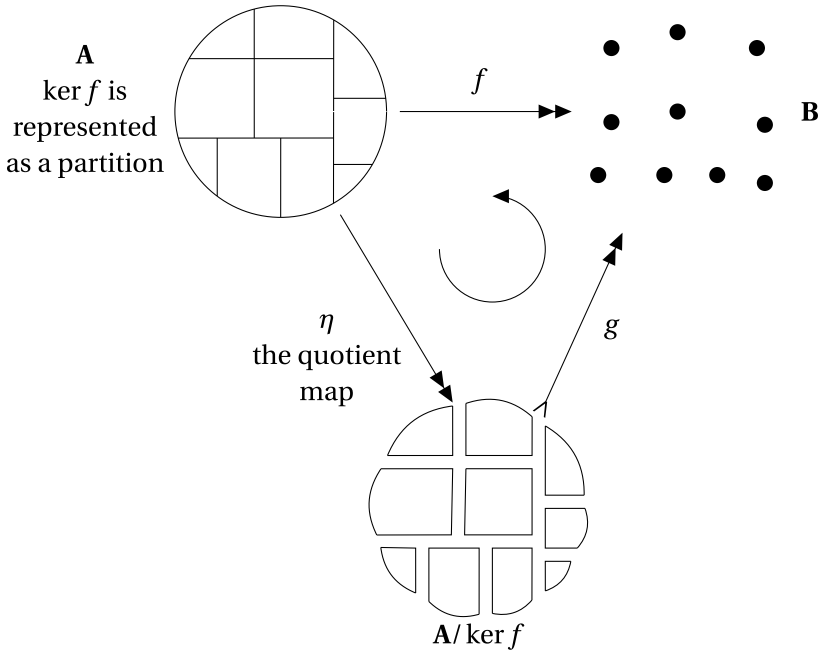
GEORGE F. MCNULTY

Algebra for First Year Graduate Students

DRAWINGS BY THE AUTHOR



UNIVERSITY OF SOUTH CAROLINA

2016

Preface

This exposition is for the use of first year graduate students pursuing advanced degrees in mathematics. In the United States, people in this position generally find themselves confronted with a battery of examinations at the beginning of their second year, so if you are among them a good part of your energy during your first year will be expended mastering the essentials of several branches of mathematics, algebra among them.

While every doctoral program in mathematics sets its own expectations, there is a fair consensus on those parts of algebra that should be part of a mathematician's repertoire. I have tried to gather here just those parts. So here you will find the basics of (commutative) rings and modules in Part I. The basics of groups and fields, constituting the content of second semester, are in Part II. The background you will need to make good use of this exposition is a good course in linear algebra and another in abstract algebra, both at the undergraduate level.

As you proceed through these pages you will find many places where the details and sometimes whole proofs of theorems will be left in your hands. The way to get the most from this presentation is to take it on with paper and pencil in hand and do this work as you go. There are also weekly problem sets. Most of the problems have appeared on Ph.D. examinations at various universities. In a real sense, the problems sets are the true heart of this presentation.

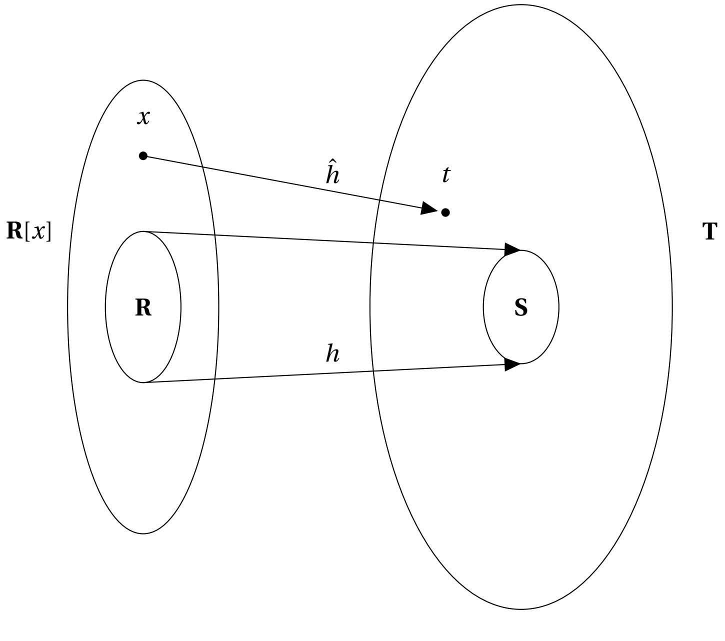
This work grew out of teaching first year graduate algebra courses. Mostly, I have done this at the University of South Carolina (but the first time I did it was at Dartmouth College and I had the delightful experience of teaching this material at the University of the Philippines). Many of the graduate students in these courses have influenced my presentation here. Before all others, I should mention Kate Scott Owens, who had the audacity to sit in the front row with her laptop, putting my classroom discussions into LaTeX on the fly. She then had the further audacity to post the results so that all the flaws and blunders I made would be available to everyone. So this effort at exposition is something in the way of self-defense.... I am deeply grateful to Jianfeng Wu and Nieves Austria McNulty for catching an impressively large number of flaws in earlier versions of this presentation. I own all those that remain.

During the final stages of preparing this account, the author was supported in part by National Science Foundation Grant 1500216.

George F. McNulty
Columbia, SC
2016

CONTENTS

Rings and Modules
Part I



LECTURE 0

The Basics of Algebraic Systems

0.1 Algebraic Systems

In your undergraduate coursework you have already encountered many algebraic systems. These probably include some specific cases, like $\langle \mathbb{Z}, +, \cdot, -, 0, 1\rangle$ which is the system of integers equipped with the usual two-place operations of addition, multiplication, the one-place operation of forming negatives, and two distinguished integers 0 and 1, which we will construe as zero-place operations (all output and no input). You have also encountered whole classes of algebraic systems such as the class of vector spaces over the real numbers, the class of rings, and the class of groups. You might even have encountered other classes of algebraic systems such as Boolean algebras and lattices.

The algebraic systems at the center of this two-semester course are rings, modules, groups, and fields. Vector spaces are special cases of modules. These kinds of algebraic systems arose in the nineteenth century and most of the mathematics we will cover was well-known by the 1930's. This material forms the basis for a very rich and varied branch of mathematics that has flourished vigorously over the ensuing decades.

Before turning to rings, modules, groups, and fields, it pays to look at algebraic systems from a fairly general perspective. Each algebraic system consists of a nonempty set of elements, like the set $\mathbb{Z}$ of integers, equipped with a system of operations. The nonempty set of elements is called the **universe** of the algebraic system. (This is a shortening of "universe of discourse".) Each of the operations is a function that takes as inputs arbitrary r-tuples of elements of the universe and returns an output again in the universe—here, for each operation, r is some fixed natural number called the **rank** of the operation. In the familiar algebraic system $\langle \mathbb{Z}, +, \cdot, -, 0, 1\rangle$, the operations of addition and multiplication are of rank 2 (they are two-place operations), the operation of forming negatives is of rank 1, and the two distinguished elements 0 and 1 are each regarded as operations of rank 0.

Aside. Let A be a set and r be a natural number. We use A^r to denote the set of all r-tuples of elements of A. An operation F of rank r on A is just a function from A^r into A. There is a curious case. Suppose A is the empty set and $r > 0$. Then A^r is also empty. A little reflection shows that the empty set is also a function from A^r into A, that is the empty set is an operation of rank r. The curiosity is that this is so for any positive natural number r. This means that the rank of this operation is not uniquely determined. We also note that A^0 actually has one element, namely the empty tuple. This means that when A is empty there can be no operations on A of rank 0. On the other hand, if A is nonempty, then the rank of every operation of finite rank is uniquely determined and A has operations of every finite rank. These peculiarities explain, to some extent, why we adopt the convention that the universe of an algebraic system should be nonempty.

The notion of the signature of an algebraic system is a useful way to organize the basic notions of our subject. Consider these three familiar algebraic systems:

$$\langle \mathbb{Z}, +, \cdot, -, 0, 1\rangle$$
$$\langle \mathbb{C}, +, \cdot, -, 0, 1\rangle$$
$$\langle \mathbb{R}_{2\times 2}, +, \cdot, -, 0, 1\rangle$$

The second system consists of the set of complex numbers equipped with the familiar operations, while the third system consists of the set of 2×2 matrices with real entries equipped with matrix multiplication, matrix addition, matrix negation, and distinguished elements 0 for the zero matrix, and 1 for the identity matrix. Notice that $+$ has a different meaning on each line displayed above. This is a customary, even well-worn, ambiguity. To resolve this ambiguity let us regard $+$ not as a two-place operation but as a symbol for a two-place operation. Then each of the three algebraic systems gives a different meaning to this symbol—a meaning that would ordinarily be understood from the context, but could be completely specified as needed. A **signature** is just a set of operation symbols, each with a uniquely determined natural number called its rank. More formally, a signature is a function with domain some set of operation symbols that assigns to each operation symbol its rank. The three algebraic systems above have the same signature.

"Algebraic system" is a mouthful. So we shorten it to "algebra". This convenient choice is in conflict with another use of this word to refer to a particular kind of algebraic system obtained by adjoining a two-place operation of a certain kind to a module.

As a matter of notation, we tend to use boldface $\mathbf{A}$ to denote an algebra and the corresponding normal-face A to denote its universe. For an operation symbol Q we use, when needed, $Q^{\mathbf{A}}$ to denote the operation of $\mathbf{A}$ symbolized by Q. We follow the custom of writing operations like $+$ between its inputs (like $5+2$), but this device does not work very well if the rank of the operation is not two. So in general we write things like $Q^{\mathbf{A}}(a_0, a_1, \dots, a_{r-1})$ where the operation symbol Q has rank r and $a_0, a_1, \dots, a_{r-1} \in A$.

Each algebra has a signature. It is reasonable to think of each algebra as one particular way to give meaning to the symbols of the signature.

Homomorphisms and their relatives

Let $\mathbf{A}$ and $\mathbf{B}$ be algebras of the same signature. We say that a function $h : A \to B$ is a **homomorphism** provided for every operation symbol Q and all $a_0, a_1, \dots, a_{r-1} \in A$, where r is the rank of Q, we have

$$h(Q^{\mathbf{A}}(a_0, a_1, \dots, a_{r-1})) = Q^{\mathbf{B}}(h(a_0), h(a_1), \dots, h(a_{r-1})).$$

That is, h preserves the basic operations. We use $h : \mathbf{A} \to \mathbf{B}$ to denote that h is a homomorphism from the algebra $\mathbf{A}$ into the algebra $\mathbf{B}$. For example, we learned in linear algebra that the determinant det is a homomorphism from $\langle \mathbb{R}_{2\times 2}, \cdot, 0, 1\rangle$ into $\langle \mathbb{R}, \cdot, 0, 1\rangle$. The key fact from linear algebra is

$$\det(AB) = \det A \det B.$$

We note in passing that the multiplication on the left (that is AB) is the multiplication of matrices, while the multiplication on the right is multiplication of real numbers.

In the event that h is a homomorphism from $\mathbf{A}$ into $\mathbf{B}$ that happens to be one-to-one we call it an **embedding** and express this in symbols as

$$h : \mathbf{A} \rightarrowtail \mathbf{B}.$$

In the event that the homomorphism h is onto B we say that $\mathbf{B}$ is a **homomorphic image** of $\mathbf{A}$ and write

$$h : \mathbf{A} \twoheadrightarrow \mathbf{B}.$$

In the event that the homomorphism h is both one-to-one and onto B we say that h is an **isomorphism** and we express this in symbols as

$$h : \mathbf{A} \rightarrowtail\!\!\!\!\twoheadrightarrow \mathbf{B}$$

or as

$$\mathbf{A} \overset{h}{\cong} \mathbf{B}.$$

It is an easy exercise, done by all hard-working graduate students, that if h is an isomorphism from $\mathbf{A}$ to $\mathbf{B}$ then the inverse function h^{-1} is an isomorphism from $\mathbf{B}$ to $\mathbf{A}$. We say that $\mathbf{A}$ and $\mathbf{B}$ are **isomorphic** and write

$$\mathbf{A} \cong \mathbf{B}$$

provided there is an isomorphism from $\mathbf{A}$ to $\mathbf{B}$.

The algebra $\langle \mathbb{R}, +, -, 0\rangle$ is isomorphic to $\langle \mathbb{R}^+, \cdot, ^{-1}, 1\rangle$, where $\mathbb{R}^+$ is the set of positive real numbers. There are isomorphisms either way that are familiar to freshmen in calculus. Find them.

A homomorphism from $\mathbf{A}$ into $\mathbf{A}$ is called an **endomorphism** of $\mathbf{A}$. An isomorphism from $\mathbf{A}$ to $\mathbf{A}$ is called an **automorphism** of $\mathbf{A}$.

Subuniverses and sublagebras

Let $\mathbf{A}$ be an algebra. A subset $B \subseteq A$ is called a **subuniverse** of $\mathbf{A}$ provided it is closed with respect to the basic operations of $\mathbf{A}$. This means that for every operation symbol Q of the signature of $\mathbf{A}$ and for all $b_0, b_1, \dots, b_{r-1} \in B$, where r is the rank of Q we have $Q^{\mathbf{A}}(b_0, b_1, \dots, b_{r-1}) \in B$. Notice that if the signature of $\mathbf{A}$ has an operation symbol c of rank 0, then $c^{\mathbf{A}}$ is an element of A and this element must belong to every subuniverse of $\mathbf{A}$. On the other hand, if the signature of $\mathbf{A}$ has no operation symbols of rank 0, then the empty set $\varnothing$ is a subuniverse of $\mathbf{A}$.

The restriction of any operation of $\mathbf{A}$ to a subuniverse B of $\mathbf{A}$ results in an operation on B. In the event that B is a nonempty subuniverse of $\mathbf{A}$, we arrive at the **subalgebra B** of $\mathbf{A}$. This is the algebra of the same signature as $\mathbf{A}$ with universe B such that $Q^{\mathbf{B}}$ is the restriction to B of $Q^{\mathbf{A}}$, for each operation symbol Q of the signature. $\mathbf{B} \leq \mathbf{A}$ symbolizes that $\mathbf{B}$ is a subalgebra of $\mathbf{A}$.

Here is a straightforward but informative exercise for hard-working graduate students. Let $\mathbb{N} = \{0, 1, 2, \dots\}$ be the set of natural numbers. Discover all the subuniverses of the algebra $\langle \mathbb{N}, +\rangle$.

Congruence relations

Let $\mathbf{A}$ be an algebra and h be a homomorphism from $\mathbf{A}$ to some algebra. We associate with h the following set, which called here the **functional kernel** of h,

$$\theta = \{(a, a') \mid a, a' \in A \text{ and } h(a) = h(a')\}.$$

This set of ordered pairs of elements of A is evidently an equivalence relation on A. That is, θ has the following properties.

(a) It is reflexive: $(a, a) \in \theta$ for all $a \in A$.

(b) It is symmetric: for all $a, a' \in A$, if $(a, a') \in \theta$,then $(a', a) \in \theta$.

(c) It is transitive: for all $a, a', a'' \in A$, if $(a, a') \in \theta$ and $(a', a'') \in \theta$, then $(a, a'') \in \theta$.

This much would be true were h any function with domain A. Because θ is a binary (or two-place) relation on A it is useful to use the following notations interchangeably.

$$\begin{gathered}(a,a')\in\theta\\ a\,\theta\,a'\\ a\equiv a' \mod \theta\end{gathered}$$

Here is another piece of notation which we will use often. For any set A, any $a \in A$ and any equivalence relation θ on A we put

$$a/\theta := \{a' \mid a' \in A \text{ and } a \equiv a' \mod \theta\}.$$

We also put

$$A/\theta := \{a/\theta \mid a \in A\}.$$

Because h is a homomorphism θ has one more important property, sometimes called the *substitution property*:

For every operation symbol Q of the signature of $\mathbf{A}$ and for all $a_0, a'_0, a_1, a'_1, \dots, a_{r-1}, a'_{r-1} \in A$, where r is the rank of Q,

if

$$\begin{gathered}a_0 \equiv a'_0 \mod \theta\\ a_1 \equiv a'_1 \mod \theta\\ \vdots\\ a_{r-1} \equiv a'_{r-1} \mod \theta\end{gathered}$$

then

$$Q^{\mathbf{A}}(a_0, a_1, \dots, a_{r-1}) \equiv Q^{\mathbf{A}}(a'_0, a'_1, \dots, a'_{r-1}) \mod \theta.$$

An equivalence relation on A with the substitution property above is called a **congruence** relation of the algebra $\mathbf{A}$. The functional kernel of a homomorphism h from $\mathbf{A}$ into some other algebra is always a congruence of $\mathbf{A}$. We will see below that this congruence retains almost all the information about the homomorphism h.

As an exercise to secure the comprehension of this notion, the hard-working graduate students should try to discover all the congruence relations of the familiar algebra $\langle \mathbb{Z}, +, \cdot \rangle$.

A comment of mathematical notation

The union $A \cup B$ and the intersection $A \cap B$ of sets A and B are familiar. These are special cases of more general notions. Let $\mathcal{K}$ be any collection of sets. The union $\bigcup \mathcal{K}$ is defined via

$$a \in \bigcup \mathcal{K} \Leftrightarrow a \in C \text{ for some } C \in \mathcal{K}.$$

Here is the special case $A \cup B$ rendered in this way

$$a \in A \cup B \Leftrightarrow a \in \bigcup \{A, B\} \Leftrightarrow a \in C \text{ for some } C \in \{A, B\} \Leftrightarrow a \in A \text{ or } a \in B.$$

Similarly, the intersection $\bigcap \mathcal{K}$ is defined via

$$a \in \bigcap \mathcal{K} \Leftrightarrow a \in C \text{ for all } C \in \mathcal{K}.$$

Notice that in the definition of the intersection, each set belonging to the collection $\mathcal{K}$ imposes a constraint on what elements are admitted to membership in $\bigcap \mathcal{K}$. When the collection $\mathcal{K}$ is empty there are no constraints at all on membership in $\bigcap \mathcal{K}$. This means $\bigcap \varnothing$ is the collection of all sets. However, having the collection of all sets in hand leads to a contradiction, as discovered independently by Ernst Zermelo and Bertrand Russell in 1899. To avoid this, we must avoid forming the intersection of empty collections. This situation is analogous to division by zero. Just as when division of numbers comes up, the careful mathematician considers the possibility that the divisor is zero before proceeding, so must the careful mathematician consider the possibility that $\mathcal{K}$ might be empty before proceeding to form $\bigcap \mathcal{K}$.

We also use the notation

$$\bigcup_{i \in I} A_i \text{ and } \bigcap_{i \in I} A_i$$

to denote the union and intersection of $\mathcal{K} = \{A_i \mid i \in I\}$. The set I here is used as a set of indices. In using this notation, we impose no restrictions on I (save that in forming intersections the set I must not be empty). In particular, we make no assumption that the set I is ordered in any way.

The familiar set builder notation, for example $\{n \mid n \text{ is a prime number}\}$, has a companion in the function builder notation. Here is an example

$$f = \langle e^x \mid x \in \mathbb{R} \rangle.$$

The function f is just the exponential function on the real numbers. We take the words “function”, “sequence”, and “system” to have the same meaning. We also use the notation $f(c)$ and f_c interchangeably when f is a function and c is a member of its domain.

0.2 Problem Set 0

Algebra Homework, Edition 0

First Week

How Is Your Linear Algebra?

Problem 0.

Classify up to similarity all the square matrices over the complex numbers with minimal polynomial $m(x) = (x-1)^2(x-2)^2$ and characteristic polynomial $c(x) = (x-1)^6(x-2)^5$.

Problem 1.

Let $T : V \to V$ be a linear transformation of rank 1 on a finite dimensional vector space V over any field. Prove that either T is nilpotent or V has a basis of eigenvectors of T.

Problem 2.

Let V be a vector space over a field K.

(a) Prove that if U_0 and U_1 are subspaces of V such that $U_0 \nsubseteq U_1$ and $U_1 \nsubseteq U_0$, then $V \neq U_0 \cup U_1$.

(b) Prove that if U_0, U_1, and U_2 are subspaces of V such that $U_i \nsubseteq U_j$ when $i \neq j$ and K has at least 3 elements, then $V \neq U_0 \cup U_1 \cup U_2$.

(c) State and prove a generalization of (b) for n subspaces.

Problem 3.

Let $\mathbf{F}$ be a field and n be a positive integer. Let A be an $n \times n$ matrix with entries from $\mathbf{F}$ such that A^n is zero but A^{n-1} is nonzero. Show that any $n \times n$ matrix B over $\mathbf{F}$ that commutes with A is contained in the span of $\{I, A, A^2, \ldots, A^{n-1}\}$.

Problem 4.

Let $\mathbf{V}$ be a nontrivial finite dimensional vector space over the complex numbers.

(a) Suppose S and T are commuting linear operators on $\mathbf{V}$. Prove that each eigenspace of S is mapped into itself by T.

(b) Now let $A_0, \ldots, A_{k-1}$ be finitely many linear operators on $\mathbf{A}$ that commute pairwise. Prove that they have a common eigenvector.

(c) Suppose that the dimension of $\mathbf{V}$ is n. Prove that there exists a nested sequence of subspaces

$$\mathbf{V}_0 \subseteq \mathbf{V}_1 \subseteq \cdots \subseteq \mathbf{V}_n = \mathbf{V}$$

where each $\mathbf{V}_j$ has dimension j and is mapped into itself by each of the operators

$$A_0, A_1, \ldots, A_{k-1}.$$

0.3 The Homomorphism Theorem for Algebras of the Simplest Signature

The simplest signature is, of course, empty—it provides no operation symbols. In this setting, algebras have nothing real to distinguish them from nonempty sets. Every function between two nonempty sets will be a homomorphism. Every subset will be a subuniverse. Every equivalence relation will be a congruence. Isomorphisms are just one-to-one correspondences between nonempty sets and two nonempty sets will be isomorphic just in case they have the same cardinality. So doing algebra in the empty signature is a branch of combinatorics. Nevertheless, there is an important lesson for us to learn here.

Let A be a nonempty set. A **partition** of A is a collection $\mathcal{P}$ of subsets of A with the following properties:

(a) Each member of $\mathcal{P}$ is a nonempty subset of A.

(b) If $X, Y \in \mathcal{P}$ and $X \neq Y$, then X and Y are disjoint.

(c) Every element of A belongs to some set in the collection $\mathcal{P}$.

You may already be familiar with the close connections among the notions of a function with domain A, of an equivalence relation on A, and of a partition of A. We present it in the following theorem:

The Homomorphism Theorem, Empty Signature Version. *Let A be a nonempty set, $f : A \twoheadrightarrow B$ be a function from A onto B, θ be an equivalence relation on A, and $\mathcal{P}$ be a partition of A. All of the following hold.*

(a) *The functional kernel of f is an equivalence relation on A.*

(b) *The collection $A/\theta = \{a/\theta \mid a \in A\}$ is a partition of A.*

(c) *The map η that assigns to each $a \in A$ the set in $\mathcal{P}$ to which it belongs is a function from A onto $\mathcal{P}$; moreover $\mathcal{P}$ is the collection of equivalence classes of the functional kernel of η.*

(d) *If θ is the functional kernel of f, then there is a one-to-one correspondence g from A/θ to B such that $f = g \circ \eta$, where $\eta\colon A \twoheadrightarrow A/\theta$ with $\eta(a) = a/\theta$ for all $a \in A$.*

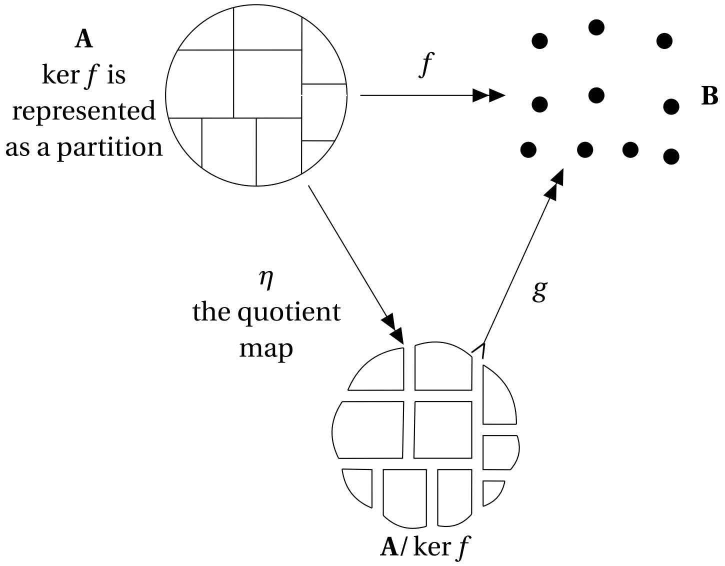


Figure 0.1: The Homomorphism Theorem

The empty signature version of the Homomorphism Theorem is almost too easy to prove. Figure 0.1 almost tells the whole story. One merely has to check what the definitions of the various notions require. The map η is called the **quotient map**. That it is a function, i.e. that

$$\{(a, X) \mid a \in A \text{ and } a \in X \in \mathcal{P}\}$$

is a function, follows from the disjointness of distinct members of the partition. That its domain is A follows from condition (c) in the definition of partition. The one-to-one correspondence g mentioned in assertion (d) of the Homomorphism Theorem is the following set of ordered pairs:

$$\{(a/\theta, f(a)) \mid a \in A\}.$$

The proof that this set is a one-to-one function from A onto B is straightforward, the most amusing part being the demonstration that it is actually a function.

