and a block of size 1 for 2 in the second. To determine which is the correct Jordan form requires doing more work, say by actually *computing* generalized eigenvectors for the eigenvalue 1. We will not do this here, as it is better left to a course in abstract linear algebra. (It has come to my attention that MATH 334 here does not always cover this, but it certainly does whenever I teach it. You can find my "Notes on Jordan Form" in the archive of MATH 334 course material on my website if you're interested in learning more.)

Applications of Jordan Form. We briefly mention two applications of Jordan normal forms. First, if A is a square matrix, one defines the *matrix exponential* e^A by

$$e^A = I + \sum_{n=1}^{\infty} \frac{1}{n!} A^n.$$

(This definition is motivated by the Maclaurian series of e^x.) The expression on the right above always converges, so the question becomes how to actually compute such a thing. It turns out that this can be done easily through the use of the Jordan normal form J of A. The basis which puts A into Jordan normal form make up the columns of an invertible matrix S satisfying $A = SJS^{-1}$,

and one can show that $e^A = Se^J S^{-1}$, so that the computation of e^A comes down to computing e^J. This in turn boils down to computing the exponential of each individual Jordan block making up J, and there is a straightforward formula for doing so. Check my "Notes on Jordan Form" mentioned above for details.

Second, and more directly relevant to stuff we did last quarter for instance, Jordan normal forms can be used to classify conjugacy classes of matrix groups, at least under the standing assumption that our base fields contains the eigenvalues of the matrices we care about. The fact is that two matrix A and B are conjugate if and only if they have the same Jordan normal form, up to the usual rearrangement of the Jordan blocks. Expressed in the language of linear algebra, the claim is that two matrices A and B are *similar* if and only if they have the same Jordan form, where to be similar means there exists exists an invertible S such that $A = SBS^{-1}$. (For instance, "diagonalizable" means similar to a diagonal matrix.) The point is that similar matrices represent the same linear transformation, only with respect to different bases. In the language of modules, similar matrices—or more generally similar linear operators—give rise to isomorphic $F[x]$-modules. Two $F[x]$-modules are isomorphic if and only if they have the same elementary divisor decomposition, which is turn is equivalent to giving the same Jordan normal form.

We actually used this fact towards the end of last quarter when classifying semidirect products, and indeed we briefly mentioned the vague concept of "Jordan form" when doing so then. The context was in trying to determine when two homomorphisms $\mathbb{Z}/q\mathbb{Z} \to \mathrm{Aut}(\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z})$ give rise to isomorphic semidirect products, with the claim we made at the time being that this was so if the images of the generators were conjugate in $\mathrm{Aut}(\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z})$. Thus to classifying such semidirect products requires understanding conjugacy classes of $\mathrm{Aut}(\mathbb{Z}/p\mathbb{Z} \times \mathbb{Z}/p\mathbb{Z}) \cong GL_2(\mathbb{Z}/p\mathbb{Z})$, and this is what Jordan normal forms allow us to do. Check the notes from last quarter to see where this was used: we made statements such as "any matrix of order 3 in $GL_2(\mathbb{Z}/2\mathbb{Z})$ is conjugate to such and such matrices", where the "such and such" matrices were Jordan forms.

Rational canonical form. We finish the quarter by briefly talking about the notion of a *rational canonical form*, which is derived from the invariant factors version of the structure theorem, in a manner analogous to how Jordan forms are derived from the elementary divisors version. (If we had one more day this winter quarter as we usually did in past winter quarters we could talk more about this, but our discussion here will be very quick.) Here we start with:

$$V \cong F[x]/(a_1(x)) \oplus \cdots \oplus F[x]/(a_m(x))$$

where $a_1(x) \mid \ldots \mid a_m(x)$ are the invariant factors of the $F[x]$-module V, with x acting by a linear operator T. Consider one such invariant factor

$$a(x) = b_0 + b_1 x + \cdots + b_{k-1}x^{k-1} + x^k.$$

The elements $1, x, \ldots, x^{k-1}$ form a basis for the factor $F[x]/(a(x))$ over F, where we use the relation $x^k = -b_0 - b_1 x - \cdots - b_{k-1}x^{k-1}$ in the quotient to deal with higher-order terms. The action of x on this basis now looks like:

$$1 \mapsto x,\ x \mapsto x^2,\ \ldots,\ x^{k-1} \mapsto x^k = -b_0 - b_1 x - \cdots - b_{k-1}x^{k-1},$$

so that x (or T on the subspace of V corresponding to this factor) sends each basis vector to the next, except for the final one, which it sends to the linear combination of the previous basis vectors

given above. The matrix of (the restriction of T) relative to this basis thus is

$$\begin{bmatrix} 0 & & & & -b_0 \\ 1 & 0 & & & -b_1 \\ & 1 & \ddots & & \vdots \\ & & \ddots & 0 & -b_{k-2} \\ & & & 1 & -b_{k-1} \end{bmatrix},$$

and doing this for each factor in the invariant factor decomposition of V produce a matrix for T on all of V which consists of blocks of this type, and this matrix is called the *rational canonical form* of T. This turns out to have the same use as one of the ones for the Jordan form mentioned before: two square matrices are similar if and only if they have the same rational canonical form.

So, the rational canonical form arises by a very similar process as the Jordan normal form. The rational canonical form does not look as "nice" as the "almost diagonal" Jordan form, but it has the benefit of always existing regardless of whether F contains all the eigenvalues of T. (The term "rational" in the name of this type of matrix comes from the fact that it exists over any field.) To give an example, suppose V has the following elementary divisors decomposition:

$$V \cong \mathbb{R}[x]/(x-1) \oplus \mathbb{R}[x]/(x-1) \oplus \mathbb{R}[x]/(x-1)^2 \oplus \mathbb{R}[x]/(x-2)^3 \oplus \mathbb{R}[x]/(x-2)^3.$$

Based on how we derived the invariant factors from the elementary divisors via the Chinese Remainder Theorem, we get that the invariant factors are:

$$(x-1) \mid (x-1)(x-2)^3 \mid (x-1)^2(x-2)^3$$

(Note that the largest invariant factor is the minimal polynomial of T, a fact which always holds.) The invariant factor decomposition of V is thus

$$\mathbb{R}[x]/(x-1) \oplus \mathbb{R}[x]/(x^4 - 7x^3 + 18x^2 - 20x + 8) \oplus \mathbb{R}[x]/(x^5 - 8x^4 + 25x^3 - 38x^2 + 28x - 8),$$

where the second and third factors come from expanding $(x-1)(x-2)^3$ and $(x-1)^2(x-2)^3$. The rational canonical form of T thus consists of a 1×1 block $\begin{bmatrix} 1 \end{bmatrix}$, and the following 4×4 and 5×5 blocks:

$$\begin{bmatrix} 0 & & & -8 \\ 1 & 0 & & 20 \\ & 1 & 0 & -18 \\ & & 1 & 7 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 0 & & & & 8 \\ 1 & 0 & & & -28 \\ & 1 & 0 & & 38 \\ & & 1 & 0 & -25 \\ & & & 1 & 8 \end{bmatrix}.$$

Truly magical stuff indeed. Thanks for reading!