0.4 Direct Products

Just as you are familiar with $A \cup B$ and $A \cap B$, you probably already know that $A \times B$ denotes the set of all ordered pairs whose first entries are chosen from A while the second entries are chosen from B. Just as we did for unions and intersections, we will extend this notion.

Let $\langle A_i \mid i \in I\rangle$ be any system of sets. We call a function $a : I \to \bigcup_{i\in I} A_i$ a **choice function** for the system $\langle A_i \mid i \in I\rangle$ provided $a_i \in A_i$ for all $i \in I$. It is perhaps most suggestive to think of a as an I-tuple (recalling that we are using function, tuple, system, and sequence interchangeably). The **direct product** of the system $\langle A_i \mid i \in I\rangle$ is just the set of all these choice functions. Here is the notation we use:

$$\prod \langle A_i \mid i \in I\rangle := \prod_{i\in I} A_i := \{a \mid a \text{ is a choice function for the system } \langle A_i \mid i \in I\rangle\}.$$

The sets A_i are called the **direct factors** of this product. If any of the sets in the system $\langle A_i \mid i \in I\rangle$ is empty, then the direct product is also empty. On the other hand, if I is empty then the direct product is $\{\varnothing\}$, since the empty set will turn out to be a choice function for the system. Notice that $\{\varnothing\}$ is itself nonempty and, indeed, has exactly one element.

Observe that $\prod\langle A, B\rangle = \{\langle a, b\rangle \mid a \in A \text{ and } b \in B\}$. This last set is, for all practical purposes, $A \times B$.

Projection functions are associated with direct products. For any $j \in I$, the j^{th} projection function p_j is defined, for all $a \in \prod\langle A_i \mid i \in I\rangle$, via

$$p_j(a) := a_j.$$

The system of projection functions has the following useful property: it **separates points**. This means that if $a, a' \in \prod\langle A_i \mid i \in T\rangle$ and $a \neq a'$, then $p_j(a) \neq p_j(a')$ for some $j \in I$. Suppose that $\langle A_i \mid i \in I\rangle$ is a system of sets, that B is some set, and that $\langle f_i \mid i \in I\rangle$ is a system of functions such that $f_i : B \to A_i$ for each $i \in I$. Define the map $h : B \to \prod\langle A_i \mid i \in I\rangle$ via

$$h(b) := \langle f_i(b) \mid i \in I\rangle.$$

Then it is easy to see that $f_i = p_i \circ h$ for all $i \in I$. If the system $\langle f_i \mid i \in I\rangle$ separates points, then the function h defined just above will be one-to-one, as all hard-working graduate students will surely check.

We form direct products of systems of algebras in the following way. Let $\langle \mathbf{A}_i \mid i \in I\rangle$ be a system of algebras, all with the same signature. We take $\prod\langle \mathbf{A}_i \mid i \in I\rangle$ to be the algebra $\mathbf{P}$ with universe $P := \prod\langle A_i \mid i \in I\rangle$ and where the operations on $\mathbf{P}$ are defined coordinatewise. This means that for each operation symbol Q and all $a_0, a_1, \ldots, a_{r-1} \in P$, where r is the rank of Q, we have

$$Q^{\mathbf{P}}(a_0, a_1, \ldots, a_{r-1}) = \left\langle Q^{\mathbf{A}_i}(a_{0,i}, a_{1,i}, \ldots, a_{r-1,i}) \mid i \in I\right\rangle.$$

To see more clearly what is intended here, suppose that Q has rank 3, that $I = \{0,1,2,3\}$, and that $a, b, c \in P$. Then

$$\begin{array}{ccllll}
a & = & \langle a_0, & a_1, & a_2, & a_3\rangle \\
b & = & \langle b_0, & b_1, & b_2, & b_3\rangle \\
c & = & \langle c_0, & c_1, & c_2, & c_3\rangle \\
\multicolumn{3}{r}{Q^{\mathbf{P}}(a,b,c)=\Big\langle Q^{\mathbf{A}_0}(a_0,b_0,c_0),} & Q^{\mathbf{A}_1}(a_1,b_1,c_1), & Q^{\mathbf{A}_2}(a_2,b_2,c_2), & Q^{\mathbf{A}_3}(a_3,b_3,c_3)\Big\rangle
\end{array}$$

In this way, the direct product of a system of algebras, all of the same signature, will be again an algebra of the common signature and it is evident that each projection map is a homomorphism from the direct product onto the corresponding direct factor. Even the following fact is easy to prove.

Fact. Let $\langle \mathbf{A}_i \mid i \in I\rangle$ be a system of algebras, all of the same signature. Let $\mathbf{B}$ be an algebra of the same signature as $\mathbf{A}$ and let $\langle f_i \mid i \in I\rangle$ be a system of homomorphisms so that $f_i : \mathbf{B} \to \mathbf{A}_i$ for all $i \in I$. Then there is a homomorphism $h : \mathbf{B} \to \prod_{i\in I} \mathbf{A}_i$ so that $f_i = p_i \circ h$ for all $i \in I$. Moreover, if $\langle f_i \mid i \in I\rangle$ separates points, then h is one-to-one.

Lecture 1

The Isomorphism Theorems

The four theorems presented today arose over a period of perhaps forty years from the mid 1890's to the mid 1930's. They emerged from group theory and the theory of rings and modules chiefly in the work of Richard Dedekind and Emmy Noether and it was Noether who gave their first clear formulation in the context of module theory in 1927. You have probably already seen versions of these theorems for groups or rings in an undergraduate abstract algebra course.

We will frame them in the broader context of algebras in general. That way it will not be necessary to do more than add a comment or two when applying them in the context of groups, rings, and modules (these being our principal focus). In addition, you will be able to apply them in the context of lattices, Boolean algebras, or other algebraic systems.

At the center of this business is the notion of a **quotient algebra**. Let $\mathbf{A}$ be an algebra and let θ be a congruence of $\mathbf{A}$. Recall that for each $a \in A$ we use a/θ to denote the congruence class $\{a' \mid a' \in A \text{ and } a \equiv a' \mod \theta\}$. Moreover, we use A/θ to denote the partition $\{a/\theta \mid a \in A\}$ of A into congruence classes. We make the quotient algebra $\mathbf{A}/\theta$ by letting its universe be A/θ and, for each operation symbol Q of the signature of $\mathbf{A}$, and all $a_0, a_1, \ldots, a_{r-1} \in A$, where r is the rank of Q, we define

$$Q^{\mathbf{A}/\theta}(a_0/\theta, a_1/\theta, \ldots, a_{r-1}/\theta) := Q^{\mathbf{A}}(a_0, a_1, \ldots, a_{r-1})/\theta.$$

Because the elements of A/θ are congruence classes, we see that the r inputs to $Q^{\mathbf{A}/\theta}$ must be congruence classes. On the left side of the equation above the particular elements a_i have no special standing—they could be replaced by any a'_i provided only that $a_i \equiv a'_i \mod \theta$. Loosely speaking, what this definition says is that to evaluate $Q^{\mathbf{A}/\theta}$ on an r-tuple of θ-classes, reach into each class, grab an element to represent the class, evaluate $Q^{\mathbf{A}}$ at the r-tuple of selected representatives to obtain say $b \in A$, and then output the class b/θ. A potential trouble is that each time such a process is executed on the same r-tuple of congruence classes, different representatives might be selected resulting in, say b', instead of b. But the substitution property, the property that distinguishes congruences from other equivalence relations, is just what is needed to see that there is really no trouble. To avoid a forest of subscripts, here is how the argument would go were Q to have rank 3. Suppose $a, a', b, b', c, c' \in A$ with

$$a/\theta = a'/\theta$$
$$b/\theta = b'/\theta$$
$$c/\theta = c'/\theta.$$

So a and a' can both represent the same congruence class—the same for b and b' and for c and c'. Another way to write this is

$$\begin{aligned} a &\equiv a' \mod \theta \\ b &\equiv b' \mod \theta \\ c &\equiv c' \mod \theta. \end{aligned}$$

What we need is $Q^{\mathbf{A}}(a,b,c)/\theta = Q^{\mathbf{A}}(a',b',c')/\theta$. Another way to write that is

$$Q^{\mathbf{A}}(a,b,c) \equiv Q^{\mathbf{A}}(a',b',c') \mod \theta.$$

But this is exactly what the substitution property provides. Hard-working graduate students will do the work to see that what works for rank 3 works for any rank.

The theorem below, sometimes called the First Isomorphism Theorem, is obtained from its version for the empty signature replacing arbitrary functions by homomorphisms and arbitrary equivalence relations by congruence relations.

The Homomorphism Theorem. *Let* $\mathbf{A}$ *be an algebra, let* $f : \mathbf{A} \twoheadrightarrow \mathbf{B}$ *be a homomorphism from* $\mathbf{A}$ *onto* $\mathbf{B}$*, and let* θ *be a congruence relation of* $\mathbf{A}$*. All of the following hold.*

(a) *The functional kernel of* f *is a congruence relation of* A*.*

(b) $\mathbf{A}/\theta$ *is an algebra of the same signature as* $\mathbf{A}$*.*

(c) *The map* η *that assigns to each* $a \in A$ *the congruence class* a/θ *is a homomorphism from* $\mathbf{A}$ *onto* $\mathbf{A}/\theta$ *and its functional kernel is* θ*.*

(d) *If* θ *is the functional kernel of* f*, then there is an isomorphism* g *from* $\mathbf{A}/\theta$ *to* $\mathbf{B}$ *such that* $f = g \circ \eta$*, where* $\eta : \mathbf{A} \twoheadrightarrow \mathbf{A}/\theta$ *with* $\eta(a) = a/\theta$ *for all* $a \in A$*.*

The proof of this theorem has been, for the most part, completed already. We just saw how to prove part (b) and part (a) was done when the notions of congruence relation and functional kernel were introduced. Even parts (c) and (d) were mostly established in the version of the theorem for algebras with empty signature. It only remains to prove that the quotient map η in part (c) and the map g in part (d) are actually homomorphisms. With the definition of how the operations in the quotient algebra work, this only requires checking that the basic operations are preserved by η and by g. This work is left to the diligent graduate students.

From parts (a) and (c) of the Homomorphism Theorem we draw the following corollary.

Corollary 1.0.1. *Let* $\mathbf{A}$ *be an algebra. The congruence relations of* $\mathbf{A}$ *are exactly the functional kernels of homomorphisms from* $\mathbf{A}$ *into algebras of the same signature as* $\mathbf{A}$*.*

It will be necessary, as we develop the theory of rings, modules, and groups, to determine whether certain equivalence relations at hand are in fact congruence relations. Of course, we can always check the conditions defining the concept of congruence relation. But sometimes it is simpler to show that the relation is actually the functional kernel of some homomorphism.

Now let us suppose that θ is a congruence of $\mathbf{A}$ and that $\mathbf{B}$ is a subalgebra of $\mathbf{A}$. By $\theta \restriction B$ we mean the restriction of θ to B. That is

$$\theta \restriction B := \theta \cap (B \times B).$$

Now θ partitions A into congruence classes. Some of these congruence classes may include elements of B while others may not. We can **inflate** B using θ to obtain the set θB of all elements of A related by θ to some element of B. That is

$$\theta B := \{a \mid a \equiv b \mod \theta \text{ for some } b \in B\}.$$

Figure 1.1 illustrates the inflation of B by θ, where we have drawn lines to indicate the partition of A into θ-classes.

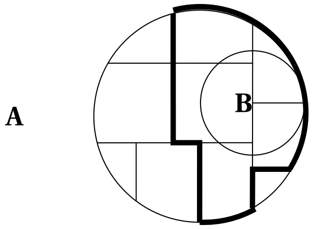


Figure 1.1: The Inflation θB of $\mathbf{B}$ by θ

The Second Isomorphism Theorem. *Let*$\mathbf{A}$ *be an algebra, let* θ *be a congruence of* $\mathbf{A}$*, and let* $\mathbf{B}$ *be a subalgebra of* $\mathbf{A}$*. Then each of the following hold.*

(a) $\theta \restriction B$ *is a congruence relation of* $\mathbf{B}$.

(b) θB *is a subuniverse of* $\mathbf{A}$.

(c) $\boldsymbol{\theta}\mathbf{B}/(\theta \restriction \theta B) \cong \mathbf{B}/(\theta \restriction B)$.

Proof. For part (a) we have to see that $\theta \restriction B$ is an equivalence relation on B and that it has the substitution property. Hard-working graduate students will check that it is indeed an equivalence relation. To see that the substitution property holds, let Q be an operation symbol. Just for simplicity, let us suppose the rank of Q is 3. Pick $a, a', b, b', c, c' \in B$ so that

$$\begin{aligned} a &\equiv a' \mod \theta \restriction B \\ b &\equiv b' \mod \theta \restriction B \\ c &\equiv c' \mod \theta \restriction B. \end{aligned}$$

We must show that $Q^{\mathbf{B}}(a,b,c) \equiv Q^{\mathbf{B}}(a',b',c') \mod \theta \restriction B$. Because all those elements come from B we see that

$$\begin{aligned} a &\equiv a' \mod \theta \\ b &\equiv b' \mod \theta \\ c &\equiv c' \mod \theta, \end{aligned}$$

and that both $Q^{\mathbf{B}}(a,b,c) = Q^{\mathbf{A}}(a,b,c)$ and $Q^{\mathbf{B}}(a',b',c') = Q^{\mathbf{A}}(a',b',c')$. It follows from the substitution property for θ that $Q^{\mathbf{A}}(a,b,c) \equiv Q^{\mathbf{A}}(a',b',c') \mod \theta$. But since both $Q^{\mathbf{A}}(a,b,c) = Q^{\mathbf{B}}(a,b,c) \in B$ and $Q^{\mathbf{A}}(a',b',c') = Q^{\mathbf{B}}(a',b',c') \in B$, we draw the desired conclusion that $Q^{\mathbf{B}}(a,b,c) \equiv Q^{\mathbf{B}}(a',b',c') \mod \theta \restriction B$.

For part (b) we have to show that θB is closed under all the basic operations of $\mathbf{A}$. So let Q be an operation symbol, which without loss of generality we assume to have rank 3. Let $a, b, c \in \theta B$. Our goal is to show that $Q^{\mathbf{A}}(a,b,c) \in \theta B$. Using the definition of θB pick $a', b', c' \in B$ so that

$$\begin{aligned} a &\equiv a' \mod \theta \\ b &\equiv b' \mod \theta \\ c &\equiv c' \mod \theta. \end{aligned}$$

Because B is a subuniverse, we see that $Q^{\mathbf{A}}(a',b',c') \in B$. Because θ is a congruence, we see that $Q^{\mathbf{A}}(a,b,b) \equiv Q^{\mathbf{A}}(a',b',c')$. Putting these together, we find that $Q^{\mathbf{A}}(a,b,c) \in \theta B$, as desired.

For part (c) we will invoke the Homomorphism Theorem. Define the map h from B to $\theta B/(\theta \restriction \theta B)$ via

$$h(b) := b/(\theta \restriction \theta B).$$

We have three contentions, namely that h is a homomorphism, that h is onto $B/(\theta \restriction \theta B)$, and that the functional kernel of h is $\theta \restriction B$. Given these, the Homomorphism Theorem provides that desired isomorphism.

To see that h is a homomorphism we have to show it respects the operations. So again take Q to be an operation symbol, of rank 3 for simplicity. Let $a,b,c \in B$. Now observe

$$\begin{aligned} h(Q^{\mathbf{B}}(a,b,c)) &= Q^{\mathbf{B}}(a,b,c)/(\theta \restriction \theta B) \\ &= Q^{\boldsymbol{\theta}\mathbf{B}}(a,b,c)/(\theta \restriction \theta B) \\ &= Q^{\boldsymbol{\theta}\mathbf{B}/(\theta \restriction \theta B)}(a/(\theta \restriction \theta B), b/(\theta \restriction \theta B), c/(\theta \restriction \theta B)) \\ &= Q^{\boldsymbol{\theta}\mathbf{B}/(\theta \restriction \theta B)}(h(a),h(b),h(c)). \end{aligned}$$

In this way we see that h respects Q. So h is a homomorphism.

To see that h is onto, let $b' \in \theta B$. Pick $b \in B$ so that $b' \equiv b \mod \theta$. We assert that $h(b) = b'/(\theta \restriction \theta B)$. So what we have to demonstrate is that

$$b/(\theta \restriction \theta B) = b'/(\theta \restriction \theta B)$$

or what is the same

$$b \equiv b' \mod \theta \restriction \theta B.$$

Now both b and b' belong to θB, so all that remains is to see that $b \equiv b' \mod \theta$. But we already know this.

Finally, we have to understand the functional kernel of h. Let $a,b \in B$ and observe

$$\begin{aligned} h(a) = h(b) &\Leftrightarrow a/(\theta \restriction \theta B) = b/(\theta \restriction \theta B) \\ &\Leftrightarrow a \equiv b \mod \theta \restriction \theta B \\ &\Leftrightarrow a \equiv b \mod \theta \restriction B. \end{aligned}$$

The last equivalence follows since a and b both belong to B. So we see that $\theta \restriction B$ is the functional kernel of h, completing the proof. □

Let $\mathbf{A}$ be an algebra and let θ and φ be congruences of $\mathbf{A}$ with $\theta \subseteq \varphi$. Let

$$\varphi/\theta := \{(a/\theta, a'/\theta) \mid a,a' \in A \text{ with } a \equiv a' \mod \varphi\}.$$

So φ/θ is a two-place relation on A/θ.

The Third Isomorphism Theorem. *Let* $\mathbf{A}$ *be an algebra and let* θ *and* φ *be congruences of* $\mathbf{A}$ *with* $\theta \subseteq \varphi$. *Then*

(a) φ/θ *is a congruence of* $\mathbf{A}/\theta$, *and*

(b) $\mathbf{A}/\theta \big/ \varphi/\theta \cong \mathbf{A}/\varphi$.

Proof. Define the function h from A/θ to A/φ so that for all $a \in A$ we have

$$h(a/\theta) := a/\varphi.$$

Here we have to worry again whether h is really a function—the definition above uses a representative element a of the congruence class a/θ to say how to get from the input to the output. What if $a/\theta = a'/\theta$?

Then $(a,a') \in \theta$. Since $\theta \subseteq \varphi$, we get $(a,a') \in \varphi$. This means, of course, that $a/\varphi = a'/\varphi$. So we arrive at the same output, even using different representatives. This means our definition is sound.

Let us check that h is a homomorphism. So let Q be an operation symbol, which we suppose has rank 3 just in order to avoid a lot of indices. Pick $a,b,c \in A$. Now observe

$$\begin{aligned} h(Q^{\mathbf{A}/\theta}(a/\theta), b/\theta, c/\theta) &= h(Q^{\mathbf{A}}(a,b,c)/\theta) \\ &= Q^{\mathbf{A}}(a,b,c)/\varphi \\ &= Q^{\mathbf{A}/\varphi}(a/\varphi, b/\varphi, c/\varphi) \\ &= Q^{\mathbf{A}/\varphi}(h(a/\theta), h(b/\theta), h(c/\theta)) \end{aligned}$$

In this way we see that h respects the operation symbol Q. We conclude that h is a homomorphism.

Notice that h is onto A/φ since any member of that set has the form a/φ for some $a \in A$. This means that $h(a/\theta) = a/\varphi$.

Now lets tackle the functional kernel of h. Let $a,b \in A$. Then observe

$$h(a/\theta) = h(b/\theta) \Leftrightarrow a/\varphi = b/\varphi \Leftrightarrow a \equiv b \mod \varphi.$$

So $(a/\theta, b/\theta)$ belongs to the functional kernel of h if and only if $a \equiv b \mod \varphi$. That is, the functional kernel of h is φ/θ. From the Homomorphism Theorem we see that φ/θ is a congruence of $\mathbf{A}/\theta$. Also from the Homomorphism Theorem we conclude that $(\mathbf{A}/\theta)/(\varphi/\theta) \cong \mathbf{A}/\varphi$. □

The set inclusion relation $\subseteq$ is a partial ordering of the congruence relations of an algebra $\mathbf{A}$. Some of the secrets of $\mathbf{A}$ can be discovered by understanding how the congruence relations are ordered. The next theorem, sometimes called the Fourth Isomorphism Theorem, is a first and useful step along this road. To understand it we need the notion of isomorphism of relational structures (as opposed to algebras). Let A and B be nonempty sets and let $\sqsubseteq$ be a two-place relation on A and $\preceq$ be a two-place relation on B. A function h from A to B is called an **isomorphism** between $\langle A, \sqsubseteq\rangle$ and $\langle B, \prec\rangle$ provided h is one-to-one, h is onto B, and for all $a,a' \in A$ we have

$$a \sqsubseteq a' \text{ if and only if } h(a) \preceq h(a').$$

As a matter of notation, let $\operatorname{Con}\mathbf{A}$ be the set of congruence relations of $\mathbf{A}$.

The Correspondence Theorem. *Let* $\mathbf{A}$ *be an algebra and let* θ *be a congruence of* $\mathbf{A}$. *Let* $P = \{\varphi \mid \varphi \in \operatorname{Con}\mathbf{A}$ *and* $\theta \subseteq \varphi\}$. *Then the map from* P *to* $\operatorname{Con}\mathbf{A}/\theta$ *that sends each* $\varphi \in P$ *to* φ/θ *is an isomorphism between the ordered set* $\langle P, \subseteq\rangle$ *and the ordered set* $\langle \operatorname{Con}\mathbf{A}/\theta, \subseteq\rangle$.

Proof. Let Ψ denote the map mentioned in the theorem. So

$$\Psi(\varphi) = \varphi/\theta$$

for all $\varphi \in \operatorname{Con}\mathbf{A}$ with $\theta \subseteq \varphi$.

To see that Ψ is one-to-one, let $\varphi, \rho \in \operatorname{Con}\mathbf{A}$ with $\theta \subseteq \varphi$ and $\theta \subseteq \rho$. Suppose that $\Psi(\varphi) = \Psi(\rho)$. This means $\varphi/\theta = \rho/\theta$. Now consider for all $a,a' \in A$

$$\begin{aligned} (a,a') \in \varphi &\Leftrightarrow (a/\theta, a'/\theta) \in \varphi/\theta \\ &\Leftrightarrow (a/\theta, a'/\theta) \in \rho/\theta \\ &\Leftrightarrow (a,a') \in \rho \end{aligned}$$

So $\varphi = \rho$. Notice that the first equivalence depends on $\theta \subseteq \varphi$ while the last depends on $\theta \subseteq \rho$. We see that Ψ is one-to-one.

To see that Ψ is onto $\operatorname{Con}\mathbf{A}/\theta$, let μ be a congruence of $\mathbf{A}/\theta$. Define

$$\varphi := \{(a,a') \mid a,a' \in A \text{ and } (a/\theta, a'/\theta) \in \mu\}.$$

This two-place relation is our candidate for a preimage of μ. First we need to see that φ is indeed a congruence of $\mathbf{A}$. The checking of reflexivity, symmetry, and transitivity is routine. To confirm the substitution property, let Q be an operation symbol (with the harmless assumption that its rank is 3). Pick $a,a',b,b',c,c' \in A$ so that

$$\begin{aligned} a &\equiv a' \mod \varphi \\ b &\equiv b' \mod \varphi \\ c &\equiv c' \mod \varphi. \end{aligned}$$

We must see that $Q^{\mathbf{A}}(a,b,c) \equiv Q^{\mathbf{A}}(a',b',c') \mod \varphi$. From the three displayed conditions we deduce

$$\begin{aligned} a/\theta &\equiv a'/\theta \mod \mu \\ b/\theta &\equiv b'/\theta \mod \mu \\ c/\theta &\equiv c'/\theta \mod \mu. \end{aligned}$$

Because μ is a congruence of $\mathbf{A}/\theta$, we obtain

$$Q^{\mathbf{A}/\theta}(a/\theta, b/\theta, c/\theta) \equiv Q^{\mathbf{A}/\theta}(a'/\theta, b'/\theta, c'/\theta) \mod \mu.$$

But given how the operations work in quotient algebras, this gives

$$Q^{\mathbf{A}}(a,b,c)/\theta \equiv Q^{\mathbf{A}}(a',b',c')/\theta \mod \mu.$$

Then the definition of φ supports the desired conclusion that $Q^{\mathbf{A}}(a,b,c) \equiv Q^{\mathbf{A}}(a',b',c') \mod \varphi$. So φ is a congruence of $\mathbf{A}$. But we also need to see that $\theta \subseteq \varphi$ to get that $\varphi \in P$. So suppose that $a,a' \in A$ with $(a,a') \in \theta$. Then $a/\theta = a'/\theta$. This entails that $(a/\theta, a'/\theta) \in \mu$ since μ is reflexive. In this way, we see that $(a,a') \in \varphi$. So $\theta \subseteq \varphi$ and $\varphi \in P$. Now consider

$$\begin{aligned} \Psi(\varphi) &= \varphi/\theta \\ &= \{(a/\theta, a'/\theta) \mid a,a' \in A \text{ and } (a,a') \in \varphi\} \\ &= \{(a/\theta.a'/\theta) \mid a,a' \in A \text{ and } (a/\theta, a'/\theta) \in \mu\} \\ &= \mu. \end{aligned}$$

In this way, we see that Ψ is onto $\operatorname{Con}\mathbf{A}/\theta$.

Last, we need to show that Ψ respects the ordering by set inclusion. So let $\varphi, \rho \in \operatorname{Con}\mathbf{A}$ with $\theta \subseteq \varphi$ and $\theta \subseteq \rho$. Let us first suppose that $\varphi \subseteq \rho$. To see that $\Psi(\varphi) \subseteq \Psi(\rho)$, let $a,a' \in A$ and notice

$$\begin{aligned} (a/\theta, a'/\theta) \in \Psi(\varphi) &\implies (a/\theta, a'/\theta) \in \varphi/\theta \\ &\implies (a,a') \in \varphi \\ &\implies (a,a') \in \rho \\ &\implies (a/\theta, a'/\theta) \in \rho/\theta \\ &\implies (a/\theta, a'/\theta) \in \Psi(\rho) \end{aligned}$$

So we find if $\varphi \subseteq \rho$, then $\Psi(\varphi) \subseteq \Psi(\rho)$. For the converse, suppose $\Psi(\varphi) \subseteq \Psi(\rho)$. Let $a,a' \in A$ and notice

$$\begin{aligned} (a,a') \in \varphi &\implies (a/\theta, a'/\theta) \in \varphi/\theta \\ &\implies (a/\theta, a'/\theta) \in \Psi(\varphi) \\ &\implies (a/\theta, a'/\theta) \in \Psi(\rho) \\ &\implies (a/\theta, a'/\theta) \in \rho/\theta \\ &\implies (a,a') \in \rho. \end{aligned}$$

So we find that if $\Psi(\varphi) \subseteq \Psi(\rho)$, then $\varphi \subseteq \rho$. So we have for all $\varphi, \rho \in P$,

$$\varphi \subseteq \rho \text{ if and only if } \Psi(\varphi) \subseteq \Psi(\rho).$$

Finally, we can conclude that Ψ is an isomorphism between our two ordered sets of congruences. □

1.1 Problem Set 1

Algebra Homework, Edition 1
Second Week
Just Some General Notions

Problem 5.
Prove that the congruence relations of **A** are exactly those subuniverses of $\mathbf{A}\times\mathbf{A}$ which happen to be equivalence relations on A.

Problem 6.
Prove that the homomorphisms from **A** to **B** are exactly those subuniverses of $\mathbf{A}\times\mathbf{B}$ which are functions from A to B.

Problem 7.
Prove that the projection functions associated with $\mathbf{A}\times\mathbf{B}$ are homomorphisms.

Problem 8.

(a) Prove that the intersection of any nonempty collection of subuniverses of **A** is again a subuniverse.

(b) Prove that the intersection of any nonempty collection of congruences of **A** is again a congruence.

Problem 9.
A collection $\mathcal{C}$ of sets is **up-directed** by $\subseteq$ provided if $U,V\in\mathcal{C}$ then there is $W\in\mathcal{C}$ such that $U,V\subseteq W$.

(a) Prove that the union on any nonempty up-directed collection of subuniverses of **A** is again a subuniverse of **A**.

(b) Prove that the union of any nonempty up-directed collection of congruences of **A** is again a congruence of **A**.

LECTURE 2

COMPREHENDING PLUS AND TIMES

2.1 WHAT A RING IS

The notion of a ring arose in the nineteenth century by generalizing a collection of specific algebraic systems built around various examples of addition and multiplication. Certainly our understanding of addition and multiplication of positive integers is very old. Eudoxus of Cnidus, a contemporary of Plato, put—in modern terms—the notions of addition and multiplication of positive real numbers on a sound basis. His work can be found in Book V of Euclid's elements. Negative numbers emerged in India and China about the time of Archimedes, but met with little welcome in the Hellenistic world. This attachment of mathematical illegitimacy to negative numbers persisted in Europe into the eighteenth century. However, by the end of the eighteenth century, not only negative real numbers but complex numbers in general were well in hand. Euler was a master of it all.

In the nineteenth century we had algebraic systems built around addition and multiplication of all of the following:

- integers
- rational numbers
- real numbers
- complex numbers
- algebraic numbers
- constructible numbers
- $n \times n$ matrices with entries selected from the systems listed above.
- polynomials with coefficients selected from certain of the systems listed above.
- functions from the reals to the reals (and similarly with the reals replaced by some other systems)
- many other examples of addition and multiplication

As that century progressed, mathematicians realized that to develop the theories of each of these particular cases, one had to duplicate more or less a lot of effort. The examples had many properties in common. So it was a matter of convenience to develop the details of many of these common properties just once, before pursuing the more specialized theory of, say, the complex numbers. This led to the notion of a ring.

The signature we use to present this notion consists of a two-place operation symbol $\cdot$ to name multiplication, a two-place operation symbol $+$ to name addition, a one-place operation symbol $-$ to denote the formation of negatives, and two constant symbols 0 and 1. A **ring** is an algebraic system of this signature in which the following equations hold true.

$$\begin{aligned} x+(y+z) &= (x+y)+z & \qquad x\cdot(y\cdot z) &= (x\cdot y)\cdot z \\ x+0 &= x & x\cdot 1 &= x \\ x+y &= y+x & 1\cdot x &= x \\ -x+x &= 0 & x\cdot(y+z) &= x\cdot y + x\cdot z \\ & & (x+y)\cdot z &= x\cdot z + y\cdot x \end{aligned}$$

This collection of equations is sometimes called the axioms of ring theory.

You see here the familiar associative, commutative, and distributive laws, as well as equations giving the behavior of 0 and 1. It is important to realize that while the commutative law for addition is included, the commutative law for multiplication is not. The absence of the commutative law for multiplication has compelled me to include two forms of the distributive law as well as two equations to capture the behavior of 1. The ring of 2 × 2 matrices with real entries is an example of a ring where the commutative law for multiplication fails. A ring in which the commutative law for multiplication holds as well is called a **commutative ring**. While there is a rich theory of rings in general, in our course almost all rings will be commutative rings.

Because the axioms of ring theory are all equations it is easy to see that every subalgebra of a ring must be a ring itself, that every homomorphic image of a ring must also be a ring, and that the direct product of any system of rings is again a ring. Because the commutative law for multiplication is also an equation, the same observations apply to commutative rings.

You should also realize that in a ring the elements named by 0 and 1 might be the same. In this event, by way of a fun exercise, you can deduce from the ring axioms that such a ring can have only one element. Evidently, all one-element rings are isomorphic and, of themselves, not very interesting. They are called *trivial rings*.

According to the definition above, every ring must have an element named by the constant symbol 1 and this element must behave as described by the equations in our list. This has been the most common convention since the 1970's. However, some considerable part of the older literature and some of the contemporary literature use a different somewhat wider notion that lacks the constant symbol 1. For example, the even integers under ordinary addition and multiplication would constitute a ring in this manner, but not in the sense that I have put forward here. In that style of exposition, what we have called "rings" are referred to as "rings with unit". Nathan Jacobson, one of the great ring theorists of the twentieth century, used the notion of ring I have adopted and referred to these other old-fashion algebraic systems as "rngs".

2.2 Congruences and Ideals on Rings

Let $\mathbf{R}$ be a ring and let θ be a congruence on $\mathbf{R}$. Recall that

$$0/\theta = \{a \mid a \in R \text{ and } a \equiv 0 \mod \theta\}$$

is the θ-congruence class containing 0. Observe that the set $0/\theta$ has each of the following properties

(a) $0 \in 0/\theta$.

(b) If $a, b \in 0/\theta$, then $a + b \in 0/\theta$.

(c) if $a \in 0/\theta$ and $r \in R$, then $ra, ar \in 0/\theta$.

To obtain (b) reason as follows

$$\begin{aligned} a &\equiv 0 \mod \theta \\ b &\equiv 0 \mod \theta \\ a + b &\equiv 0 + 0 \mod \theta \\ a + b &\equiv 0 \mod \theta \end{aligned}$$

The third step uses the key substitution property of congruence relations, whereas the fourth step use the equation $0 + 0 = 0$, which follows easily from the ring axioms.

To obtain (c) reason as follows

$$\begin{aligned} a &\equiv 0 \mod \theta \\ r &\equiv r \mod \theta \\ ar &\equiv 0r \mod \theta \\ ar &\equiv 0 \mod \theta \end{aligned}$$

The second step uses the fact that congruence relations, being special equivalence relations, are reflexive. The last step uses the equation $0x = 0$, which can be deduced from the ring axioms. A similar line of reasoning produces the conclusion

$$ra \equiv 0 \mod \theta.$$

Any subset $I \subseteq R$ that has the three attributes listed above for $0/\theta$ is called an **ideal** of the ring **R**. This means that I is an ideal of **R** if and only if

(a) $0 \in I$.

(b) If $a, b \in I$, then $a + b \in I$.

(c) if $a \in I$ and $r \in R$, then $ra, ar \in I$.

So we have taken the definition of ideal to allow us to observe that in any ring **R**

if θ is a congruence relation of **R**, then $0/\theta$ is an ideal of **R**.

That is, every congruence relation gives rise to an ideal.

The converse is also true. Let **R** be a ring and let I be an ideal of **R**. Define

$$\theta_I := \{(a, b) \mid a, b \in R \text{ and } a - b \in I\}.$$

The eager graduate students should check that θ_I is indeed a congruence relation of **R**. Actually, the theorem below tells a fuller tale and its proof, which only requires pursuing all the definitions involved, is left to delight the graduate students.

Theorem on Ideals and Congruences. *Let* **R** *be any ring, let* θ *be a congruence relation of* **R** *and let* I *be any ideal of* **R**. *All of the following hold.*

(a) $0/\theta$ *is an ideal of* **R**.

(b) θ_I *is a congruence relation of* **R**.

(c) $I = 0/(\theta_I)$.

(d) $\theta = \theta_{0/\theta}$.

(e) *The collection of all ideals of* **R** *is ordered by* $\subseteq$ *and the map* $I \mapsto \theta_I$ *is an isomorphism of the ordered set of all ideals of* **R** *with the ordered set of all congruence relations of* **R**.

The significance of this theorem is that when dealing with rings we can replace the study of congruence relations with the study of ideals. After all, each congruence θ is a set of ordered pairs, that is $\theta \subseteq R \times R$.; whereas each ideal I is merely a set of elements of R, that is $I \subseteq R$. Of course, there are places, in ring theory, where congruence relations are more convenient than ideals, so we need to remember both.

Here is some notation for using ideals in place of congruences. Let **R** be any ring and let θ and I be a congruence relation and an ideal that correspond to each other, let $a, b \in R$ and let J be an ideal of **R** so that $I \subseteq J$.

$$\begin{aligned} \mathbf{R}/I &:= \mathbf{R}/\theta \\ a + I &:= a/\theta = \{a + b \mid b \in I\} \\ J/I &:= \{b + I \mid b \in J\} = 0/(\theta_J/\theta_I) \\ a \equiv b \mod I \text{ means } a &\equiv b \mod \theta \\ &\text{means } a - b \in I \end{aligned}$$

The graduate student should work out the details to see that these conventions really do the job. Incidentally, the notation $a + I$ is a special case of $U + V := \{u + v \mid u \in U \text{ and } v \in V\}$, where $U, V \subseteq R$.

Suppose that **R** is a ring and $h : \mathbf{R} \to \mathbf{S}$ is a homomorphism. The **kernel** of h is the following set

$$\ker h := \{a \mid a \in R \text{ and } h(a) = 0\}.$$

The graduate students should check that if θ denotes that functional kernel of h, then

$$\ker h = 0/\theta.$$

So $\ker h$ is an ideal of **R** and the congruence corresponding to this ideal is the functional kernel of h.

2.3 The Isomorphism Theorems for Rings

With this sort of lexicon in hand, all the isomorphism theorems can be rendered into ring theoretic versions, with no need for further proofs. Here they are.

The Homomorphism Theorem, Ring Version. *Let* **R** *be a ring, let* $f : \mathbf{R} \twoheadrightarrow \mathbf{S}$ *be a homomorphism from* **R** *onto* **S**, *and let* I *be an ideal of* **R**. *All of the following hold.*

(a) *The kernel of* f *is an ideal of* **R**.

(b) $\mathbf{R}/I$ *is a ring.*

(c) *The map* η *that assigns to each* $a \in R$ *the congruence class* $a + I$ *is a homomorphism from* **R** *onto* $\mathbf{A}/I$ *and its kernel is* I.

(d) *If* I *is the kernel of* f, *then there is an isomorphism* g *from* $\mathbf{R}/I$ *to* **S** *such that* $f = g \circ \eta$.

The Second Isomorphism Theorem, Ring Version. *Let* $\mathbf{R}$ *be a ring, let* I *be an ideal of* $\mathbf{R}$*, and let* $\mathbf{S}$ *be a subring of* $\mathbf{R}$*. Then each of the following hold.*

(a) $I \cap S$ *is an ideal of* $\mathbf{S}$.

(b) $I + S$ *is a subuniverse of* $\mathbf{R}$.

(c) $(\mathbf{I}+\mathbf{S})/I \cong \mathbf{S}/(I \cap S)$.

The Third Isomorphism Theorem, Ring Version. *Let* $\mathbf{R}$ *be a ring and let* I *and* J *be ideals of* $\mathbf{R}$ *with* $I \subseteq J$*. Then*

(a) J/I *is an ideal of* $\mathbf{R}/I$*, and*

(b) $\mathbf{R}/I \Big/ J/I \cong \mathbf{R}/J$.

The Correspondence Theorem, Ring Version. *Let* $\mathbf{R}$ *be a ring and let* I *be an ideal of* $\mathbf{R}$*. Let* $P = \{J \mid J \text{ is an ideal of } \mathbf{R} \text{ and } I \subseteq J\}$*. Then the map from* P *to the ordered set of ideals of* $\mathbf{R}/I$ *that sends each* $J \in P$ *to* J/I *is an isomorphism between the ordered set* $\langle P, \subseteq \rangle$ *and the ordered set of ideals of* $\mathbf{R}/I$*.*

2.4 Dealing with Ideals

Let $\mathbf{R}$ be a ring. Then R and $\{0\}$ will be ideals of $\mathbf{R}$. (They might be the same ideal, but only if $\mathbf{R}$ is a one-element ring). By a **proper ideal** of $\mathbf{R}$ we mean one that is different from R. By a **nontrivial ideal** we mean one that is different from $\{0\}$. The collection of all ideals of $\mathbf{R}$ is ordered by $\subseteq$. Under this ordering, $\{0\}$ is the unique least ideal and R is the unique largest ideal.

Let $\mathbf{R}$ be a ring and let $\mathcal{K}$ be any nonempty collection of ideals of $\mathbf{R}$. It is a routine exercise (why not put pen to paper?) that $\bigcap\mathcal{K}$ is also an ideal of $\mathbf{R}$ and this ideal is the greatest (in the sense of $\subseteq$) ideal included in every ideal belonging to $\mathcal{K}$. So every nonempty collection of ideals has a greatest lower bound in the ordered set of ideals. Let $W \subseteq R$ and take $\mathcal{K} = \{I \mid I \text{ is an ideal of } \mathbf{R} \text{ and } W \subseteq I\}$. Then $\bigcap\mathcal{K}$ is the smallest ideal of $\mathbf{R}$ that includes W. This ideal is denoted by (W) and is called the **ideal generated by** W.

Unlike the situation with intersection, when $\mathcal{K}$ is a nonempty collection of ideals of the ring $\mathbf{R}$ it is usually not the case that the union $\bigcup\mathcal{K}$ will turn out to be an ideal. However, $(\bigcup\mathcal{K})$ will be an ideal—indeed, it is the least ideal in the ordered set of ideals that includes every ideal in $\mathcal{K}$.

So the collection of all ideals of any ring is an ordered set with a least member, a greatest member, and every nonempty collection of ideals has both a greatest lower bound and a least upper bound. Such ordered sets are called **complete lattice-ordered sets**.

While in general the union of a collection of ideals is unlikely to be an ideal, there are collections for which the union is an ideal. A collection $\mathcal{K}$ of ideals is said to be **updirected** provided if $I, J \in \mathcal{K}$, then there is $K \in \mathcal{K}$ so that $I \subseteq K$ and $J \subseteq K$.

Theorem 2.4.1. *Let* $\mathbf{R}$ *be a ring and let* $\mathcal{K}$ *be a nonempty updirected collection of ideals of* $\mathbf{R}$*. Then* $\bigcup\mathcal{K}$ *is an ideal of* $\mathbf{R}$*.*

Proof. First observe that $0 \in \bigcup\mathcal{K}$, since $\mathcal{K}$ is nonempty and every ideal must contain 0.

Now suppose that $a, b \in \bigcup\mathcal{K}$. Pick $I, J \in \mathcal{K}$ so that $a \in I$ and $b \in J$. Because $\mathcal{K}$ is updirected, pick $K \in \mathcal{K}$ so that $I \cup J \subseteq K$. So $a, b \in K$. Because K is an ideal, we see $a + b \in K \subseteq \bigcup\mathcal{K}$.

Finally, suppose $a \in \bigcup\mathcal{K}$ and $r \in R$. Pick $I \in \mathcal{K}$ so that $a \in I$. Then $ar, ra \in I$ since I is an ideal. Hence $ar, ra \in \bigcup\mathcal{K}$

In this way, we see that $\bigcup\mathcal{K}$ is an ideal. □

One kind of updirected set is a chain. The collection $\mathcal{C}$ is a **chain** of ideals provided for all $I, J \in \mathcal{C}$ either $I \subseteq J$ or $J \subseteq I$. As a consequence, we see that the union of any nonempty chain of ideals is again an ideal.

A little reflection shows that this result is not particularly ring theoretic. In fact, for algebras generally, the union of any updirected collection of congruence relations is again a congruence relation.

Now let **R** be a ring and $W \subseteq R$. The ideal (W) that is generated by W was defined in what might be called a shrink wrapped manner as the intersection of all the ideals containing W. It is also possible to describe this ideal by building it up from W in stages using the following recursion.

$$\begin{aligned} W_0 &:= W \cup \{0\} \\ W_{n+1} &:= W_n \cup \{ra \mid r \in R \text{ and } a \in W_n\} \cup \{ar \mid r \in R \text{ and } a \in W_n\} \cup \{a+b \mid a, b \in W_n\} \\ &\quad \text{for all natural numbers } n. \end{aligned}$$

Notice $W \subseteq W_0 \subseteq W_1 \subseteq W_2 \subseteq \dots$ and each set along this chain repairs potential failures of the earlier sets along the chain to be ideals. It does this by adding new elements. Unfortunately, these new elements, while they repair earlier failures may introduce failures of their own. For this reason the construction continues through infinitely many stages. Now let $W_\omega := \bigcup_{n\in\omega} W_n$ be the union of this chain of sets. Our expectation is that all the failures have been fixed and that W_ω is an ideal. The eager graduate students are invited to write out a proof of this. But more is true. Actually, $W_\omega = (W)$. Here are some suggestions for how to prove this. To establish $W_\omega \subseteq (W)$ prove by induction on n that $W_n \subseteq I$ for every ideal I that includes W. Observe that $(W) \subseteq W_\omega$ once we know that W_ω is an ideal that includes W.

This process that shows that shrink wrapping and building up from the inside works not only here in the context of ideals, but in several other contexts as well.

A more transparent version of the building up from the inside is available in our particular context. By a *combination* of W over **R** we mean an element of the form

$$r_0 w_0 s_0 + r_1 w_1 s_1 + \dots + r_{n-1} w_{n-1} s_{n-1}$$

where n is a natural number, $r_0, s_0, r_1, s_1, \dots, r_{n-1}, s_{n-1} \in R$, and $w_0, w_1, \dots, w_{n-1} \in W$. In case $n = 0$, we take the element represented to be the zero of the ring. It is straightforward, with the help of the distributive laws, to see that the set of all combinations of W over **R** is an ideal that includes the subset W. An induction on the length of combinations shows that all these combinations belong to (W). So the set of all combinations of W over **R** must be the ideal (W) generated be W. It is important to observe that in the combination displayed above we have not assumed that the w_i's are distinct. In commutative rings it is only necessary to consider combinations of the form

$$r_0 w_0 + r_1 w_1 + \dots + r_{n-1} w_{n-1}.$$

Moreover, in this case we can insist that the w_i's be distinct. In particular, if **R** is commuative , $w \in R$, and I is an ideal of **R**, then the ideal $(\{w\} \cup I)$ generated by the element w and the ideal I consists of all elements of the form

$$rw + u \text{ where } r \in R \text{ and } u \in I.$$

2.5 Problem Set 2

Algebra Homework, Edition 2
Third Week
Prime Ideals

Problem 10.

(a) Let I and J be ideals of a commutative ring **R** with $I + J = R$. Prove that $IJ = I \cap J$.

(b) Let I, J, and K be ideals of a principal ideal domain. Prove that $I \cap (J + K) = I \cap J + I \cap K$.

Problem 11.
Let **R** be a commutative ring and I be a proper prime ideal of **R** such that $\mathbf{R}/I$ satisfies the descending chain condition on ideals. Prove that $\mathbf{R}/I$ is a field.

Problem 12.
Let **R** be a commutative ring and I be an ideal which is contained in a prime ideal P. Prove that the collection of prime ideals contained in P and containing I has a minimal member.

Problem 13.
Let X be a finite set and let **R** be the ring of functions from X into the field $\mathbb{R}$ of real numbers. Prove that an ideal M of **R** is maximal if and only if there is an element $a \in X$ such that

$$M = \left\{ f \mid f \in R \text{ and } f(a) = 0 \right\}.$$

Problem 14.
Let **R** be a commutative ring and suppose the I, J, and K are ideals of **R**. Prove that if $I \subseteq J \cup K$, then $I \subseteq J$ or $I \subseteq K$.

LECTURE 3

RINGS LIKE THE INTEGERS

3.1 INTEGRAL DOMAINS

The ring $\langle \mathbb{Z}, +, \cdot, -, 0, 1\rangle$ of integers is one of the most familiar mathematical objects. Its investigation lies at the heart of number theory that, together with geometry, is among the oldest parts of mathematics. This ring is commutative and has a host of other very nice properties. Among these is that the product of any two nonzero integers must itself be nonzero. This property may fail, even in rings closely connected to the ring of integers. For example, let **R** be the direct square of the ring of integers. The elements of this ring will be ordered pairs of integers with the ring operations defined coordinatewise. That is

$$\begin{aligned}(a,b)+(c,d) &= (a+c,b+d)\\ (a,b)\cdot(c,d) &= (ac,bd)\\ -(a,b) &= (-a,-b)\end{aligned}$$

The zero of **R** is the pair $(0,0)$ while the unit (the one) is $(1,1)$. But observe that the product of $(1,0)$ with $(0,1)$ is $(1\cdot 0, 0\cdot 1) = (0,0)$.

A ring **D** is called an **integral domain** provided

(a) **D** is a commutative ring,

(b) 0 and 1 name different elements of D, and

(c) If $a,b \in D$ and $a \neq 0 \neq b$, then $ab \neq 0$.

Integral domains used to be called by a more charming name: domains of integrity. Condition (b) above is equivalent to the stipulation that integral domains must have at least two elements. Condition (c) can be replaced by either of the following conditions.

(c′) If $a,b \in D$ and $ab = 0$, then either $a = 0$ or $b = 0$.

(c″) If $a,b,c \in D$ with $a \neq 0$ and $ab = ac$, then $b = c$

Condition (c′) is just a contrapositive form of Condition (c). Condition (c″) is the familiar cancellation law. The graduates student can find amusement by showing the equivalence of this condition.

While, as observed above, the direct product of a system of integral domains need not be an integral domain (is it ever?), every subring of an integral domain will be again an integral domain. What about homomorphic images of integral domains? Well, the trivial one-element ring is a homomorphic image of every ring, including every integral domain, and the trivial ring is not an integral domain. But suppose **D** is an integral domain and h is a homomorphism mapping **D** onto the nontrivial ring **S**. Must **S** be an integral domain? Certainly, conditions (a) and (b) hold for **S**. Consider a concrete example. Let I be the set of integers that are multiples of 4. It is easy to check that I is an ideal of the ring of integers. The quotient ring $\mathbb{Z}/I$ has just four elements:

$$0+I \qquad 1+I \qquad 2+I \quad \text{and} \quad 3+I.$$

In the quotient ring we have the product $(2+I)\cdot(2+I) = 2\cdot 2+I = 4+I = 0+I$. This violates condition (c) in the definition of integral domain. So while some homomorphic images of some integral domain will be integral domains, it is not true generally. Perhaps some property of the ideal I would ensure that the quotient ring is an integral domain.

Let **R** be a commutative ring and let I be an ideal of **R**. I is said to be a **prime ideal** provided

- I is a proper ideal of **R** [that is, $I \neq R$], and
- if $a, b \in R$ with $ab \in I$, then either $a \in I$ or $b \in I$.

The graduate students can prove the following theorem by chasing definitions.

Theorem 3.1.1. *Let* **R** *be a commutative ring and let I be an ideal of* **R**. **R**/I *is an integral domain if and only if I is a prime ideal of* **R**.

Suppose **R** is a ring. Consider the list of elements of R below:

$$1, 1+1, 1+1+1, 1+1+1+1, \ldots.$$

This looks like a list of the positive integers, but we mean something different. The element 1 is the unit of multiplication in **R** and + names the addition operation in **R**. The ring **R** may not contain any integers at all. The list above might even be finite, depending on the ring **R**. If the list is infinite we say that **R** has **characteristic** 0. If the list is finite, then (as pigeons know) two distinct members of this list must actually be the same element. That is

$$\underbrace{1+\cdots+1}_{n \text{ times}} = \underbrace{1+\cdots+1}_{n \text{ times}} + \underbrace{1+\cdots+1}_{k \text{ times}}$$

for some positive natural numbers n and k. This entails that

$$0 = \underbrace{1+\cdots+1}_{k \text{ times}}$$

for some positive natural number k. In this case, we say that the **characteristic** of **R** is the smallest such positive natural number. On reflection, it might have been better to say that rings of characteristic 0 had infinite characteristic. However, the use of characteristic 0 for this notion is so well entrenched that we are stuck with it.

The characteristic of a ring **R** is a useful invariant of **R**. It will play a prominent role in the spring semester during our development of the theory of fields. Observe that every finite ring must have a characteristic that is not 0. Because 1 must belong to every subring of **R**, we see that all the subrings of **R** have the same characteristic as **R**. On the other hand, the homomorphic images of **R** may have characteristic differing from the characteristic of **R**. To begin with, trivial rings have characteristic 1 (these are the only rings of characteristic 1) and trivial rings are homomorphic images of every ring. The ring of integers has characteristic 0, but $\mathbb{Z}/(6)$ evidently has characteristic 6. On the other hand, it is easy to verify (do it, why not?) that the characteristic of a homomorphic image of $\mathbf{R}$ can be no larger than the characteristic of $\mathbf{R}$ (well, taking 0 to be larger than all the positive natural numbers. . .). We leave it to the eager graduate students to figure out the characteristic of $\mathbf{R} \times \mathbf{S}$ when the characteristic of $\mathbf{R}$ is r and the characteristic of $\mathbf{S}$ is s.

Here is a useful fact.

Fact. Let $\mathbf{D}$ be an integral domain. The characteristic of $\mathbf{D}$ is either 0 or it is a prime number.

We won't prove this, but here is a hint as to why an integral domain cannot have characteristic 6.

$$0 = 1+1+1+1+1+1 = 1 \cdot (1+1+1) + 1 \cdot (1+1+1) = (1+1) \cdot (1+1+1).$$

3.2 Principal Ideal Domains

A route to a deeper understanding of the ring of integers is to investigate the congruence relations of this ring. This is the route chosen by Gauss in his 1801 masterpiece *Disquistiones Arithmeticæ.* Of course, we see that the investigation of congruences of a ring amounts to the investigation of its ideals. The notion of an ideal of a ring arose in the work of Kummer, Kronecker, and Dedekind in the second half of the nineteenth century to be refined still later by Hilbert and by Emmy Noether. Still, the discoveries of Gauss needed changes of only the most modest kind to fit with the later theoretical apparatus.

We begin with an important observation that surely must have been known to Euclid.

A Key Fact About the Integers. *Let d be any nonzero integer and let n be any integer. There are unique integers q and r satisfying the following constraints:*

(a) *$n = qd + r$, and*

(b) *Either $r = 0$ or $0 < r < |d|$.*

Graduate students with itchy fingers who turn their hands to this are advised that there are two things to show: the *existence* of integers q and r and the *uniqueness* of these integers. Here is a hint. Consider the set $\{|n - xd| \mid x \in \mathbb{Z}\}$. This is a set of natural numbers. It is nonempty (why?). Every nonempty set of natural numbers has a least element.

The uniquely determined integers q and r mentioned in this Key Fact are called the *quotient* of n upon division by d and the *remainder* of n upon division by d, respectively. We will also call r the *residue* of n upon division by d.

Let I be any nontrivial ideal of the ring of integers. Since I is not trivial, it must have a member other than 0 and, because I is an ideal, there must be a positive integer in I. Hence there must be a least positive integer d in I. Now let $n \in I$ be chosen arbitrarily. Using the Key Fact, pick integers q and r so that

(a) $n = qd + r$, and

(b) Either $r = 0$ or $0 < r < |d|$.

Then $r = n - qd$. Notice that $n, d \in I$ because that's the way we chose them. So $r = n - qd \in I$ because I is an ideal. But $0 < r < |d| = d$ is impossible, by the minimality of the choice of d. So we conclude that $r = 0$ and therefore that n is a multiple of d. Thus

$$I = \{qd \mid q \in \mathbb{Z}\} = (d).$$

So we have the conclusion that every ideal of the ring of integers is generated by some one of its members (and, in fact, by the smallest postive integer belonging to the ideal if the ideal in not trivial).

A **principal ideal domain** is an integral domain for which every ideal is generated by some one of its members. In an arbitrary ring, we will say an ideal is **principal** provided it is generated by some one of its members. So a principal ideal domain is an integral domain for which every ideal is principal.

The ring of integers is a principal ideal domain. Many interesting properties of the ring of integers also hold for principal ideal domains in general. This includes the powerful Fundamental Theorem of Arithmetic:

> *Every nonzero integer, other than* 1 *and* -1*, can be written in a unique way as a product of primes.*

In order to formulate this result for rings more generally, we need to introduce some further notions.

A **unit** in a commutative ring is an element u such that there is an element v in the ring so that $uv = 1 = vu$. So a unit is just an element with a multiplicative inverse. The units of the ring of integers are just 1 and -1. (Notice the appearance of these numbers in the statement above.) Two elements a and b of a commutative ring are said to be **associates** provided $au = bu$ for some unit u. It is routine (and you know the routine when the word routine comes up in these notes...) to show that relation "is an associate of" is an equivalence relation on any commutative ring. We will use $a \sim b$ to denote that a and b are associates. Do you think $\sim$ is a congruence relation on the ring?

An element a of a commutative ring is said to be **irreducible** provided it is neither 0 nor a unit and if $a = bc$ for some elements b and c in the ring, then either b is a unit or c is a unit. So irreducible elements of a ring are the ones that cannot be factored, except in some trivial manner. (Observe that $2 = (-1)\cdot(-1)\cdot 1$ is a factorization of the integer 2 is such a trivial manner.).

An integral domain **D** is said to be a **unique factorization domain** provided

(a) Every nonzero nonunit in D can be expressed as a (finite) product of irreducibles.

(b) If m and n are natural numbers and $a_0, a_1, \dots, a_{m-1} \in D$ and $b_0, b_1, \dots, b_{n-1} \in D$ are irreducibles such that
$$a_0 a_1 \dots a_{m-1} \sim b_0 b_1 \dots b_{n-1},$$
then $m = n$ and there is a permutation σ of $\{0, 1, \dots, m-1\}$ so that
$$a_i \sim b_{\sigma(i)} \text{ for all } i \text{ with } 0 \le i < m.$$

The point of the permutation σ is that we don't really want to consider $2\cdot 3$ and $3\cdot 2$ as distinct factorizations of 6. Observe that stipulation (a) asserts the existence of a factorization into irreducibles, while stipulation (b) asserts the uniqueness of such factorization.

The Fundamental Theorem of Arithmetic asserts that the ring of integers is a unique factorization domain. So is every principal ideal domain and that is what we tackle below.

You might wonder that we have used the word "irreducible" instead of "prime" in formulating these notions. (You might also be wondering now if prime ideals have anything to do with primes....) Euclid realized long ago that an irreducible (positive) integer p had the property

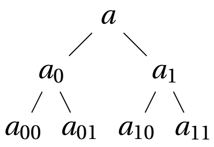
$$\text{If } p \mid ab, \text{ then either } p \mid a \text{ or } p \mid b.$$

Here $p \mid a$ means that p divides a—-that is, $a = pc$ for some integer c. The divisibility relation, denoted by $\mid$, makes sense in any ring and we use it without further comment.

We will say that an element p of a commutative ring is **prime** provided p is neither 0 nor a unit and for all a and b in the ring
$$\text{If } p \mid ab, \text{ then either } p \mid a \text{ or } p \mid b.$$

The **primeness condition** is just that every irreducible element is prime. Incidentally, the converse is always true in any integral domain: a prime element is always irreducible. Indeed, if a is prime and $a = bc$, then we see that either $a \mid b$ or $a \mid c$. Consider, for instance, the first alternative. Pick d so that $b = ad$. Then $a \cdot 1 = bc = adc$. Now cancel a (we are in a integral domain) to obtain $1 = dc$. This means that c is a unit. The second alternative is similar.

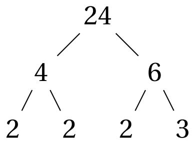
An attempt to factor a nonzero nonunit a into irreducibles might look like this:



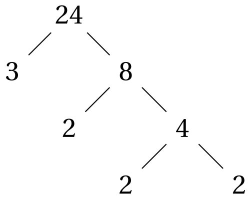
This tree represents two steps in an attempt to factor a.

$$a = a_0 a_1$$
$$a_0 = a_{00} a_{01}$$
$$a_1 = a_{10} a_{11}$$

So we have the factorization $a = a_{00} a_{01} a_{10} a_{11}$. The diagram displayed is a tree (look at it while standing on your head) with three levels. Each node branches into two succeeding nodes (except the nodes on the bottom level). This tree has four branches that start at the root (a) and extend down to the bottom level. Now our intention is that all the nodes should themselves be nonzero nonunits. So if we run into an irreducible then we will not attempt to factor it. Here is a tree showing a factorization of the integer 24.



Suppose we try again. Here is another way to factor 24.



These trees and their labellings reflect the actual processes of the factorizations. We see that they are not unique. But the irreducibles (counting how often they appear but not the order of their appearance) is unique. In each of these trees, every node has either 0 or 2 succeeding nodes, since multiplication is a two-place operation. In any case, each node has only finitely many nodes as immediate successors. We say the tree is *finitely branching.* There is a useful combinatorial fact about trees that comes into play.

König's Infinity Lemma. *Any finitely branching tree with infinitely many nodes must have an infinite branch.*

Proof. We can build the desired infinite branch by the following recursion.

Let a_0 be the root of the tree. There are only finitely many nodes immediately below a_0. Every node, apart from a_0 lies somewhere below a_0. Since the union of finitely many finite sets is always finite, there must be a node immediately below a_0 which itself has infinitely many nodes below it. Let a_1 be such a node. Now apply the same reasoning to a_1 to obtain a node immediately below a_1 that is above infinitely many nodes. Continuing in this way, we obtain a branch $a_0, a_1, a_2, \ldots$ that is infinite. □

The graduate students should be a bit unhappy with the informality of this proof. For one thing, it describes an infinite process. For another it is not terribly specific about how to pick any of the nodes along the infinite branch, apart from a_0. Producing the infinite branch requires making infinitely many choices. These issues might be addressed in two stages. The first stage would secure the validity of definition by recursion. To see what is at issue consider the following familiar definition of the factorial function.

$$\begin{aligned} 0! &= 1 \\ (n+1)! &= n!(n+1) \text{ for all natural numbers } n \end{aligned}$$

The issue is two-fold: first, is there any function, here indicated by !, that fulfills the two conditions laid out above? Second, is there exactly one such function? After all definitions should be, well, definite. Here is a slightly more general situation. Suppose that a is a member of some set U and h is a function from $U \times \mathbb{N}$ into U. Is there exactly one function f from the natural numbers to U satisfying the following constraints?

$$\begin{aligned} f(0) &= a \\ f(n+1) &= h(f(n), n+1) \text{ for all natural numbers } n. \end{aligned}$$

The answer to this question is YES. It is among the simplest cases of a theorem known as the Recursion Theorem. You might try to prove this—remember there is an existence part and a uniqueness part. Induction may help in your proof.

After securing some version of the Recursion Theorem in the first stage, the second stage of cleaning up König's Infinity Lemma is to remove the ambiguity about how to pick the "next element of the infinite branch". This amounts to producing a suitable function to play the role of h in your definition by recursion. Here is what you need h to accomplish. Call a node in the tree *good* provide there are infinitely many nodes beneath it. Given a good node c we see that the set of good nodes immediately beneath it is always a nonempty set. We want $h(c, n+1)$ to pick some element of this nonempty set. (In our case, h turns out not to depend on its second input.) Functions like h always exist. They are called choice functions.

A commutative ring has the **divisor chain condition** provided whenever $a_0, a_1, a_2, \ldots$ are elements of the ring so that $a_{k+1} \mid a_k$ for all natural numbers k, then there is a natural number n so that $a_n \sim a_{n+k}$ for all natural numbers k. This means that, ignoring the distinction between associates, every descending divisor chain is finite.

Theorem Characterizing Unique Factorization Domains. *Let* **D** *be an integral domain.* **D** *is a unique factorization domain if and only if* **D** *has both the primeness condition and the divisor chain condition.*

Proof. First, suppose that **D** has the divisor chain condition and the primeness condition. Let $a \in D$ be any nonzero nonunit. Consider any factorization tree with root a. This tree is finitely branching (in fact, the branching is bounded by 2) and it cannot have any infinite branch, according the the divisor chain condition. By König the factorization tree is finite. So we see that a can be written as a product of irreducibles.

Now let $a_0 \ldots a_{m-1} \sim b_0 b_1 \cdots b_{n-1}$ be products of irreducibles. We assume, without loss of generality, that $n \le m$. We will deduce the required uniqueness by induction on m. Leaving in the hands of the capable graduate students the base step ($m = 0$) of the inductive argument, we turn to the inductive step. Let $m = k+1$. Now since a_k is irreducible, the primeness condition ensures that it is also prime. Evidently, $a_k \mid b_0 \ldots b_{n-1}$. A little (inductive) thought shows us that since a_k is prime there must be $j < n$ so that $a_k \mid b_j$. Since b_j is irreducible, we find that $a_k \sim b_j$. Using the cancellation law (we are in an integral domain!) we see that

$$a_0 a_1 \ldots a_{k-1} \sim b_0 \ldots b_{j-1} b_{j+1} \cdots b_{n-1}$$

or something easier if $j = n-1$. The left side has $k = m-1$ factors in the product whereas the right side has $n-1$ factors. Applying the induction hypothesis, we find that $m-1 = n-1$ (and hence that $m = n$) and we can pick a one-to-one map σ' from $\{0, 1, \ldots, m-2\}$ onto $\{0, 1, \ldots, j-1\} \cup \{j+1, \ldots, n-2\}$ so that

$$a_i \sim b_{\sigma'(i)} \text{ for all } i < m-1.$$

Now extend σ' to the set $\{0,1,2,\ldots,m-1\}$ by putting $\sigma(m-1)=j$. Then σ is a permutation of $\{0,1,2,\ldots,m-1\}$ that fulfills the uniqueness requirement. So **D** is a unique factorization domain.

Second, suppose for the converse, that **D** is a unique factorization domain. Let us check the divisor chain condition. Let $\cdots \mid a_2 \mid a_1 \mid a_0 = a$ be a divisor chain that is proper in the sense that no entry in the chain is an associate of any other entry. We must show that this chain is finite. For i less than the length of our chain, pick b_{i+1} so that $a_i = b_{i+1}a_{i+1}$. (This will be a proper factorization with neither a_{i+1} nor b_{i+1} being units.) Let $a = c_0 \ldots c_{n-1}$ be a factorization of a into irreducibles. Suppose, for contradiction, that our divisor chain has more than n entries. Notice

$$c_0c_1\ldots c_{n-1} = a = b_0b_1\ldots b_{n-1}b_na_n.$$

Each of $b_0,\ldots,b_n$ as well as a_n can be written as a product of irreducibles. Clearly the right side of the equation above has more factors that the left side. This violates the unique factorization property, providing the contradiction we seek. So we find that every unique factorization domain has the divisor chain condition.

To see that primeness condition, suppose $a,b,c\in D$ where a is irreducible and $a \mid bc$. Pick $d \in D$ so that $bc = ad$. Factor $b = b_0\ldots b_{m-1}, c = c_0\ldots c_{n-1}$ and $d = d_0\ldots d_{\ell-1}$ into irreducibles. This gives

$$b_0\ldots b_{m-1}c_0\ldots c_{n-1} = ad_0\ldots d_{\ell-1}$$

By the uniqueness of factorizations, there must be j so that either $a \sim b_j$ (and $j<m$) or $a\sim c_j$ (and $j<n$). In the first alternative, we get $a \mid b$ while in the second we get $a \mid c$. □

Example. The ring $\mathbb{Z}[\sqrt{-5}]$ is an integral domain that is not a unique factorization domain.

Proof. The ring $\mathbb{Z}[\sqrt{-5}]$ is, by definition, the smallest subring of the field $\mathbb{C}$ of complex numbers that includes $\mathbb{Z}\cup\{\sqrt{-5}\}$. Since it is a subring of a field it must be an integral domain. You probably see that

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

To see that $\mathbb{Z}[\sqrt{-5}]$ is not a unique factorization domain consider the following factorizations of 9.

$$9 = 3\cdot 3 = (2+\sqrt{-5})(2-\sqrt{-5}).$$

What we need is to show that each of $3, 2+\sqrt{-5}$, and $2-\sqrt{-5}$ are irreducible and the none of these is an associate of any other of them. We do this with the help of a function $N\colon \mathbb{Z}[\sqrt{-5}]\to\mathbb{N}$ defined by

$$N(a+b\sqrt{-5}) = a^2+5b^2 \text{ for all integers } a \text{ and } b.$$

This function has the following nice properties.

- $N(0)=0$.
- $N(1)=1$.
- $N(rt)=N(r)N(t)$ for all $r,t\in\mathbb{Z}[\sqrt{-5}]$.

Functions with these nice properties are sometimes called *norms*.

First, let's determine the units of $\mathbb{Z}[\sqrt{-5}]$. Suppose that u is unit and pick v so that $uv=1$. Then

$$1 = N(1) = N(uv) = N(u)N(v).$$

Since N outputs natural numbers, we see that $N(u)=1$. Pick integers a and b so that $u = a+b\sqrt{-5}$. Then

$$1 = N(u) = N(a+b\sqrt{-5}) = a^2+5b^2.$$

Notice that $5b^2$ cannot be 1. It follows that $b = 0$ and $a = 1$ or $a = -1$. This means that our unit u is either 1 or -1. So we find that the units of $\mathbb{Z}[\sqrt{-5}]$ are just 1 and -1. It follows at once that none of $3, 2+\sqrt{-5}$, and $2-\sqrt{-5}$ is an associate of any other of them.

It remains to see that our three members listed of $\mathbb{Z}[\sqrt{-5}]$ are irreducible. Below is an argument for $2+\sqrt{-5}$. I leave the other two listed elements in the capable hands of the graduate students. Pick $r, t \in \mathbb{Z}[\sqrt{-5}]$ so that $2+\sqrt{-5} = rt$. We need to see that one of r and t is a unit. So consider

$$9 = 4+5 = N(2+\sqrt{-5}) = N(rt) = N(r)N(t).$$

The only possibilities for $N(r)$ are $1, 3$, and 9. If $N(r) = 1$, then, as we saw above, r must be a unit. Likewise, if $N(r) = 9$, then $N(t) = 1$ and t is a unit. So it only remains to consider the case that $N(r) = 3$. Pick integers a and b so that $r = a+b\sqrt{-5}$. Then $3 = N(r) = N(a+b\sqrt{-5}) = a^2+5b^2$. The only possibility for b is 0, since otherwise a^2+5b^2 must be at least 5. But then $3 = a^2$. Since there is no integer a whose square is 3, we reject the alternative that $N(r) = 3$.

In this way, we see that 9 has two quite distinct factorizations into irreducibles in $\mathbb{Z}[\sqrt{-5}]$. So $\mathbb{Z}[\sqrt{-5}]$ is not a unique factorization domain. □

The Fundamental Factorization Theorem for Principal Ideal Domains. *Every principal ideal domain is a unique factorization domain.*

Proof. We just need to demonstrate that every principal ideal domain has both the primeness condition and the divisor chain condition.

Let **D** be a principal ideal domain and suppose that $a \in D$ is irreducible and that $a \mid bc$ where $b, c \in D$. We must argue that either $a \mid b$ or $a \mid c$. So let us reject the first alternative: we assume $a \nmid b$. Let $M = (a)$. My contention is that M is maximal among proper ideals. Certainly, $M \neq D$ since a is not a unit. So M is a proper ideal. Suppose that I is an ideal that includes M. Since I is a principal ideal pick d to be a generator of of I. Now $a \in M \subseteq I = (d)$. So a is a multiple of d. That is, $a = dw$ for some w. Since a is irreducible, either d is a unit, in which case $I = D$, or w is a unit, in which case $M = I$. In this way, we see that M is maximal. Since we have $a \nmid b$ we see that $b \notin M$. So the ideal (a, b) generated by a and b must be all of D. This means $1 \in (a, b)$. So pick $x, y \in D$ so that

$$1 = xa + yb.$$

This yields $c = xac + ybc$. But $bc \in M = (a)$ since $a \mid bc$ and $xac \in (a)$ as well. So $c \in (a)$, since (a) is an ideal. This means that $a \mid c$ and so the primeness condition holds.

Consider the divisor chain condition. Suppose that $\cdots \mid a_2 \mid a_1 \mid a_0 = a$ is a proper divisor chain in **D**. Then

$$(a_0) \subsetneqq (a_1) \subsetneqq (a_2) \subsetneqq \dots$$

is a properly increasing chain of ideals in **D**. Let I be the union of this chain. We know that the union of any chain of ideals is again an ideal. So I is an ideal. Let d be a generator I. Pick a natural number k so that $d \in (a_k)$. Then $I = (d) \subseteq (a_k) \subseteq I$. Thus, $I = (a_k)$ and the chain of ideals displayed above must be finite. This means our original divisor chain must also be finite, proving the divisor chain condition. □

So we have an immediate corollary.

The Fundamental Theorem of Arithmetic. *The ring of integers is a unique factorization domain.*

Actually, the line of reasoning we have just described is a kind of reorganization of the reasoning in Gauss's Disquisitiones.

We can extract from our proof of the Fundamental Factorization Theorem for Principal Ideal Domains the following result:

> In a principal ideal domain every prime ideal is maximal among all the proper ideals.

To see it, let P be a prime ideal of the principal ideal domain $\mathbf{D}$ and pick a so that $P = (a)$. Observe that $b \in P$ if and only if $a \mid b$ for all $b \in D$. This allows us to conclude that a is prime. By a contention mentioned in the proof of the Fundamental Factorization Theorem, we see that P is a maximal ideal.

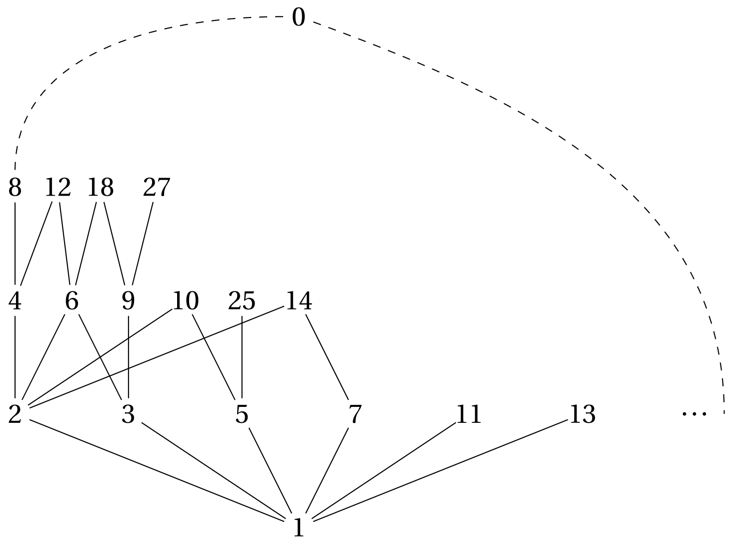
The converse, that every maximal proper ideal is prime, holds in every commutative ring $\mathbf{R}$. For suppose M is a maximal proper ideal and $ab \in M$. In case $b \notin M$ we have $(b, M) = R$. So we can pick $u \in R$ and $w \in M$ so that $1 = ub + w$. It follows that $a = uab + aw$. Since both $ab \in M$ and $w \in M$ we conclude that $a \in M$, as desired.

Of course, this more general business is only interesting if there are significant examples of principal ideal domains other than the ring of integers. There are and we will meet some others soon.

3.3 Divisibility

We have already used divisibility above. Given a commutative ring $\mathbf{R}$ we say that an element a **divides** an element b provided there is an element c so that $ac = b$. We denote this relation by $a \mid b$. Observe that $\mid$ is a two-place relation on R. Moreover, all the graduate students will see easily that this relation is both reflexive and transitive. It just misses being an order relation because it fails the antisymmetry property—that is, $a \mid b$ and $b \mid a$ may hold even though $a \neq b$. For example, $1 \mid -1$ and $-1 \mid 1$ in the ring of integers, but $-1 \neq 1$. Suppose $\mathbf{R}$ is an integral domain and $a \mid b$ and $b \mid a$. Pick elements u and v so that $au = b$ and $bv = a$. Then we have $a(uv) = a$. This means that either $a = 0$ or $uv = 1$. In the first alternative we find that $b = 0$ as well, so that $a = 0 = b$, while in the second alternative we see that $a \sim b$. So in either case a and b are associates. The relation of association is an equivalence relation on R and up to this equivalence relation, the divisibility relation is an ordering.

Let us suppose that $\mathbf{R}$ is an integral domain and consider the divisibility ordering $\mid$ on (the $\sim$ classes of) R. The element 0 is the largest element in this ordering since $a \mid 0$ for all a (because $a \cdot 0 = 0$). Likewise the element 1 is the least element of this ordering (well, actually we are ordering the $\sim$ classes and we really mean the set of units is the least thing...). The figure below sketches part of the divisibility ordering on the natural numbers (these are the representatives of the $\sim$-classes of the integers by taking the nonnegative member of each class).



A Finite Fragment of the Divisibility Relation on the Natural Numbers

The set of natural numbers ordered by divisibility has some properties that may be discerned from this diagram (or perhaps more easily if more of the diagram were to be filled in...). As noted, it has a least element and a greatest element. Also the elements are evidently organized into levels, depending on the number of factors occurring in their decomposition into primes. So $8, 12, 18$, and 27 belong on the same level since they each have 3 factors in their decompositions:

$$\begin{aligned} 8 &= 2 \cdot 2 \cdot 2 \\ 12 &= 2 \cdot 2 \cdot 3 \\ 18 &= 2 \cdot 3 \cdot 3 \\ 27 &= 3 \cdot 3 \cdot 3 \end{aligned}$$

In this way, 1 is the only element at level 0, which suggests we might think it has 0 factors in its decomposition into primes. The primes themselves occupy level 1, and so on. In addition to the top element 0, there will be countably many levels—one level for each natural number—and each level, apart from level 0, is itself countably infinite (an extension of a famous result of Euclid: the graduate students are invited to prove this extension).

Another thing to notice is that any two elements, for example 6 and 9, have a greatest lower bound (in this case 3) and a least upper bound (in this case 18). Ordered sets in which every pair of distinct elements has both a least upper bound and a greatest lower bound are called **lattice-ordered sets**. It is important to realize that the words “greatest” and “least” refer to divisibility and *not* to that other familiar order $\leq$. In rings, we issue special names. Given elements a and b of a commutative ring we say that an element d is a **greatest common divisor** of a and b provided

- $d \mid a$ and $d \mid b$, and
- if $e \mid a$ and $e \mid b$, then $e \mid d$.

You should notice that greatest common divisors are not unique—both 3 and -3 are greatest common divisors of 6 and 9. However, in any integral domain, any two greatest common divisors of a and b must be associates. Likewise, we say an element ℓ is a **least common multiple** of a and b provided

- $a \mid \ell$ and $b \mid \ell$, and
- if $a \mid m$ and $b \mid m$,then $\ell \mid m$.

Like greatest common divisors, least common multiples need not be unique. In integral domains, they are unique up to association. We say that the elements a and b are **relatively prime** provided 1 is a greatest common divisor of a and b.

While in the ring of integers, it is easy to see that greatest common divisors and least common multiple always exist, this is less obvious for other rings. After some reflection, you can convince yourselves that the existence of greatest common divisors and least common multiples can be established with the help of the Fundamental Theorem of Arithmetic. Only a bit more reflection leads us to be conclusion that greatest common divisors and least common multiples always exist in unique factorization domains.

It takes a bit more work (but what else should the graduate students be doing?) to establish the following fact.

Fact. Let **D** be an integral domain. If any two elements of D have a greatest common divisor, then **D** has the primeness condition.

This means that in the Theorem Characterizing Unique Factorization Domains we can replace the primeness condition with the condition that any pair of elements have a greatest common divisor.

3.4 The Chinese Remainder Theorem

The Chinese Remainder Theorem, the focus of this section, appeared in its earliest known form in China in the 3rd century C.E. and, after various refinements, it has taken its place among the most widely known theorems of number theory. It actually holds in all commutative rings and even in some much broader contexts. In its most familiar form, it deals with the simultaneous solution of certain congruences with respect to pairwise relatively prime moduli. To frame this for commutative rings in general we will replace the integer moduli by ideals. Suppose that a and b are relatively prime integers. Observe that the ideal $(a)+(b)$ must have a generator d since the ring of integers is a principal ideal domain. Thus $(a)+(b)=(d)$. Because $(a)\subseteq(d)$ we see that $d \mid a$. Likewise, $d \mid b$. So d is a common divisor of a and b. It must be a greatest common divisor of a and b since (d) is the least ideal that contains both (a) and (b). But recall that a and b are relatively prime. Hence $(d)=(1)$. So we can draw two conclusions:

$$1 = au+bv \quad \text{for some integers } u \text{ and } v$$

and that $(a)+(b)=\mathbb{Z}$. Actually, either of these conclusions imply that a and b are relatively prime.

The Chinese Remainder Theorem. *Let* $\mathbf{R}$ *be a commutative ring and let* $I_0, I_1, \dots, I_{n-1}$ *be finitely many ideals of* $\mathbf{R}$ *such that*

$$I_j + I_k = R \quad \textit{for all } j,k<n \textit{ with } j\neq k.$$

Let $a_0, a_1, \dots, a_{n-1} \in R$. *There is some* $b \in R$ *such that*

$$\begin{aligned} b &\equiv a_0 \mod I_0\\ b &\equiv a_1 \mod I_1\\ &\vdots\\ b &\equiv a_{n-1} \mod I_{n-1}. \end{aligned}$$

Proof. The first interesting case happens when $n=2$. Let us examine it. Since $I_0+I_1=R$ we can pick $r_0\in I_0$ and $r_1\in I_1$ so that $1=r_0+r_1$. Put $b=r_0a_1+r_1a_0$. Then observe

$$\begin{aligned} b = r_0a_1+r_1a_0 &\equiv 0\cdot a_1 + 1\cdot a_0 \mod I_0\\ &\equiv a_0 \mod I_0\\ b = r_0a_1+r_1a_0 &\equiv 1\cdot a_1 + 0\cdot a_0 \mod I_1\\ &\equiv a_1 \mod I_1 \end{aligned}$$

So the stipulations of the theorem are strong enough to assert that each *pair* of the listed congruences can be satisfied by some appropriately chosen element b.

Now for each j with $0<j<n$, we have that $I_0+I_j=R$. So pick $s_j\in I_0$ and $t_j\in I_j$ so that $1=s_j+t_j$. Then we obtain

$$1=(s_1+t_1)(s_2+t_2)\dots(s_{n-1}+t_{n-1}).$$

Using the laws of commutative rings and the properties of ideals we can expand this to obtain

$$1 = s + t_1t_2\dots t_{n-1}$$

where $s\in I_0$. Notice that $t_1t_2\dots t_{n-1}\in \bigcap_{0<j<n} I_j$. This means that

$$I_0 + \bigcap_{0<j<n} I_j = R.$$

As we observed above, there is an element $d_0 \in R$ so that

$$\begin{aligned} d_0 &\equiv 1 \mod I_0 \\ d_0 &\equiv 0 \mod \bigcap_{0<j<n} I_j. \end{aligned}$$

Since $\bigcap_{0<j<n} I_j \subseteq I_k$ for all k with $0 < k < n$, we find that

$$\begin{aligned} d_0 &\equiv 1 \mod I_0 \\ d_0 &\equiv 0 \mod I_j \text{ for all } j \text{ with } 0 \neq j < n. \end{aligned}$$

We can apply this reasoning that worked for the index 0 to any of the indices. In this way, for each $k < n$ we can have $d_k \in R$ such that

$$\begin{aligned} d_k &\equiv 1 \mod I_k \\ d_k &\equiv 0 \mod I_j \text{ for all } j \text{ with } k \neq j < n. \end{aligned}$$

Now put $b = \sum_{k<n} d_k a_k$. Then for all $j < n$ we obtain

$$\begin{aligned} b &= d_0 a_0 + \cdots + d_{j-1} a_{j-1} + d_j a_j + d_{j+1} a_{j+1} + \cdots + d_{n-1} a_{n-1} \\ b &\equiv 0 \cdot a_0 + \cdots + 0 \cdot a_{j-1} + 1 \cdot a_j + 0 \cdot a_{j+1} + \cdots + 0 \cdot a_{n-1} \mod I_j \\ b &\equiv a_j \mod I_j \end{aligned}$$

as desired. □

We can cast the Chinese Remainder Theorem as a structure theorem for commutative rings. Recall that in § 1.3 we discussed direct products of algebraic systems in general. For commutative rings we can enhance the fact at the end of that section.

The Chinese Remainder Theorem: Structural Version. *Let* $\mathbf{R}$ *be a commutative ring and* $I_0, I_1, \ldots, I_{n-1}$ *be a finite list of ideals of* $\mathbf{R}$. *Then*

$$\mathbf{R}\Big/\bigcap_{j<n} I_j \text{ is embeddable into } \mathbf{R}/I_0 \times \cdots \times \mathbf{R}/I_{n-1}.$$

Moreover, if $I_j + I_k = R$ *for all* j *and* k *with* $j, k < n$ *and* $j \neq k$, *then the embedding is an isomorphism.*

Proof. We need a map h from $\mathbf{R}$ into the direct product whose kernel is $\bigcap_{j<n} I_j$. Then we can invoke the Homomorphism Theorem to obtain the desired embedding. The map is the one that comes most easily to hand:

$$h(a) := (a + I_0, a + I_1, \ldots, a + I_{n-1}) \text{ for all } a \in R.$$

This map is assembled from the quotient maps. It is routine to demonstrate that it is a homomorphism and that its kernel is $\bigcap_{j<n} I_j$. An appeal to the Homomorphism Theorem gives us the desired embedding. So the first part of this theorem just rests on general considerations. The power resides in the "moreover" part of the statement. For that, what is needed is to see that h maps R onto the direct product.

Consider any element of the direct product. It has the form

$$(a_0 + I_0, a_1 + I_1, \ldots, a_{n-1} + I_{n-1}).$$

We must see that there is a $b \in R$ so that

$$h(b) = (a_0 + I_0, a_1 + I_1, \ldots, a_{n-1} + I_{n-1}).$$

Given the definition of h, we see that this is the same as finding a $b \in R$ so that

$$b + I_j = a_j + I_j \text{ for all } j < n.$$

In other words, that

$$b \equiv a_j \mod I_j \text{ for all } j < n.$$

Of course, this is precisely what the Chinese Remainder Theorem does for us. □

3.5 PROBLEM SET 3

ALGEBRA HOMEWORK, EDITION 3
FOURTH WEEK
MORE IDEALS

PROBLEM 15.
Let **R** be a commutative ring and let n be a positive integer. Let $J, I_0, I_1, \ldots, I_{n-1}$ be ideals of **R** so that I_k is a prime ideal for every $k < n$ and so that $J \subseteq I_0 \cup \cdots \cup I_{n-1}$. Prove that $J \subseteq I_k$ for some $k < n$.

PROBLEM 16.
Let **R** be a nontrivial commutative ring and let J be the intersection of all the maximal proper ideals of **R**. Prove that $1 + a$ is a unit of **R** for all $a \in J$.

PROBLEM 17.
Let **R** be a commutative ring. Define

$$N := \{a \mid a \in R \text{ and } a^n = 0 \text{ for some positive integer } n\}.$$

(a) Prove that N is an ideal of **R**.

(b) Prove that $N \subseteq P$ for every prime ideal P of **R**.

PROBLEM 18.
Let **R** be a commutative ring. An ideal Q of **R** is called **primary** provided, if $ab \in Q$, then $a \in Q$ or $b^n \in Q$ for some natural number n. Prove that if Q is a primary ideal and $Q = I_0 \cap \cdots \cap I_{n-1}$ where each I_j is a prime ideal, the $Q = I_j$ for some $j < n$.

PROBLEM 19.
Let **R** be a unique factorization domain.

(a) Let u be any element of **R**. Prove that there are only finitely many principal ideals that contain u.

(b) Prove that **R** satisfies the ascending chain condition,

Lecture 4

Zorn's Lemma

Zorn's Lemma is a transfinite existence principle which has found a number of useful and informative applications in algebra and analysis. While the Lemma bears the name of Max Zorn, equal credit should be extended to Felix Hausdorff and Kazimierz Kuratowski who found closely related results decades before Zorn published applications of the Lemma in 1935. Indeed, Zorn supplied no proof, rather he said that he would give one in a later paper, which he never published.

A **chain** or **linearly ordered set** is just a partially ordered set in which any two elements are comparable. We also refer to any subset of a partially ordered set as a **chain** when it is linearly ordered by the ordering inherited from the larger ordered set. This means that, where a and b are elements of the chain and $\leq$ denotes the order relation, we have either $a \leq b$ or $b \leq a$. Let C be a subset of a partially ordered set P and $b \in P$. We say that b is an **upper bound** of C provided $a \leq b$ for all $a \in C$. We say b is a **strict upper bound** provided $a < b$ for all $a \in C$. An element d is **maximal** in C if $d \in C$ and whenever $d \leq a \in C$ it follows that $d = a$.

Zorn's Lemma. *Let P be a partially ordered set and suppose that every chain in P has an upper bound in P. Then P has a maximal member.*

Proof. Let g be a function which chooses an element from each nonempty subset of P. That is the domain of g is the collection of nonempty subsets of P and $g(D) \in D$ for each nonempty subset $D \subseteq P$. The function g, which is called a *choice function*, exists according to the Axiom of Choice.

Denote the ordering on P by $\leq$. For each set $C \subseteq P$ let $\widehat{C}$ denote the set of all strict upper bounds of C. Notice that the empty set $\varnothing$ is a chain in P. According to our hypothesis it must have an upper bound in P. Since $\varnothing$ is empty this upper bound must be a proper upper bound. This means $\widehat{\varnothing}$ is nonempty. (Hence, P is nonempty.)

We will say that $K \subseteq P$ is a g**-chain** provided

- K is not empty.
- K is a chain.
- if $C \subseteq K$ and C has a strict upper bound in K, then $g(\widehat{C})$ is a minimal member of $\widehat{C} \cap K$.

Here is a useful fact about how elements in g-chains compare.

Fact. *Let K and J be g-chains so that $a \in K - J$ and $b \in J$. Then $b < a$.*

Proof. Let $C = \{d \mid d \in K \cap J \text{ and } d < a\}$. So C has a strict upper bound in K. Since K is a g-chain, we have $g(\widehat{C})$ is a minimal member of $\widehat{C} \cap K$. Also, $g(\widehat{C}) \leq a$. Now $\widehat{C} \cap J$ must be empty, since otherwise $g(\widehat{C}) \in K \cap J$, entailing that $g(\widehat{C}) \neq a$ since $a \notin J$, putting $g(\widehat{C}) \in C$, which is impossible. So if $b \in J$ then there is $d \in C$ with $b \leq d < a$. Hence, $b < a$. □

Claim. *The union of any nonempty collection of g-chains is a g-chain.*

Proof. Let L be a union of some nonempty family $\mathcal{F}$ of g-chains. We first have to check that L is linearly ordered. So suppose $a, b \in L$. Pick $K, J \in \mathcal{F}$ so that $a \in K$ and $b \in J$. We need to show that a and b are comparable. We might as well consider that $a \notin J$, since if $a \in J$ we see, J being a chain, that a and b are comparable. But the Fact above then tells us that $b < a$, so a and b are comparable. This means that L is a chain.

Of course, L is not empty since it is union of a nonempty collection of nonempty sets. So it remains to verify the last condition in the definition of g-chain. To this end, let $C \subset L$ such that C has a strict upper bound $b \in L$. Pick $J \in \mathcal{F}$ so that $b \in J$. To see that $C \subseteq J$, pick $a \in C$ and, for contradiction, suppose $a \notin J$. Pick $K \in \mathcal{F}$ so that $a \in K$. Now the Fact yields $b < a$. But here we also have $a < b$. So we find $C \subseteq J$. Since J is a g-chain, we have $g(\widehat{C})$ is a minimal member of $\widehat{C} \cap J$. But we need to see that $g(\widehat{C})$ is a minimal member of $\widehat{C} \cap L$. Suppose not. Pick $a' \in \widehat{C} \cap L$ so that $a' < g(\widehat{C})$. To simplify notation, let $g(\widehat{C}) = b'$. So $a' < b'$. Now pick $K' \in \mathcal{F}$ so that $a' \in K'$. Now the Fact above again yields $b' < a'$, which is contrary to $a' < b'$. This verifies for L the last condition in the definition of g-chain. So L is a g-chain, as claimed. □

Now let M be the union of the collection of all g-chains. Were $\widehat{M}$ nonempty we could form $M \cup \{g(\widehat{M})\}$, which would be a chain properly extending M. A routine check of the definition shows that $M \cup \{g(\widehat{M})\}$ would again be a g-chain. This produces $g(\widehat{M}) \in M \cup \{g(\widehat{M})\} \subseteq M$. But we know $g(\widehat{M}) \notin M$. So $\widehat{M}$ must be empty. So M has no strict upper bounds. But by the hypothesis, every chain has an upper bound. So M must have a largest element m. That is $a \leq m$ for all $a \in M$. As there is no strict upper bound of M, there can be no element which is strictly above m. That is m is the maximal element we seek.

Zorn's Lemma is proven. □

Here is Felix Hausdorff's version, published in 1914.

Hausdorff's Maximality Principle. *In a partially ordered set every chain is included in a maximal chain.*

And here is the version of Kazimierz Kuratowski, published in 1922.

Kuratowski's Maximality Principle. *Every collection of sets, ordered by the inclusion relation $\subseteq$ that has the property that every well-ordered chain of sets in the collection has an upper bound in the collection, must have a maximal member.*

Each of these three maximality principles can be proved with the help of any of the others. It will be inviting to the graduate students to work out the proofs.

There are many maximality principles and other kinds of assertions that certain kinds of infinite sets must exist. There is one more worth mentioning here, since it is sometimes easier to apply than Zorn's Lemma.

To state it, we need another notion. A collection $\mathcal{F}$ of sets is said to have **finite character** provided $X \in \mathcal{F}$ if and only if every finite subset of X belongs to $\mathcal{F}$.

The Teichmüller-Tukey Maximality Principle. *Every nonempty collection of sets that has finite character must have a maximal element.*

Oswald Teichmüller published his paper in 1939 and, independently, John Tukey published his in 1940. An immediate example of a collection of finite character is the collection of all linearly independent subsets of a vector space. This leads almost at once to a proof that every vector space has a basis.

With Zorn's Lemma in hand, the graduate students should be able to deduce the Teichmüller-Tukey Lemma.

Lecture 5

Rings like the Rationals

5.1 Fields

In the commutative ring of rational numbers every nonzero element has a multiplicative inverse—every nonzero element is a unit. Other rings you are acquainted with have this property as well. A **field** is a nontrivial commutative ring in which every nonzero element has a multiplicative inverse. Fields evidently satisfy the cancellation law. So every field is an integral domain. Moreover, since every nonzero element is a unit we see that every nontrivial ideal of a field must actually be the whole field. In other words, every field has exactly two ideals: the trivial ideal and the whole field. Both of these ideals are principal ideals, so every field is a principal ideal domain. So every field is also a unique factorization domain, but in itself, this is not too interesting since fields have no nonzero nonunits to factor and there are no irreducible elements.

Theorem 5.1.1. *Let* $\mathbf{R}$ *be a commutative ring and let* I *be an ideal of* $\mathbf{R}$*. Then*

(a) $\mathbf{R}$ *is a field if and only if* $\mathbf{R}$ *has exactly two ideals.*

(b) $\mathbf{R}/I$ *is a field if and only if* I *is maximal among all the proper ideals of* $\mathbf{R}$*.*

Proof. For part (a) we have already observed the implication from left to right. For the converse, suppose $\mathbf{R}$ has exactly two ideals and let $a \in R$ be nonzero. We have to show that a is invertible. The ideal (a) must be the whole of R, so in particular $1 \in (a)$. This means we can (and do) pick $b \in R$ so that $1 = ab$. So b is the desired inverse of a.

Part (b) is an immediate consequence of part (a) and the Correspondence Theorem. □

To simplify the language we will say that I is a **maximal ideal** of the ring $\mathbf{R}$ provided

(a) I is a proper ideal of $\mathbf{R}$, and

(b) Either $I = J$ or $J = R$ whenever J is an ideal of $\mathbf{R}$ with $I \subseteq J$.

So the theorem above asserts, in part, that, for a commutative ring $\mathbf{R}$, we have that $\mathbf{R}/I$ is a field if and only if I is a maximal ideal of $\mathbf{R}$.

The Maximal Ideal Theorem.

(a) *Every proper ideal of a ring is included in a maximal ideal of the ring.*

(b) *Every nontrivial commutative ring has a homomorphic image that is a field.*

Proof. For part (a) let I be a proper ideal of the ring $\mathbf{R}$. Let

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

Any maximal element of $\mathcal{F}$ will be a maximal ideal that includes I. We invoke Zorn's Lemma to see that $\mathcal{F}$ has a maximal member. Indeed, suppose $\mathcal{C}$ is a chain included in $\mathcal{F}$. If $\mathcal{C}$ is empty, then I will be an upper bound of $\mathcal{C}$. So we suppose that $\mathcal{C}$ is not empty. Observe that $\bigcup\mathcal{C}$ is an ideal of $\mathbf{R}$ since it is a union of a chain of ideals. Plainly, $I \subseteq \bigcup\mathcal{C}$. Finally, were $\bigcup\mathcal{C}$ not proper we would have $1 \in \bigcup\mathcal{C}$. But that would mean that $1 \in J$ for some $J \in \mathcal{C}$. However, the members of $\mathcal{C}$ belong to $\mathcal{F}$ so they are proper ideals. So we find that $\bigcup\mathcal{C}$ is a proper ideal of $\mathbf{R}$ that includes I. This means that $\bigcup\mathcal{C}$ belongs to $\mathcal{F}$. That is, every nonempty chain in $\mathcal{F}$ has an upper bound in $\mathcal{F}$. According to Zorn, $\mathcal{F}$ must have maximal members. This establishes part (a).

Part (b) is an immediate consequence of part (a) and the first theorem in this section. □

The Maximal Ideal Theorem was proven by Wolfgang Krull in 1929.

5.2 Fields of Fractions

In addition to the field of rational numbers, you are also acquainted with the field of real numbers, as well as the field of complex numbers. We also have in hand finite fields like $\mathbb{Z}/(p)$, where p is a prime number. This is because we know that (p) is a prime ideal of $\mathbb{Z}$ and we know that in a principal ideal domain prime ideals are maximal. Some of these fields don't seem much like the field of rational numbers. We know there is a close connection between the integers and the rationals. We can build the field of rationals from the ring of integers. An interesting thing is that the same procedure can be applied to any integral domain to produce a closely associated field. Here is how.

Fix an integral domain $\mathbf{D}$ throughout this section. The idea is to enhance D by adjoining all the multiples of the multiplicative inverses of the nonzero elements of D. There is a little wrinkle in this process. When we do this for the integers we have to throw in $\frac{1}{4}$ to ensure that 4 will have a multiplicative inverse and then we have to throw in $\frac{2}{4} = 2 \cdot \frac{1}{4}$. Of course, we have to identify $\frac{2}{4}$ and $\frac{1}{2}$. There is a two-step process to smooth out this wrinkle.

Let $E = \{(a,b) \mid a,b \in D \text{ with } b \neq 0\}$. On E define the binary relation $\asymp$ by

$$(a,b) \asymp (c,d) \text{ if and only if } ad = bc$$

for all $(a,b),(c,d) \in E$. The eager graduate students will write out a proof that $\asymp$ is an equivalence relation on E. As a second step, we name the equivalence classes in a convenient manner. For $a,b \in D$ with $b \neq 0$ we put

$$\frac{a}{b} := \{(c,d) \mid (c,d) \in E \text{ and } (a,b) \asymp (c,d)\}.$$

So we have

$$\frac{a}{b} = \frac{c}{d} \text{ if and only if } ad = bc,$$

for all $a,b,c,d \in D$ with $b \neq 0 \neq d$.

Let $F' = \{\frac{a}{b} \mid a, b \in D \text{ with } b \neq 0\}$. Our plan is to make F' into a field by defining the ring operations in some appropriate manner. Here is how. For all $\frac{a}{b}, \frac{c}{d} \in F$ let

$$\begin{aligned}
\frac{a}{b} + \frac{c}{d} &:= \frac{ad+cb}{bd} \\
\frac{a}{b} \cdot \frac{c}{d} &:= \frac{ac}{bd} \\
-\frac{a}{b} &:= \frac{-a}{b} \\
0^* &:= \frac{0}{1} \\
1^* &:= \frac{1}{1}
\end{aligned}$$

The last two equations define the one and the zero of the ring of fractions. Of course, these definitions are very familiar from the days in school when we learned how to deal with fractions. It is worth noting that the soundness to these definitions depends on the fact that **D** is an integral domain—to ensure that $bd \neq 0$ when $b \neq 0 \neq d$. Here is the question:

> Is the algebra $\langle F', +, \cdot, -, 0^*, 1^* \rangle$ really a field?

Unfortunately, we seem to be forced to check all the equations defining commutative rings as well as checking that every nonzero element has a multiplicative inverse. This checking is tedious but must be done (by the graduate students!). The most strenuous case is checking the associative law for addition. Here is a verification of a distributive law to show how it is done.

$$\begin{aligned}
\frac{a}{b}\Big(\frac{c}{d} + \frac{e}{f}\Big) &= \frac{a}{b}\frac{cf+ed}{df} \\
&= \frac{a(cf+ed)}{b(df)} \\
&= \frac{a(cf)+a(ed)}{b(df)} \\
&= \frac{((ac)f+(ae)d)\cdot 1}{((bd)f)\cdot 1} \\
&= \frac{(ac)f+(ae)d}{(bd)f}\frac{1}{1} \\
&= \frac{(ac)f+(ae)d}{(bd)f}\frac{b}{b} \\
&= \frac{(ac)(bf)+(ae)(bd)}{(bd)(bf)} \\
&= \frac{ac}{bd} + \frac{ae}{bf} \\
&= \frac{a}{b}\frac{c}{d} + \frac{a}{b}\frac{e}{f}
\end{aligned}$$

In the reasoning above, we used $\frac{1}{1} = \frac{b}{b}$ where we know $b \neq 0$. This is a little lemma that is helpful in the other parts of the proof.

Let $D' = \{\frac{a}{1} \mid a \in D\}$. It is easy to check that D' is a subuniverse of the field $\langle F', +, \cdot, -, 0^*, 1^* \rangle$ and that the map sending $a \mapsto \frac{a}{1}$ for $a \in D$ is an embedding of **D** into the field. But we would rather regard **D** as a subring of its field of fractions, just as we regard $\mathbb{Z}$ as a subring of $\mathbb{Q}$. We accomplish this by letting

$$F := D \cup (F' \setminus \{\frac{a}{1} \mod a \in D\}).$$

We have to define the operations on F. Here is how to define addition for $u, v \in F$.

$$u + v := \begin{cases} u + v & \text{if } u, v \in D \\ u + v & \text{if } u, v \in F' \\ \frac{ub+a}{b} & \text{if } u \in D \text{ and } v = \frac{a}{b} \in F' \\ \frac{a+vb}{b} & \text{if } v \in D \text{ and } u = \frac{a}{b} \in F' \end{cases}$$

The first two lines of this may look a bit strange. The + in the first case refers to the addition in **D**, whereas on the second line the + refers to the addition defined above over F'. The other operations can be defined in a similar fashion. In effect, what we have done is a bit of transplant surgery. We have sliced out D' and put D in its place making sure to stitch things up so the operations work right. The result is a field **F** that has **D** as a subring. This field **F** is called the **field of fractions** of the integral domain **D**.

We have provided one construction that starts with an integral domain **D** and ends up with an extension **F** that can be rightfully called a "field of fractions". However, it should be clear that this construction is not really unique—it is possible to make small changes that will produce other fields that could also be called fields of fractions but that are technically different from the one we have just constructed. There is, however, a strong uniqueness result for fields of fractions.

Theorem on the Existence and Uniqueness of Fields of Fractions. *Let* **D** *be any integral domain. There is a field* **F** *such that* **D** *is a subring of* **F**, *and moreover, if* **S** *is any ring and* **K** *is any field so that* **S** *is a subring of* **K** *and if* $h : \mathbf{D} \to \mathbf{S}$ *is any isomorphism from* **D** *onto* **S**, *then* h *has a unique extension to an embedding of* **F** *into* **K**.

Proof. We already established the existence of a field **F** of fractions. Suppose the field **K** and the embedding h are given to us. We define the extension $\hat{h}$ from F into K as follows. For any $u \in F$ let

$$\hat{h}(u) = \begin{cases} h(u) & \text{if } u \in D \\ h(a)(h(b))^{-1} & \text{if } u = \frac{a}{b} \notin D \end{cases}$$

In the second alternative, $h(b)$ will be a nonzero element of K and it will have a multiplicative inverse in K, which we have denoted by $(h(b))^{-1}$. It is a routine work for the delight of the graduate students to demonstrate that $\hat{h}$ is actually an embedding. Staring hard at the definition of $\hat{h}$ should suggest a proof that this is the only way to get such an extension. □

So the field of fractions of an integral domain **D** is, in the sense described above, the smallest field extending **D**.

5.3 Problem Set 4

Algebra Homework, Edition 4

Fifth Week

Fields

Problem 20.
Use Zorn's Lemma (or one of the other maximality principles) to give a clean proof of the König Infinity Lemma.

Problem 21.
Let **F** be a field and let $p(x) \in \mathbf{F}[x]$ be a polynomial of degree n. Prove that $p(x)$ has at most n distinct roots in **F**.

Problem 22.
Let **R** be a commutative ring and let $a \in R$ with $a^n \neq 0$ for every natural number n. Prove that **R** has an ideal P such that each of the following properties holds:

(a) $a^n \notin P$ for every natural number n, and

(b) for all ideals I of **R**, if $P \subseteq I$ and $P \neq I$, then $a^n \in I$ for some natural number n.

Problem 23.
Let **F** be a field and let $\mathbf{F}^*$ be its (multiplicative) group of nonzero elements. Let **G** be any finite subgroup of $\mathbf{F}^*$. Prove that **G** must be cyclic.

Problem 24.
Suppose that **D** is a commutative ring such that $\mathbf{D}[x]$ is a principal ideal domain. Prove that **D** is a field.

Lecture 6

Rings of Polynomials

6.1 Polynomials over a Ring

$5x^3 + 3x^2 - 7x + 1$ is a polynomial with integer coefficients. Our experience in school and even through calculus leads us to think of polynomials as functions, but here in algebra we take a different view. We consider that polynomials are formal expressions that describe functions. We regard polynomials as certain kinds of strings of symbols. We could also regard the polynomial at the start of this paragraph as a polynomial over the ring $\mathbb{Z}/(8)$. That ring has just 8 elements and there are only 8^8 one-place operations on the underlying set $\{0,1,2,3,4,5,6,7\}$. However, there is a countable infinity of polynomials, some of each degree, with coefficients in that ring. This means that some (actually many) polynomials will name the same function.

The interesting thing about treating polynomials as strings of symbols is that we can define an addition and a multiplication, as well as the formation of negatives and in this way produce a ring. We know well how to add and multiply polynomials in a formal manner—having had lots of drill in Algebra I. To help in formalizing addition and multiplication, it is convenient to write polynomials backwards from how most of us were taught. In fact, it is reasonable to imagine each polynomial as an infinitely long expression where after some point all the coefficients are 0 (and so have been neglected...).

Here is how addition works, of course.

$$\begin{array}{ccccccccc}
a_0 & + & a_1 x & + & a_2 x^2 & + \cdots + & a_n x^n \\
b_0 & + & b_1 x & + & b_2 x^2 & + \cdots + & b_n x^n \\
\hline
(a_0 + b_0) & + & (a_1 + b_1)x & + & (a_2 + b_2)x^2 & + \cdots + & (a_n + b_n)x^n
\end{array}$$

Notice that while this looks like we have assumed that the polynomials are both of degree n, we have not made such an assumption. Some (or all) of the coefficients above can be 0. So this description of addition works for all polynomials. It is important to realize that the $+$'s occurring in the parentheses on the last line actually refer to the addition in the ring of coefficients. So the idea is that, unlike the other $+$'s, which are formal symbols, those in the parentheses should actually be executed to produce elements of the ring of coefficients to get the coefficients of the sum of the polynomials.

Multiplication is more complicated.

$$\begin{array}{ccccccccc}
a_0 & + & a_1x & + & a_2x^2 & +\cdots+ & a_nx^n \\
b_0 & + & b_1x & + & b_2x^2 & +\cdots+ & b_nx^n \\
\hline
(a_0b_0) & + & (a_0b_1+a_1b_0)x & + & (a_0b_2+a_1b_1+a_2b_0)x^2 & +\cdots+ & (\sum_{i+j=n} a_ib_j)x^n
\end{array}$$

In general, the k^{th} coefficient is

$$\sum_{i+j=k} a_ib_j.$$

Here is a smaller example

$$\begin{aligned}
(a_0+a_1x)(b_0+b_1x+b_2x^2) &= a_0(b_0+b_1x+b_2x^2)+a_1x(b_0+b_1x+b_2x^2)\\
&= a_0b_0+a_0b_1x+a_0b_2x^2+a_1b_0x+a_1b_1x^2+a_1b_2x^3\\
&= a_0b_0+(a_0b_1+a_1b_0)x+(a_0b_2+a_1b_1)x^2+a_1b_2x^3
\end{aligned}$$

This looks like a deduction—like a proof the the formula at the start is equal to the last formula. It looks like a string of uses of the distributive, associative, and commutative laws. But it is not really a deduction. We would first have to see that those laws actually hold. Rather, the display above is the basis for the definition of multiplication given above. But this does show that while we didn't allow the commutative law to sneak into the calculation of the coefficients, we have somehow assumed here that the variable x commutes with everything.

Given a ring $\mathbf{R}$, we make the **ring $\mathbf{R}[x]$ of polynomials with coefficients from $\mathbf{R}$** by imposing the addition and multiplication described above on the set of polynomials. The zero of the ring of polynomials is the polynomials where all the coefficients are 0. The one of this ring is the polynomial with constant coefficient 1 and all other coefficients 0. Forming negatives of polynomials we leave to the imagination of the graduate students.

Well, is $\mathbf{R}[x]$ really a ring? We need to check the equations that we used to define the notion of a ring. The equations only involving $+, -$ and 0 are easy. The associative law for multiplication and the distributive laws are messy and best not displayed in public (but the disciplined graduate students will not flinch from checking this stuff). Notice that $\mathbf{R}$ is a subring of $\mathbf{R}[x]$.

The **zero polynomial** is the one whose coefficients are all 0. Every nonzero polynomial

$$a_0+a_1+\cdots+a_nx^n$$

has a rightmost coefficient that is not 0. This coefficient is the **leading coefficient** of the polynomial and the exponent of the associated x is called the **degree** of the polynomial. It is convenient to assign no degree to the zero polynomial.

If the sum of two polynomials is not the zero polynomial then the degree of the sum can be no larger than the maximum of the degree of the summands. Likewise, if the product of two polynomials is not the zero polynomial, then the degree of the product is no larger than the sum of the degrees of two factors. If $\mathbf{R}$ is an integral domain, then the degree of the product of nonzero polynomials in $\mathbf{R}[z]$ is the sum of the degrees of the factors.

Once we are convinced that $\mathbf{R}[x]$ is a ring we can repeat the construction to form the ring $\mathbf{R}[x][y]$. Here is a member of $\mathbb{Z}[x][y]$.

$$(1+2x+3x^3)+(2-x)y+(5+x^3)y^2$$

Observe that the coefficients of this polynomial (namely, the parts in parentheses) are members of $\mathbf{R}[x]$. We identify this polynomial with

$$1+2x+2y-xy+5y^2+3x^3+x^3y^2$$

Now notice that the polynomial below

$$(1+2y+5y^2)+(2-y)x+(3+y^2)x^3$$

is a member of $\mathbb{Z}[y][x]$ that we also identify with

$$1+2x+2y-xy+5y^2+3x^3+x^3y^2$$

By similar reasoning we identify $\mathbb{Z}[x][y]$ with $\mathbb{Z}[y][x]$. We use the notation $\mathbb{Z}[x,y]$ to denote this ring. More generally, we arrive at the polynomial ring $\mathbf{R}[x_0,x_1,\dots,x_n]$ in any finite number of variables. It is even possible to consider rings of polynomials over infinite sets of variables, although we will not pursue this.

Here are some easily deduced facts.

Fact. Let **R** be a ring. $\mathbf{R}[x]$ is a commutative ring if and only if **R** is a commutative ring.

Fact. Let **R** be a ring. $\mathbf{R}[x]$ is an integral domain if and only if **R** is an integral domain.

A very useful result about rings of polynomials is next.

The Homomorphism Extension Property for $\mathbf{R}[x]$. *Let* $\mathbf{R},\mathbf{S}$, *and* **T** *be rings so that* **S** *is a subring of* **T** *and let h be a homomorphism from* **R** *onto* **S**. *For any $t\in T$ there is exactly one homomorphism $\hat{h}$ extending h that maps* $\mathbf{R}[x]$ *into* **T** *such that* $\hat{h}(x)=t$.

This theorem is illustrated in Figure 6.1.

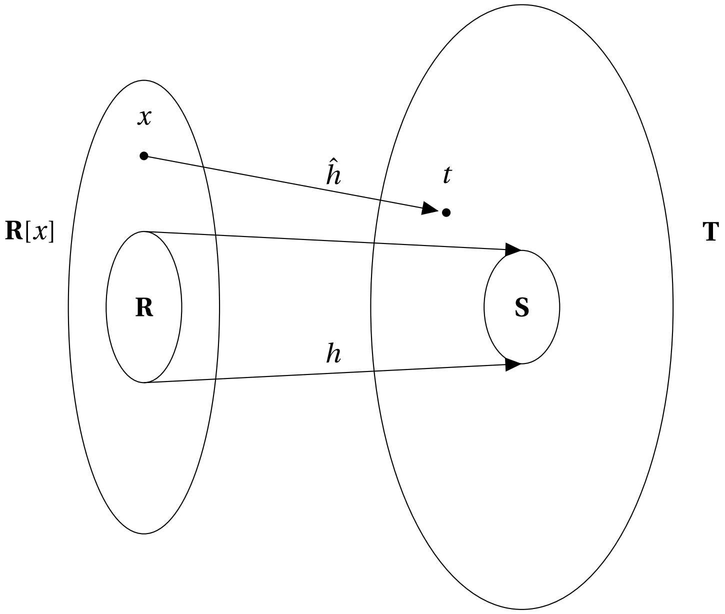


Figure 6.1: The Homomorphism Extension Property

Proof. Consider an arbitrary polynomial

$$p(x)=a_0+a_1x+\cdots+a_nx^n.$$

Were there to be any extension $\hat{h}$ of h as desired, then we would have to have

$$\begin{aligned}\hat{h}(p(x)) &= \hat{h}(a_0)+\hat{h}(a_1\hat{h}(x)+\cdots+\hat{h}(a_n)(\hat{h}(x))^n\\ &= h(a_0)+h(a_1)t+\cdots+h(a_n)t^n.\end{aligned}$$

In this way we see that there can be at most one possibility for $\hat{h}$. Moreover, we can define the desired extension by

$$\hat{h}(p(x)) := h(a_0) + h(a_1)t + \cdots + h(a_n)t^n.$$

The only issue is whether this function is actually a homomorphism. This we leave in the capable hands of the graduate students. □

An interesting special case of this theorem is when $\mathbf{R} = \mathbf{S}$ and h is just the identity map. In that case the extension $\hat{h}$ gives us

$$\hat{h}(p(x)) = \hat{h}(a_0 + a_1 x + \cdots + a_n x^n) = a_0 + a_1 t + \cdots + a_n t^n = p(t).$$

Notice that the $p(x)$ on this line is a polynomial whereas the $p(t)$ is an element of T. If we construe the polynomial $p(x)$ as a name for a function from T to T, then what $\hat{h}$ does is *evaluate* the named function at the input t. For this reason, $\hat{h}$, which depends on t, is called an *evaluation map*. In this context, we say that t is a **root** of $p(x)$ provided $\hat{h}(p(x)) = 0$; that is provided $p(t) = 0$ in $\mathbf{T}$.

We saw a key fact about the integers that had to do with quotients and remainders. This very useful fact led us to the conclusion that the ring of integers is a principal ideal domain. Something like this fact holds for polynomial rings.

Theorem on Quotients and Remainders for Polynomials. *Let* $\mathbf{R}$ *be a commutative ring, let* $d(x) \in \mathbf{R}[x]$ *be a nonzero polynomial, and let* b *be the leading coefficient of* $d(x)$. *Let* $f(x) \in \mathbf{R}[x]$ *be any polynomial. There is a natural number* k *and there are polynomials* $q(x)$ *and* $r(x)$ *such that*

(a) $b^k f(x) = q(x)d(x) + r(x)$ *and*

(b) *either* $r(x)$ *is the zero polynomial or* $\deg r(x) < \deg d(x)$.

Moreover, given such a k *the polynomials* $q(x)$ *and* $r(x)$ *are unique, provided* $\mathbf{R}$ *is an integral domain.*

Proof. Observe that if the degree of $d(x)$ is larger than the degree of $f(x)$, then we can take $r(x) = f(x)$ and $q(x) = 0$ and we can put $k = 0$. So the existence part of this theorem only needs a proof when $\deg f(x) \geq \deg d(x)$. We prove the existence part of the theorem by induction on $\deg f(x)$.

Base Step: $\deg f(x) = \deg d(x)$
Let a be the leading coefficient of $f(x)$. Put $k = 1, q(x) = a$, and $r(x) = bf(x) - ad(x)$. This works.

Inductive Step
We suppose that $\deg f(x) = n+1 > \deg d(x)$. Let m be the degree of $d(x)$. Once more let a be the leading coefficient of $f(x)$. Observe

$$\hat{f}(x) := bf(x) - ax^{n+1-m}xd(x)$$

is a polynomial of degree no more than n. We can apply the induction hypothesis to obtain a natural number ℓ and polynomials $\hat{q}(x)$ and $r(x)$ so that

(a) $b^\ell \hat{f}(x) = \hat{q}(x)d(x) + \hat{r}(x)$ and

(b) $r(x)$ is the zero polynomial or $\deg r(x) < \deg d(x)$.

But this entails

$$\begin{aligned} b^{\ell+1} f(x) &= \hat{q}(x)d(x) + ax^{n+1-m}d(x) + r(x) \\ &= (\hat{q}(x) + ax^{n+1-m})d(x) + r(x). \end{aligned}$$

Taking $q(x) := \hat{q}(x) = ax^{n+1-m}$ establishes the inductive step.

So the existence part of the theorem is finished. For the uniqueness part, we suppose that $\mathbf{R}$ is an integral domain and we take k to be a fixed natural number and $f(x)$, $q_0(x)$, $q_1(x)$, $r_0(x)$, and $r_1(x)$ to be polynomials such that

(a) $b^k f(x) = q_0(x)d(x) + r_0(x)$,

(b) $r_0(x)$ is the zero polynomial or $\deg r_0(x) < \deg d(x)$,

(c) $b^k f(x) = q_1(x) + r_1(x)$, and

(d) $r_1(x)$ is the zero polynomial or $\deg r_1(x) < \deg d(x)$.

It follows that

$$(q_0(x) - q_1(x))d(x) = r_1(x) - r_0(x).$$

Now the polynomial on the right is either the zero polynomial or it has degree less than the degree of $d(x)$. The polynomial on the left is either the zero polynomial or it has degree at least the degree of $d(x)$. It follows that both sides of this equation are the zero polynomial. In particular, $r_0(x) = r_1(x)$. (At this point we have yet to invoke the fact that **R** is an integral domain.) So we have

$$(q_0(x) - q_1(x))d(x) = 0.$$

We know that $d(x)$ is not the zero polynomial. Since $\mathbf{R}[x]$ is an integral domain, we find that $q_0(x) = q_1(x)$, as desired. □

Here are three important immediate corollaries of this theorem.

Corollary 6.1.1. *Let* **R** *be a commutative ring and let* $d(x) \in \mathbf{R}[x]$ *be a nonzero polynomial whose leading coefficient is a unit. Let* $f(x) \in \mathbf{R}[x]$ *be any polynomial. There are polynomials* $q(x)$ *and* $r(x)$ *such that*

(a) $f(x) = q(x)d(x) + r(x)$ *and*

(b) *either* $r(x)$ *is the zero polynomial or* $\deg r(x) < \deg d(x)$.

Moreover, the polynomials $q(x)$ *and* $r(x)$ *are unique, provided* **R** *is an integral domain.*

Corollary 6.1.2. $\mathbf{F}[x]$ *is a principal ideal domain provided* **F** *is a field. Hence* $\mathbf{F}[x]$ *is a unique factorization domain, provided* **F** *is a field.*

Corollary 6.1.3. *Let* **R** *be a commutative ring, let* $f(x) \in \mathbf{R}[x]$ *be a polynomial with coefficients in* R *and let* $r \in R$*. Then* r *is a root of* $f(x)$ *if and only if* $(x - r) \mid f(x)$.

The second of the corollaries displayed above can be deduced in the same manner that we used to establish that $\mathbb{Z}$ is a principal ideal domain.

There is one more general observation to make here.

The Binomial Theorem holds in every commutative ring.

This means that in any commutative ring we have

$$(x + y)^n = \sum_{k \le n} \binom{n}{k} x^k y^{n-k} \text{ for all elements } x \text{ and } y \text{ of the ring.}$$

This must be understood carefully. The binomial coefficient $\binom{n}{k}$ that appear here are positive natural numbers, not elements of the ring at hand. We must understand them as indicating repeated additions within the ring. That is we take $\binom{n}{k}$ to be

$$\underbrace{1 + \cdots + 1}_{\binom{n}{k} \text{ times}}.$$

With this in mind, it is routine to see that only the laws of commutative rings are needed to establish the Binomial Theorem. Now notice that $n \mid \binom{n}{k}$ for all k such that $0 < k < n$, while $\binom{n}{0} = 1 = \binom{n}{n}$. This observation yields

Fact. Let **R** be a commutative ring of characteristic n. Then for all $x, y \in R$

$$(x+y)^n = x^n + y^n.$$

Moreover, the map sending $a \mapsto a^n$ for all $a \in R$ is a homomorphism.

This map, used in later parts of algebra, is called the **Frobenius map**.

Finally, we know that $\mathbb{Z}$ is an integral domain and so $\mathbb{Z}[x]$ is also an integral domain. However, even though $\mathbb{Z}$ is a principal ideal domain, it turns out that $\mathbb{Z}[x]$ is *not* a principal ideal domain. Establishing this fact is a task left to the graduate students in one of the Problem Sets. Even though $\mathbb{Z}[x]$ is not a principal ideal domain, it turns out that it is still a unique factorization domain and that while some of the ideals of $\mathbb{Z}[x]$ cannot be generated by some single element, it is nevertheless true that all the ideals of $\mathbf{Z}[x]$ can be generated by some *finite set* of elements. These are consequences of more general theorems (of Gauss and Hilbert) that are the primary objectives of this sequence of lectures.

6.2 Polynomials over a Unique Factorization Domain

We know that both $\mathbb{Z}$ and $\mathbf{F}[x]$, where **F** is a field, are principal ideal domains, but that neither $\mathbb{Z}[x]$ nor $\mathbf{F}[x, y]$ are principal ideal domains. It is a theorem of Gauss that $\mathbb{Z}[x]$ is nevertheless a unique factorization domain. With little change to the proof of Gauss, we have the following theorem.

Unique Factorization Theorem for Polynomials over a Unique Factorization Domain. *Let* **D** *be a unique factorization domain. Then* $\mathbf{D}[x]$ *is also a unique factorization domain.*

The proof of this theorem depends on three lemmas. Let **F** be the field of fractions of the unique factorization domain **D**. The diagram below may help to understand how the proof will work.

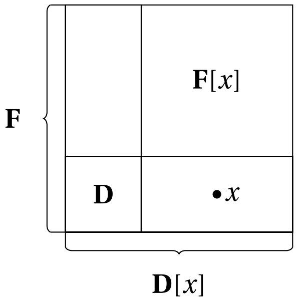


Figure 6.2: Linking **D** and $\mathbf{F}[x]$

The information we start with is that **D** is a unique factorization domain, that **F** is the field of fractions of **D** (and therefore closely linked to **D**), and that $\mathbf{F}[x]$ is also a unique factorization domain. Observe that

$$\mathbf{D} \subseteq \mathbf{D}[x] \subseteq \mathbf{F}[x].$$

When Gauss tackled this problem, he had $\mathbb{Z}$ in place of **D** and $\mathbb{Q}$ in place of **F**.

We will say that a polynomial in $\mathbf{D}[x]$ is **primitive** provided there is no irreducible of **D** that divides all of the coefficients of the polynomial.

Lemma A.
Let **D** *be a unique factorization domain and let* **F** *be its field of fractions. Let* $p(x)$ *be a nonzero polynomial with coefficients in* **F**. *There is an element* $c \in F$ *and a primitive polynomial* $q(x) \in \mathbf{D}[x]$ *such that* $p(x) = cq(x)$. *Moreover, up to multiplication by units of* **D**, *the element* c *and the coefficients of* $q(x)$ *are unique.*

Proof. Pick $a_0, b_0, a_1, b_1, \ldots, a_n, b_n \in D$ so that

$$p(x) = \frac{a_0}{b_0} + \frac{a_1}{b_1}x + \cdots + \frac{a_n}{b_n}x^n.$$

Let $b = b_0 b_1 \cdots b_n$. Then

$$bp(x) = c_0 + c_1 x + \cdots + c_n x^n$$

for certain elements $c_0, c_1, \ldots, c_n \in D$. Let d be a greatest common divisor of $c_0, c_1, \ldots, c_n$. Factoring d out of $c_0 + c_1 x + \cdots + c_n x^n$ leaves a primitive polynomial $q(x)$ such that

$$bp(x) = dp(x).$$

Let $c = \frac{d}{b}$. Then $p(x) = cp(x)$, establishing the existence part of Lemma A. This argument was just clearing the denominators and factoring the remaining coefficients as much as possible.

Now consider the uniqueness assertion. Suppose $c, c^* \in F$ and $q(x), q^*(x) \in D[x]$ with both $q(x)$ and $q^*(x)$ primitive such that

$$cq(x) = p(x) = c^* q^*(x).$$

Pick $r, s, r^*, s^* \in D$ so that $c = \frac{r}{s}$ and $c^* = \frac{r^*}{s^*}$ and so that r and s are relatively prime as are r^* and s^*. So we have

$$s^* r q(x) = s r^* q^*(x).$$

Now let t be any prime in $\mathbf{D}$ so that $t \mid s^*$. We know that t cannot divide r^* and it cannot divide each of the coefficients of $q^*(x)$ since that polynomial is primitive. Therefore it must divide s. So, factoring s^* into primes, we see that $s^* \mid s$. In a like manner, we can conclude that $s \mid s^*$. This means that s and s^* are associates. Pick a unit u of $\mathbf{D}$ so that $s = s^* u$. So we find after cancellation

$$rq(x) = ur^* q^*(x).$$

Applying the same reasoning to r^* and r, we can find a unit v of $\mathbf{D}$ so that $r^* = vr$. This gives

$$q(x) = uvq^*(x) \text{ and } c = \frac{r}{s} = \frac{ru}{su} = \frac{ru}{s^*} = \frac{rvu}{s^* v} = \frac{r^* u}{s^* v} = \frac{r^*}{s^*}\frac{u}{v} = c^* \frac{u}{v}.$$

But both uv and $\frac{u}{v}$ are units of $\mathbf{D}$. This finishes the proof of uniqueness up to multiplication by units of $\mathbf{D}$. □

An immediate consequence of Lemma A is that if $p(x)$ and $q(x)$ are primitive polynomials in $\mathbf{D}[x]$ that are associates in $\mathbf{F}[x]$, then they are already associates in $\mathbf{D}[x]$.

Gauss's Lemma. *Let* $\mathbf{D}$ *be a unique factorization domain. The product of two primitive polynomials in* $\mathbf{D}[x]$ *is again primitive.*

Proof. Let $f(x)$ and $g(x)$ be primitive and put $h(x) = f(x)g(x)$. Suppose, for the sake of contradiction, that t is an irreducible of $\mathbf{D}$ that divides all the coefficients of $h(x)$. So t is prime and the ideal (t) is a prime ideal. This means that $\mathbf{D}/(t)$ is an integral domain. Let η denote the quotient map from $\mathbf{D}$ to $\mathbf{D}/(t)$. Using the Homomorphism Extension Theorem for Polynomials, we know there is a unique homomorphism $\hat{\eta} : \mathbf{D}[x] \to \mathbf{D}/(t)[x]$ so that $\hat{\eta}(x) = x$. What $\hat{\eta}$ does is simply apply η to each of the coefficients of the polynomial given as input.

Now observe in $\mathbf{D}/(t)[x]$ we have

$$\begin{aligned} 0 &= \hat{\eta}(h(x)) \text{ since each coefficient of } h(x) \text{ is divisible by } t. \\ &= \hat{\eta}(f(x)g(x)) \\ &= \hat{\eta}(f(x))\hat{\eta}(g(x)) \end{aligned}$$

But t cannot divide all the coefficients of $f(x)$ nor all the coefficients of $g(x)$, since these polynomials are primitive. So $\hat{\eta}(f(x)) \neq 0 \neq \hat{\eta}(g(x))$ in $\mathbf{D}/(t)[x]$. Since $\mathbf{D}/(t)[x]$ is an integral domain, we have uncovered a contradiction. So $h(x)$ must be primitive. □

The proof just given is certainly in the fashion of the 20th century. Here is a proof that appeals directly to basic principles. It is much more like the reasoning of Gauss.

A more basic proof of Gauss's Lemma. Let

$$f(x) = a_0 + a_1 x + \cdots + a_n x^n$$
$$g(x) = b_0 + b_1 x + \cdots + b_m x^m$$

be primitive polynomials. Put $h(x) = f(x)g(x) = c_0 + c_1 x + \cdots + c_{n+m}x^{n+m}$. We know that

$$c_k = \sum_{i+j=k} a_i b_j.$$

Let t be a prime of $\mathbf{D}$. Pick ℓ as small as possible so that $t \nmid a_\ell$ and pick r as small as possible so that $t \nmid b_r$. We can do this since $f(x)$ and $g(x)$ are primitive. Then

$$c_{\ell+r} = (a_0 b_{\ell+r} + a_1 b_{\ell+r-1} + \cdots + a_{\ell-1} b_{r+1}) + a_\ell b_r + (a_{\ell+1} b_{r-1} + \cdots + a_{\ell+r} b_0).$$

(Be generous in understanding this equation. Depending on the values of ℓ and r some terms in the first and last pieces of the sum may be missing.) Then t divides the first and last pieces of this sum, but not the middle term $a_\ell b_r$. This means that t cannot divide $c_{\ell+r}$. Hence, no prime can divide all the coefficients of $h(x)$. So $h(x)$ must be primitive. □

Lemma B.
Let $\mathbf{D}$ *be a unique factorization domain and let* $\mathbf{F}$ *be its field of fractions. If* $f(x) \in \mathbf{D}[x]$ *is irreducible and of positive degree, then* $f(x)$ *is also irreducible in* $\mathbf{F}[x]$

Proof. Observe that $f(x)$ must be primitive since it is of positive degree and irreducible in $\mathbf{D}[x]$. Now suppose that $f(x) = g(x)h(x)$ for some polynomials $g(x), h(x) \in F[x]$. According to Lemma A, we can pick $c \in F$ and primitive polynomials $g^*(x)$ and $h^*(x)$ so that $f(x) = cg^*(x)h^*(x)$. Pick $a, b \in D$ so that a and b are relatively prime and $c = \frac{a}{b}$. Then

$$bf(x) = a\big(g^*(x)h^*(x)\big).$$

Gauss's Lemma tells us that $g^*(x)h^*(x)$ is primitive. So the uniqueness assertion of Lemma A gives us two unit u and v of $\mathbf{D}$ such that

$$ub = a \text{ and } f(x) = vg^*(x)h^*(x).$$

Since $f(x)$ is irreducible in $\mathbf{D}[x]$ it must be that one of $g^*(x)$ and $h^*(x)$ is a unit and thus has degree 0. But then one of $g(x)$ and $h(x)$ must also have degree 0 and be, therefore, a unit of $\mathbf{F}$. This means that $f(x)$ is irreducible in $\mathbf{F}[x]$. □

Here is a proof of the Unique Factorization Theorem for Polynomials over a Unique Factorization Domain.

Proof. Let $f(x) \in D[x]$ be a nonzero polynomial. We begin by letting c be a greatest common divisor of the coefficients of $f(x)$ we obtain a primitive polynomial $g(x)$ so that

$$f(x) = cg(x).$$

Either c is a unit of $\mathbf{D}$ or else we can factor it into irreducibles over $\mathbf{D}$. Observe that apart from units $g(x)$ has no factors of degree 0 in $\mathbf{D}[x]$ since $g(x)$ is primitive. Thus any proper factorization of $g(x)$ over $\mathbf{D}[x]$ must produce factors of properly smaller degree. In this way we see that $f(x)$ can be factored into irreducibles over $\mathbf{D}[x]$.

To see that the factorization of $f(x)$ is unique suppose

$$f(x) = c_0c_1\cdots c_m g_0(x)g_1(x)\cdots g_n(x)$$
$$f(x) = d_0d_1\cdots d_k h_0(x)h_1(x)\cdots h_\ell(x)$$

are factorization of $f(x)$ into irreducibles over $\mathbf{D}[x]$ so that $c_0, c_1, \ldots, c_m, d_0, d_1, \ldots d_k$ are irreducibles of degree 0 while the remaining irreducible factors have positive degree. Irreducibles in $\mathbf{D}[x]$ of positive degree are primitive. Using Gauss's Lemma and Lemma A, we see that

(a) $c_0c_1\cdots c_m$ and $d_0d_1\cdots d_k$ are associates over $\mathbf{D}$. Since $\mathbf{D}$ is a unique factorization domain, we find that $m = k$ and, perhaps after some reindexing, c_i and d_i are associates for all $i \le m$.

(b) $g_0(x)g_1(x)\cdots g_n(x)$ and $h_0(x)h_1(x)\cdots h_\ell(x)$ are associates over $\mathbf{D}[x]$ and hence over $\mathbf{F}[x]$. By Lemma B, these polynomials are irreducible over $\mathbf{F}[x]$. Because $\mathbf{F}[x]$ is a unique factorization domain, we find that $n = \ell$ and, after a suitable reindexing, that $g_j(x)$ and $h_j(x)$ are associates over $\mathbf{F}[x]$, for all $j \le n$. But the $g_j(x)$'s and the $h_j(x)$'s are primitive, so by Gauss's Lemma and Lemma A they must also be associates over $\mathbf{D}[x]$.

This establishes the uniqueness of the factorization. □

An easy induction shows that

if $\mathbf{D}$ is a unique factorization domain, then so is $\mathbf{D}[x_0, x_1, \ldots, x_{n-1}]$.

Eisenstein's Criteria. *Let* $\mathbf{D}$ *be a unique factorization domain and let* $\mathbf{F}$ *be its field of fractions. Let* $f(x) = a_0 + a_1x + \cdots + a_nx^n \in D[x]$ *where* $a_n \neq 0$ *and n is positive. If there is an irreducible* $p \in D$ *such that*

1. $p \mid a_i$ *for all* $i < n$,
2. $p \nmid a_n$, *and*
3. $p^2 \nmid a_0$,

then $f(x)$ *is irreducible in* $\mathbf{F}[x]$. *If, in addition,* $f(x)$ *is primitive, then* $f(x)$ *is irreducible in* $\mathbf{D}[x]$.

Proof. First suppose that $f(x)$ is primitive and that it satisfies the given criteria. Suppose $f(x) = g(x)h(x)$ is a factorization of $f(x)$ over $\mathbf{D}[x]$. Let $g(x) = b_0 + b_1x + \ldots$ and $h(x) = c_0 + c_1x + \ldots$. Then $a_0 = b_0c_0$. Now $p \mid a_0 = b_0c_0$ but $p^2 \nmid b_0c_0$. So p divides exactly one of b_0 and c_0. It is harmless to suppose that $p \mid b_0$ but $p \nmid c_0$. Now p cannot divide all the coefficients of $g(x)$ since then it would divide all the coefficients of $f(x)$, even a_n. Pick k as small as possible so that $p \nmid b_k$. Observe that

$$a_k = b_0c_k + b_1c_{k-1} + \cdots + b_{k-1}c_1 + b_kc_0.$$

Now $p \mid b_0c_k + b_1c_{k-1} + \cdots + b_{k-1}c_1$ but $p \nmid b_k$ and $p \nmid c_0$. Since p is prime we get $p \nmid b_kc_0$. But this implies that $p \nmid a_k$. We conclude that $k = n$. But this means that $\deg f(x) = \deg g(x)$ and that $\deg h(x) = 0$. So $h(x) \in D$. Since $f(x) = g(x)h(x)$ and $f(x)$ is primitive, we find that $h(x)$ must actually be a unit of $\mathbf{D}$. So $f(x)$ is irreducible in $\mathbf{D}[x]$. By Lemma B it is also irreducible in $\mathbf{F}[x]$.

Now consider the case when $f(x)$ is not primitve. Let c be the greatest common divisor of the coefficients of $f(x)$. So $f(x) = cf^*(x)$ where $f^*(x)$ primitive. Now observe that $p \nmid c$ since $c \mid a_n$. By the primeness of p, it follows that $f^*(x)$ satisfies Eisenstein's Criteria for the prime p. Hence $f^*(x)$ is irreducible in $\mathbf{D}[x]$ and hence in $\mathbf{F}[x]$ by Lemma B. But c is a unit of $\mathbf{F}[x]$ so $f(x)$ is an associate over $\mathbf{F}[x]$ of an irreducible. This makes $f(x)$ irreducible over $\mathbf{F}[x]$, as desired. □

Here is an example of what is at stake. The polynomial $6+3x$ has integer coefficients and it satisfies Eisenstein's Criteria with $p=2$. So it is irreducible over $\mathbb{Q}[x]$ by Eisenstein (but really, every polynomial of degree 1 is irreducible over $\mathbb{Q}[x]$). However, $6+3x = 3(2+x)$ is a proper factorization over $\mathbb{Z}[x]$ since 3 is not a unit for $\mathbb{Z}[x]$. Of course, $6+3x$ is also not primitive.

Eisenstein's Criteria is one of a large assortment of techniques for showing that polynomials are irreducible—especially polynomials in rings like $\mathbb{Z}[x], \mathbb{Z}[x,y], \ldots$.

There are a number of devices that can extend the range of polynomials to which Eisentstein's Criteria apply. For example, for any prime p the polynomial

$$f(x) = x^{p-1} + x^{p-2} + x^{p-3} + \cdots + x + 1$$

turns out to be irreducible over $\mathbb{Z}$. Of course, Eisenstein's Criteria does not apply to this polynomial. However, we can use the homomorphism extension property of polynomial rings to obtain a homomorphism from $\mathbb{Z}[x]$ to $\mathbb{Z}[x]$ that fixes each integers and sends $x \mapsto x+1$. The graduate students will see that this map is actually an automorphism that sends $f(x)$ to $f(x+1)$. So $f(x)$ will be irreducible if and only if $f(x+1)$ is irreducible. The idea is to apply Eisenstein's Criteria to $f(x+1)$. This means we need to figure out the coefficients of $f(x+1)$. Here are a couple of hints, for the hard working graduate students.

$$\begin{aligned}
(x-1)f(x) &= x^p - 1 \\
xf(x+1) &= (x+1)^p - 1 \\
&= \left(\sum_{k\leq p} \binom{p}{k} x^k\right) - 1 \\
f(x+1) &= \sum_{0<k\leq p} \binom{p}{k} x^{k-1}.
\end{aligned}$$

6.3 Hilbert's Basis Theorem

Now we know that if $\mathbf{D}$ is a principal ideal domain, then $\mathbf{D}[x]$ is a unique factorization domain, even though it might not be a principal ideal domain. Here we will see that, in some measure, $\mathbf{D}[x]$ retains some features of a principal ideal domain.

We will call a ring $\mathbf{R}$ **Noetherian** provided every ideal of $\mathbf{R}$ is finitely generated. So every principal ideal domain is Noetherian.

Theorem Characterizing Noetherian Rings. *Let* $\mathbf{R}$ *be a ring. The following are logically equivalent.*

(a) $\mathbf{R}$ *is a Noetherian ring.*

(b) *Every ascending chain of ideals of* $\mathbf{R}$ *is finite.*

(c) *Every nonempty collection of ideals of* $\mathbf{R}$ *has a maximal member with respect to the ordering by inclusion.*

Proof.

(a)⇒(b)
Suppose $I_0 \subseteq I_1 \subseteq I_2 \subseteq \ldots$ is an ascending chain of ideals of $\mathbf{R}$. Then $\bigcup_{i\in\mathbb{N}} I_i$ is also an ideal of $\mathbf{R}$. Because $\mathbf{R}$ is Noetherian there is a finite set X so that $(X) = \bigcup_{i\in\mathbb{N}} I_i$. Because X is finite and the ideals form a chain, there must be a natural number k so that $X \subseteq I_k$. But then

$$I_k \subseteq \bigcup_{i\in\mathbb{N}} I_i = (X) \subseteq I_k.$$

It follows that $I_k = I_{k+1} = I_{k+2} = \dots$. So the ascending chain of ideals is finite.

(b)⇒(c)
Let $\mathcal{F}$ be a nonempty family of ideals of $\mathbf{R}$. Since every ascending chain of ideals of $\mathbf{R}$ is finite, it follows that every chain of ideals in $\mathcal{F}$ has an upper bound in $\mathcal{F}$. By Zorn's Lemma, $\mathcal{F}$ must have maximal members.

(c)⇒(a)
Let I be an ideal of $\mathbf{R}$. Let $\mathcal{F} = \{J \mid J \subseteq I \text{ and } J \text{ is a finitely generated ideal}\}$. Let M be a maximal member of $\mathcal{F}$. Then $M \subseteq I$. Were $M \neq I$ we could pick $a \in I \setminus M$. But then $M \subsetneqq (M \cup \{a\}) \subseteq I$. Since $(M \cup \{a\})$ is finitely generated, this violates the maximality of M. So $I = M$, which is finitely generated. □

Hilbert's Basis Theorem. *If* $\mathbf{R}$ *is a commutative Noetherian ring, then so is* $\mathbf{R}[x]$.

Proof. Let I be any ideal of $\mathbf{R}[x]$ and let m be any natural number. Define

$$I(m) := \{a \mid a \text{ is the leading coefficient of a polynomial of degree } m \text{ that belongs to } I\} \cup \{0\}$$

The graduate students should routinely check that $I(m)$ is always an ideal of $\mathbf{R}$. It should also be clear that $I(m) \subseteq I(m+1)$.

Fact. Suppose I and J are ideals of $\mathbf{R}[x]$ with $I \subseteq J$. If $I(m) = J(m)$ for all natural numbers m, then $I = J$.

To establish this fact one should consider $f(x) \in J$ with the object of proving that $f(x) \in I$. This can be done by induction on the degree of $f(x)$. This induction is left for the pleasure of the graduate students.

Now consider an ascending chain $I_0 \subseteq I_1 \subseteq I_2 \subseteq \dots$ of ideals of $\mathbf{R}[x]$. There is an associated grid of ideals on $\mathbf{R}$.

$$\begin{array}{ccccccc}
\vdots & & \vdots & & \vdots & & \\
\cup & & \cup & & \cup & & \\
I_0(2) & \subseteq & I_1(2) & \subseteq & I_2(2) & \subseteq & \cdots \\
\cup & & \cup & & \cup & & \\
I_0(1) & \subseteq & I_1(1) & \subseteq & I_2(1) & \subseteq & \cdots \\
\cup & & \cup & & \cup & & \\
I_0(0) & \subseteq & I_1(0) & \subseteq & I_2(0) & \subseteq & \cdots
\end{array}$$

The family $\mathcal{F} = \{I_i(j) \mid i\,j \in \mathbb{N}\}$ displayed on this grid is a nonempty family of ideals of $\mathbf{R}$. It must have a maximal member, say $I_n(m)$. Each of the finitely many rows associated an argument j with $j \leq m$ is an ascending chain and can only extend to the right finitely far. Let ℓ be a natural number large enough so that none of these finitely many rows extends beyond ℓ steps. Notice that $n \leq \ell$. Then $I_\ell(i) = I_{\ell+k}(i)$ for all $i \leq m$ and for all natural numbers k, while $I_\ell(i) = I_n(m) = I_{\ell+k}(i)$ whenever $i > m$. Now the Fact asserted above tells us that $I_\ell = I_{\ell+k}$ for all natural numbers k. So the ascending chain $I_0 \subseteq I_1 \subseteq I_2 \subseteq \cdots$ is finite, as desired. □

It follows that if $\mathbf{R}$ is any commutative Noetherian ring, then $\mathbf{R}[x_0, x_1, \dots, x_{n-1}]$ is also a commutative Neotherian ring. This theorem has a fundamental role to play in commutative algebra and algebraic geometry. The proof I gave above has the charm of an illuminating diagram, but it doesn't allow us to lay our hands directly on a finite generating set for an ideal I of $\mathbf{R}[x]$. Coupled with the proof of the Fact embedded in our proof, some headway could be made in this direction.

6.4 Problem Set 5

Algebra Homework, Edition 5
Sixth Week
Rings of Polynomials

Problem 25.
Is the polynomial $y^3 - x^2y^2 + x^3y + x + x^4$ irreducible in $\mathbb{Z}[x, y]$?

Problem 26.
Let $\mathbf{R}$ be a principal ideal domain, and let I and J be ideals of $\mathbf{R}$. IJ denotes the ideal of $\mathbf{R}$ generated by the set of all elements of the form ab where $a \in I$ and $b \in J$. Prove that if $I + J = R$, then $I \cap J = IJ$.

Problem 27.
Let $\mathbf{D}$ be a unique factorization domain and let I be a nonzero prime ideal of $\mathbf{D}[x]$ which is minimal among all the nonzero prime ideals of $\mathbf{D}[x]$. Prove that I is a principal ideal.

Problem 28.
Let $\mathbf{D}$ be a subring of the field $\mathbf{F}$. An element $r \in F$ is said to be **integral over D** provided there is a monic polynomial $f(x) \in \mathbf{D}[x]$ such that r is a root of $f(x)$. For example, the real number $\sqrt{2}$ is integral over the ring of integers since it is a root of $x^2 - 2$.

Now suppose $\mathbf{D}$ is a unique factorization domain and $\mathbf{F}$ is its field of fractions. Prove that the set of elements of $\mathbf{F}$ that are integral over $\mathbf{D}$ coincides with D itself.

Problem 29.
Let $\mathbf{R}$ be a commutative ring and let $\mathbf{S}$ be a subring of $\mathbf{R}$ so that $\mathbf{S}$ is Noetherian. Let $a \in R$ and let $\mathbf{S}'$ be the subring of $\mathbf{R}$ generated by $S \cup \{a\}$. Prove that $\mathbf{S}'$ is Noetherian.

LECTURE 7

MODULES, A GENERALIZATION OF VECTOR SPACES

7.1 MODULES OVER A RING

A vector space over a field **F** is a set of vectors, including a zero vector, that has a two-place operation for vector addition, a one-place operation for forming the negative of a vector, a one-place operation for each element of F that can be used to scale vectors. Most of us were brought up to consider a kind of two-place operation for multiplying a vector by a scalar. For example, in the standard 2-dimensional vector space over the reals, when the vector $(2,6)$ is multiplied by the scalar 0.5 the resulting vector is $(1,3)$, a vector pointing in the same direction as $(2,6)$ but it is scaled down—it is only half as long. The one-place operation that sends each real pair (a,b) to $(0.5a,0.5b)$ precisely captures the effect of multiplication by the scalar 0.5. Of course, the advantage to us of construing scalar multiplication as a system of one-place operations is that then vector spaces fit into our overall view of algebraic systems in general.

Let **F** be a field. We say that $\langle V,+,-,0,r\cdot\rangle_{r\in F}$ is a **vector space over F** provided all of the equations below hold.

$$\begin{aligned} x+(y+z) &= (x+y)+z & \qquad (r+s)\cdot x &= rx+sx\\ x+y &= y+x & r(x+y) &= rx+ry\\ -x+x &= 0 & (rs)x &= r(sx)\\ x+0 &= x & 1x &= x \end{aligned}$$

for all $x,y,z\in V$ and $r,s\in F$.

We have followed the customary practice of using the same symbol $+$ to denote both the addition in the field of scalars and the addition in the vector space. Really, they are different in all but a few cases. The same might be said for using juxtaposition to denote the multiplication in the ring and the (one-place functional) action of a scalar on a vector. In the equations above rs is the product in the ring whereas $r(sx)$ means the action, consecutively, of two scalings.

We obtain the notion of a module over a ring by replacing the field **F** with an arbitrary ring **R**. So let **R** be a ring. We say that $\langle V,+,-,0,r\cdot\rangle_{r\in R}$ is a **module over R** provided all of the equations below hold.

$$\begin{aligned} x+(y+z) &= (x+y)+z & (r+s)\cdot x &= rx+sx \\ x+y &= y+x & r(x+y) &= rx+ry \\ -x+x &= 0 & (rs)x &= r(sx) \\ x+0 &= x & 1x &= x \end{aligned}$$

for all $x,y,z \in V$ and $r,s \in R$. Some people would say *left unitary* **R***-module* for this notion. The "left" comes from writing the scalars on the left—there is a companion notion of right modules. The "unitary" comes from the stipulation $1x = x$. Many of the most striking properties of vector spaces rely on the fact that every nonzero element of a field is a unit. Still, modules in general retain some of the nice features of vector spaces.

There is another source of modules. Let I be an ideal of **R**. Then $\langle I,+,-,0,r\cdot\rangle_{r\in R}$ is clearly an **R**-module. This would even be true if I were a "left" ideal of **R**. Indeed, the left ideals of **R** are essentially the same as the submodules of **R**. Below we will only be concerned with **R**-modules when **R** is a commutative ring. In this case, the ideals of **R** and the submodules of **R** coincide.

In fact, we will be almost exclusively concerned with modules whose underlying ring is a principal ideal domain. The familiar ring $\mathbb{Z}$ of integers and rings of the form $\mathbf{F}[x]$, where **F** is a field, are examples of principal ideal domains. Reflect a moment on the $\mathbb{Z}$-modules. Did you notice that the investigation of the $\mathbb{Z}$-modules differs in no important way from the investigation of Abelian groups?

Let **V** be a finite dimensional vector space over a field **F**. The linear operators (alias endomorphisms) of **V** can be acted on in an obvious way by the polynomials in $\mathbf{F}[x]$. Under this action, the linear operators of **V** form a module over $\mathbf{F}[x]$. Investigation of the structure of such modules leads to some of the deeper results in linear algebra.

7.2 Free Modules

A module **F** over a nontrivial ring **R** is said to be **free on** a set $B \subseteq F$ provided for every **R**-module **M** and every function $\varphi : B \to M$ there is a unique homomorphism $\psi : \mathbf{F} \to \mathbf{M}$ that extends φ. We will say that **F** is a **free R-module** provided it is free on some set. In the context of vector spaces, we know that every vector space has a basis B and that any map from B into another vector space over the same field extends uniquely to a linear transformation. This means that every vector space is free on any of its bases. This fails for modules in general. The free modules are much more like vector spaces.

The Uniqueness Theorem for Free Modules. *Let* **F** *be a module over a nontrivial ring* **R** *that is free on B and let* $\mathbf{F}^*$ *be a module over* **R** *that is free on B^*. If $|B| = |B^*|$, then* **F** *and* $\mathbf{F}^*$ *are isomorphic.*

Proof. Let φ be a one-to-one map from B onto B^*. Let ψ extend φ to a homomorphism from **F** into $\mathbf{F}^*$. Likewise let ψ^* extend φ^{-1} to a homomorphism from $\mathbf{F}^*$ into **F**. Then $\psi \circ \psi^*$ is an endomorphism of **F** extending the identity map on B. The identity on F is also such an endomorphism. By the uniqueness of such extensions, we find $\psi \circ \psi^*$ is that identity map on F. Likewise, $\psi^* \circ \psi$ is the identity map on F^*. So ψ is an isomorphism from **F** onto $\mathbf{F}^*$ and ψ^* is its inverse. □

Observe that **R** is an **R**-module that is free on $\{1\}$ and that the trivial **R**-module is free on $\varnothing$. We will see that free **R**-modules have a simple form. For this we employ the notion of direct sum of modules. Let $\mathbf{M}_i$ be an **R**-module for each $i \in I$, where I is any set. We define the **direct sum**

$$\bigoplus_{i\in I} M_i := \{\langle v_i | i \in I\rangle | v_i \in M_i \text{ for all } i \in I \text{ and all but finitely many of } v_i\text{'s are } 0\}$$

It is routine to check that this set is a subuniverse of $\prod_{i\in I}\mathbf{M}_i$. So $\bigoplus_{i\in I}\mathbf{M}_i$ is an **R**-module.

The Structure Theorem for Free Modules. *Suppose* **F** *is a module over a nontrivial ring* **R** *that is free on* B*. For each* $b \in B$ *let* $\mathbf{R}_b = \mathbf{R}$*. Then* **F** *is isomorphic to* $\bigoplus_{b\in B} \mathbf{R}_b$*.*

Proof. All we need to do is prove that $\bigoplus_{b\in B} R_b$ is free on a set of cardinality $|B|$. The set we are after is the set of all B-tuples that have 1 in exactly one position and 0 in all other positions. This is the "standard" basis familiar from linear algebra. The graduate students should enjoy filling in the rest of this proof. □

As in vector spaces, in modules generally we will say that a set X is **linearly independent** provided that for any finitely many distinct $v_0, v_1, \dots, v_{n-1} \in X$ if $a_0 v_0 + a_1 v_1 + \dots + a_{n-1} v_{n-1} = 0$, then $a_0 = a_1 = \dots = a_{n-1} = 0$. In any module **M**, a linearly independent subset that generates **M** is said to be a **basis** for **M**.

Theorem Characterizing Free Modules. *Let* **R** *be a nontrivial ring and* **F** *be an* **R***-module.* **F** *is a free* **R***-module if and only if* **F** *has a basis.*

Proof. Suppose that **F** is a module over **R** that is free on B. Let **M** be the submodule of **F** generated by B. I leave it to the graduate students to check that **M** is also free on B. So there is an isomorphism from **M** onto **F** that extends the identity map on B. But any such extension must fix each element of M since **M** is generated by B. This means that $M = F$, and so we see that B generates **F**. Next, observe that the subset of $\bigoplus_{b\in B} R_b$ consisting of those B-tuples with exactly one entry 1 and the rest 0 is evidently linearly independent. But $\bigoplus_{b\in B} \mathbf{R}_b$ and **F** are isomorphic via an isomorphism that sends our linear independent subset of the direct sum to B. As the image of a linearly independent set under an isomorphism is again linearly independent, we find the B is linearly independent. Therefore, B is a basis for **F**.

Now suppose that B is a basis for **F**. Just as in linear algebra, we can show that every element of F can be expressed uniquely as a linear combination of elements of B. Suppose that **M** is an **R**-module and let $\varphi : B \to M$. Define $\psi : F \to M$ via

$$\psi(w) := a_0\varphi(v_0) + \dots + a_n\varphi(v_n)$$

for all $w \in F$, where $a_0 v_0 + \dots + a_n v_n$ is the unique linear combination of distinct elements $v_0, \dots, v_n \in B$ that represents w. It is routine to prove that ψ is a homomorphism. So **F** is free on B. □

One of the most useful features of vector spaces is that any two bases of the same space have the same cardinality. This gives us a notion of dimension in vector spaces. This property is lost in some free modules. On the other hand, it is often true.

The Dimension Theorem for Free Modules. *Let* **R** *be a ring that has a division ring as a homomorphic image. Any two bases of a free* **R***-module have the same cardinality.*

Proof. Let I be an ideal of **R** so that $\mathbf{R}/I$ is a division ring. Let **F** be a free **R**-module with basis B. Let E be the collection of all elements that can be written as linear combinations of elements of F using only coefficients from I. You should check that E is closed under the module operations, so **E** is a submodule of **F**.

Observe that $\mathbf{F}/E$ can be construed as an $\mathbf{R}/I$-module in a natural way. (Hint: define $(a+I)(v+E)$ to be $av+E$. Be sure to check that this definition is sound.) Now let $B^* = \{v+E \mid v \in B\}$.

We want to demonstrate that B^* is linearly independent for the $\mathbf{R}/I$-module $\mathbf{F}/E$. Suppose $v_0 + E, v_1 + E, \dots, v_n + E$ are distinct members of $B^* = B/E$. Take $a_0, \dots, a_n \in R$. Observe the following sequence of steps.

$$\begin{aligned} 0 + E &= \sum_{i\le n} (a_i + I)(v_i + E) \\ &= \sum_{i\le n} (a_i v_i + E) \\ &= \left(\sum_{i\le n} a_i v_i\right) + E \end{aligned}$$

This means that if $0+E=\sum_{i\leq n}(a_i+I)(v_i+E)$, then $\sum_{i\leq n} a_i v_i \in E$. By the definition of E, there are $w_0,\dots,w_m \in F$ and $c_0,\dots,c_m \in I$ so that

$$\sum_{i\leq n} a_i v_i = \sum_{j\leq m} c_j w_j.$$

Now because B generates $\mathbf{F}$ we see that each of the w_j's can be written as a linear combination of elements of B. This entails that $\sum_{j\leq m} c_j w_j$ can be rewritten as a linear combination of elements of B with the coefficients all belonging to I. Let $\sum_{k\leq\ell} d_k u_k$ be such a linear combination. But now

$$0=\sum_{i\leq n} a_i v_i - \sum_{k\leq\ell} d_k u_k.$$

The expression on the right can be rewritten as a linear combination of distinct elements of B. The coefficients of the linear combination can be of three forms:

$$a_i - d_k \qquad \text{or} \qquad a_i \qquad \text{or} \qquad -d_k$$

depending on whether $v_i = u_k$. All of these coefficients must be 0. Notice that in the first alternative that we get $a_i \in I$ and in the second $a_i = 0 \in I$. So we find that $a_i \in I$ for all i. This means that $a_i + I = 0 + I$ for all i and concludes the proof that B^* is linearly independent.

That B^* generates $\mathbf{F}/E$ follows easily from the fact that B generates $\mathbf{F}$. So B^* is a basis of $\mathbf{F}/E$.

Contention. $|B| = |B^*|$.

In fact, the quotient map that send $v \mapsto v+E$ for $v \in B$ is one-to-one. To see this, suppose $v, v' \in B$ and $v+b = v'+E$. Then $v-v' \in E$. By the same device we used above, we can write $v-v'$ as a linear combination of distinct elements of B with coefficients drawn from I. Let $\sum_{k\leq\ell} d_k u_k$ be such a linear combination. This gives

$$0 = v' - v + \sum_{k\leq\ell} d_k u_k.$$

Notice that v' and v might well appear among the u_k's, but none of the d_k's is a unit of $\mathbf{R}$ since I must be a proper ideal. Nevertheless, rewriting the right side as a linear combination of distinct elements of B must result in all the coefficients being 0. This can only happen if $v' = v$, establishing our contention.

At this point we know that every basis of $\mathbf{F}$ has the same cardinality as some basis of $\mathbf{F}/E$. So the last thing we need is that any two bases of a free module over a division ring have the same cardinality. Proving this only requires a careful examination of any standard proof that any two bases of a vector space have the same cardinality. One must see that the commutative property of multiplication in the field plays no role in such a proof. It also pays to notice the role division has to play in such a proof. So commit due diligence on some linear algebra book to complete this proof. □

The unique dimension guaranteed by the theorem above is called the **rank** of the free modules.

By the Maximal Ideal Theorem, we know that any nontrivial commutative ring $\mathbf{R}$ has a maximal ideal I and so $\mathbf{R}/I$ is actually a field. So we have the following corollary.

Corollary 7.2.1. *Let* $\mathbf{R}$ *be a nontrivial commutative ring. Any two bases of the same free* $\mathbf{R}$*-module must have the same cardinality.*

Suppose $\mathbf{R}$ is a nontrivial commutative ring and $\mathbf{F}$ is a free $\mathbf{R}$-module. By the **rank** of $\mathbf{F}$ we mean the cardinality of any base of $\mathbf{F}$.

7.3 Problem Set 6

Algebra Homework, Edition 6
Seventh Week
Ideals Yet Again

Problem 30.

(a) Prove that $(2, x)$ is not a principal ideal of $\mathbb{Z}[x]$.

(b) Prove that (3) is a prime ideal of $\mathbb{Z}[x]$ that is not a maximal ideal of $\mathbb{Z}[x]$.

Problem 31.
Show that any integral domain satisfying the descending chain condition on ideals is a field.

Problem 32.
Prove the following form of the Chinese Remainder Theorem: Let **R** be a commutative ring with unit 1 and suppose that I and J are ideals of R such that $I + J = R$. Then

$$\mathbf{R}\Big/(I \cap J) \cong \mathbf{R}/I \times \mathbf{R}/J.$$

Problem 33.
Prove that there is a polynomial $f(x) \in \mathbb{R}[x]$ such that

(a) $f(x) - x$ belongs to the ideal $(x^2 + 2x + 1)$;

(b) $f(x) - x^2$ belongs to the ideal $(x - 1)$, and

(c) $f(x) - x^3$ belongs to the ideal $(x^2 - 4)$.

Lecture 8

Submodules of Free Modules over a PID

The objective here is to prove that, over a principal ideal domain, every submodule of a free module is also a free module and that the rank of a free module is always at least as large of the ranks of its submodules.

So let **R** be a (nontrivial) principal ideal domain. We know that **R** is a free **R**-module of rank 1. What about the submodules of **R**? Suppose **E** is such a submodule. It is clear that E is an ideal and, in fact, that the ideals of **R** coincide with the submodules of **R**. In case **E** is trivial (that is the sole element of E is 0) we see that **E** is the free **R**-module of rank 0. So consider the case that **E** is nontrivial. Since **R** is a principal ideal domain we pick $w \neq 0$ so that **E** is generated by w. That is $E = \{rw \mid r \in R\}$. Since we know that **R** has $\{1\}$ as a basis, we see that the map that sends 1 to w extends to a unique module homomorphism from **R** onto **E**. Indeed, notice $h(r \cdot 1) = r \cdot h(1) = rw$ for all $r \in R$. But the homomorphism h is also one-to-one since

$$\begin{aligned} h(r) &= h(s) \\ rh(1) &= sh(1) \\ rw &= sw \\ r &= s \end{aligned}$$

where the last step follows because integral domains satisfy the cancellation law and $w \neq 0$. In this way we see that **E** is isomorphic to the free **R**-module of rank 1. We also see that $\{w\}$ is a basis for **E**.

So we find that at least all the submodules of the free **R**-module of rank 1 are themselves free and have either rank 0 or rank 1. We can also see where the fact that **R** is a principal ideal domain came into play.

The Freedom Theorem for Modules over a PID.
Let **R** *be a principal ideal domain, let* **F** *be a free* **R***-module and let* **E** *be a submodule of* **F**. *Then* **E** *is a free* **R***-module and the rank of* **E** *is no greater than the rank of* **F**.

Proof. Since trivial modules (those whose only element is 0) are free modules of rank 0, we suppose below that **E** is a nontrivial module. This entails that **F** is also nontrivial.

Let B be a basis for **F** and $C \subseteq B$. Because **F** is not the trivial module, we see that B is not empty. Let $\mathbf{F}_C$ be the submodule of **F** generated by C. Let $\mathbf{E}_C = \mathbf{E} \cap \mathbf{F}_C$. Evidently, C is a basis for $\mathbf{F}_C$. To see that $\mathbf{E}_C$ is free we will have to find a basis for it.

Suppose, for a moment, that C has been chosen so that $\mathbf{E}_C$ is known to be free and that $w \in B$ with $w \notin C$. Put $D := C \cup \{w\}$. Consider the map defined on D into R that sends all the elements of C to 0 and that sends w to 1. This map extends uniquely to a homomorphism φ from $\mathbf{F}_D$ onto **R** and it is easy to check (as hardworking graduate student will) that the kernel of φ is just F_C. By the Homomorphism Theorem, we draw the conclusion that $\mathbf{F}_D/F_C$ is isomorphic to $\mathbf{R}$ and that it is free of rank 1. What about $\mathbf{E}_D/E_C$? Observe that $E_C = E \cap F_C = E \cap F_D \cap F_C = E_D \cap F_C$. So we can apply the Second Isomorphism Theorem:

$$\mathbf{E}_D/E_C = \mathbf{E}_D/(E_D \cap F_C) cong (\mathbf{E}_D + \mathbf{F}_C)/F_C.$$

But $(\mathbf{E}_D + \mathbf{F}_C)/F_C$ is a submodule of $\mathbf{F}_D/F_C$. This last is a free $\mathbf{R}$-module of rank 1. We saw above that every submodule of a free $\mathbf{R}$-module of rank 1 must be itself a free $\mathbf{R}$-module and have rank either 0 or 1. In this way, we find that either $\mathbf{E}_D = \mathbf{E}_C$ (in the rank 0 case) or else $\mathbf{E}_D/E_C$ is a free $\mathbf{R}$-module of rank 1. Let us take up this latter case. Let X be a basis for $\mathbf{E}_C$, which we assumed, for the moment, was free. Pick $u \in E_D$ so that $\{u/E_C\}$ is a basis for $\mathbf{E}_D/E_C$.

We contend that $X \cup \{u\}$ is a basis for $\mathbf{E}_D$. To establish linear independence, suppose that $x_0, \dots, x_{n-1}$ are distinct elements of X, that $r_0, \dots, r_n \in R$ and that

$$0 = r_0x_0 + \cdots + r_{n-1}x_{n-1} + r_nu.$$

First notice that

$$r_n(u/E_C) = r_nu/E_C = (r_0x_0 + \cdots + r_{n-1}x_{n-1} + r_nu)/E_C = 0/E_C.$$

Since $\{u/E_C\}$ is a basis for $\mathbf{E}_D/E_C$, we must have $r_n = 0$. This leads to

$$0 = r_0x_0 + \cdots + r_{n-1}x_{n-1}.$$

But now since X is a basis for $\mathbf{E}_C$ we see that $0 = r_0 = \cdots = r_{n-1}$. So we find that $X \cup \{u\}$ is linearly independent.

To see that $X \cup \{u\}$ generates E_D, pick $z \in E_D$. Since $\{u/E_C\}$ is a basis for $\mathbf{E}_D/E_C$, pick $r \in R$ so that

$$z/E_C = ru/E_C.$$

This means that $z - ru \in E_C$. But X is a basis of $\mathbf{E}_C$. So pick $x_0, \dots, x_{n-1} \in X$ and $r_0, \dots, r_{n-1} \in R$ so that

$$z - ru = r_0x_0 + \cdots + r_{n-1}x_{n-1}.$$

Surely this is enough to see that z is in the submodule generated by $X \cup \{u\}$. So this set generates $\mathbf{E}_D$ and we conclude that it must be a basis of $\mathbf{E}_D$.

In this way we see that for $C \subseteq D \subseteq B$ where D arises from adding an element to C, if $\mathbf{E}_C$ is free, then so is $\mathbf{E}_D$ and that either $E_D = E_C$ or a basis for $\mathbf{E}_D$ can be produced by adding just one element to a basis for $\mathbf{E}_C$.

With this in mind, we can envision a procedure for showing that $\mathbf{E}$ is free and its rank cannot be larger than that of $\mathbf{F}$. Notice that $E = E \cap F = E \cap F_B$. So $\mathbf{E} = \mathbf{E}_B$. The idea is simple. We will start with $\varnothing \subseteq B$. We observe that $\mathbf{F}_\varnothing = \mathbf{E}_\varnothing$ is the module whose sole element is 0. It is free of rank 0. Next we select an element $w \in B$ and form $\varnothing \cup \{w\} = \{w\}$. We find that $\mathbf{E}_{\{w\}}$ is free of rank 0 or rank 1. We select another element and another and another... until finally all the elements of B have been selected. At this point we would have E_B is free and its rank can be no more than the total number of elements we selected, namely $|B|$ which is the rank of $\mathbf{F}$.

To carry out this program, in case B were finite or even countable, we could mount a proof by induction. You can probably see how it might be done. But we want to prove this for arbitary sets B. We could still pursue this inductive strategy openly by well-ordering B and using transfinite induction. By using the well-ordering we would always know what was meant by "pick the next element of B."

Instead, we will invoke Zorn's Lemma to short-circuit this rather long induction.

Let $\mathcal{F} = \{f \mid f$ is a function with $\operatorname{dom} f \subseteq B$ and $\operatorname{range} f$ a basis for $\mathbf{E}_{\operatorname{dom} f}\}$. Recalling that functions are certain kinds of sets of order pairs, we see that $\mathcal{F}$ is paritally ordered by set inclusion. Maybe it helps to realize that asserting that $f \subseteq g$ is the same as asserting that g extends f. We note that $\mathcal{F}$ is not empty since the empty function (the function with empty domain) is a member of $\mathcal{F}$. To invoke Zorn's Lemma, let $\mathcal{C}$ be any chain included in $\mathcal{F}$. Let $h = \bigcup \mathcal{C}$. Evidently $f \subseteq h$ for all $f \in \mathcal{C}$. So h is an upper bound of $\mathcal{C}$. We contend that $h \in \mathcal{F}$. We ask the hard-working graduate students to check that the union of any chain of functions is itself a function. Once you do that bit of work, it should be evident that $\operatorname{dom} h = \bigcup\{\operatorname{dom} f \mid f \in \mathcal{C}\}$ and that $\operatorname{range} h = \bigcup\{\operatorname{range} f \mid f \in \mathcal{C}\}$. So it remains to show that $\operatorname{range} h$ is a basis for $\mathbf{E}_{\operatorname{dom} h}$. To see that $\operatorname{range} h$ is a generating set, let z be an arbitrary element of $E_{\operatorname{dom} h} = E \cap F_{\operatorname{dom} h}$. Hence z must be generated by some finitely many elements belong in $\operatorname{dom} h$. So there are finitely many functions $f_0, \dots, f_{n-1} \in \mathcal{C}$ so that z is generated by finitely many elements of $\operatorname{dom} f_0 \cup \cdots \cup \operatorname{dom} f_{n-1}$. But $\operatorname{dom} f_0, \dots, \operatorname{dom} f_{n-1}$, rearranged in some order, forms a chain under inclusion. So $z \in F_{\operatorname{dom} f_\ell}$ for some $\ell < n$. Hence $z \in E_{\operatorname{dom} f_\ell}$. But $\operatorname{range} f_\ell$ is a basis for $\mathbf{E}_{\operatorname{dom} f_\ell}$. Because $\operatorname{range} f_\ell \subseteq \operatorname{range} h$ we find that $\operatorname{range} h$ has enough elements to generate z. Since z was an arbitrary element of $E_{\operatorname{dom} h}$ we conclude that $\operatorname{range} h$ generates $E_{\operatorname{dom} h}$. It remains to show that $\operatorname{range} h$ is linearly independent. But $\operatorname{range} h$ is the union of the chain $\{\operatorname{range} f \mid f \in \mathcal{C}\}$. I ask the hard-working graduate students to prove that the union of any chain of linearly independent sets must also be linearly independent. Once you have done this you will be certain that h belongs to $\mathcal{F}$. By Zorn, let g be a maximal element of $\mathcal{F}$.

We would be done if $\operatorname{dom} g = B$, since then $E = E \cap F = E \cap F_B = E_B = E_{\operatorname{dom} g}$. In which case, $\operatorname{range} g$ would be a basis for $\mathbf{E}$ and $\operatorname{rank} \mathbf{E} = |\operatorname{range} g| \leq |\operatorname{dom} g| = |B| = \operatorname{rank} \mathbf{F}$.

Consider the possibility that $\operatorname{dom} g$ is a proper subset of B. Put $C = \operatorname{dom} g$ and put $X = \operatorname{range} g$. Let $w \in B$ with $w \notin \operatorname{dom} g$. Put $D = C \cup \{w\}$. As we have seen above, either $E_D = E_C$ or $X \cup \{u\}$ is a basis for $\mathbf{E}_D$, for some appropriately chosen u. We can now extend g to a function g' by letting $g'(w)$ be any element of $\operatorname{range} g$ in the case when $E_D = E_C$ and by letting $g'(w) = u$ in the alternative case. In this way, $g' \in \mathcal{F}$, contradicting the maximality of g. So we reject this possibility.

This completes the proof. □

Corollary 8.0.1. *Let* $\mathbf{R}$ *be a principal ideal domain. Every submodule of a finitely generated* $\mathbf{R}$*-module must itself be finitely generated.*

Proof. Suppose $\mathbf{M}$ is an $\mathbf{R}$-module generated by n elements. Let $\mathbf{N}$ be a submodule of $\mathbf{M}$.

Now let $\mathbf{F}$ be the free $\mathbf{R}$-module with a basis of n elements. There is a function that matches this basis with the generating set of $\mathbf{M}$. So, appealing to freeness, there is a homomorphism h from $\mathbf{F}$ onto $\mathbf{M}$. Let $E = \{v \mid v \in F \text{ and } h(v) \in N\}$. It is straightforward to check (will you do it?) that E is closed under the module operations. So we get a submodule $\mathbf{E}$ of $\mathbf{F}$. Moreover, the restriction of h to E is a homomorphism from $\mathbf{E}$ onto $\mathbf{N}$. But by our theorem $\mathbf{E}$ is generated by a set with no more than n elements. Since the image, under a homomorphism, of any generating set for $\mathbf{E}$ must be a generating set of $\mathbf{N}$ (can you prove this?), we find that $\mathbf{N}$ is finitely generated. □

8.1 Problem Set 7

Algebra Homework, Edition 7
Eighth Week
More on Polynomials and then Some

Problem 34.
Let **D** be an integral domain and let $c_0, \dots, c_{n-1}$ be n distinct elements of D. Further let $d_0, \dots, d_{n-1}$ be arbitrary elements of D. Prove there is at most one polynomial $f(x) \in D[x]$ of degree $n-1$ such that $f(c_i) = d_i$ for all $i < n$.

Problem 35.
Let **F** be a field and let $c_0, \dots, c_{n-1}$ be n distinct elements of F. Further let $d_0, \dots, d_{n-1}$ be arbitrary elements of F. Prove there is at least one polynomial $f(x) \in F[x]$ of degree n such that $f(c_i) = d_i$ for all $i < n$.

Problem 36.
Let **R** be the following subring of the field of rational functions in 3 variables with complex coefficients:

$$R = \left\{ \frac{f}{g} : f, g \in \mathbb{C}[x, y, z] \text{ and } g(1,2,3) \neq 0 \right\}$$

Find 3 prime ideals P_1, P_2, and P_3 in R with

$$0 \subsetneqq P_1 \subsetneqq P_2 \subsetneqq P_3 \subsetneqq R.$$

Problem 37.
Let **R** be a commutative ring. An **R**-module **P** is said to be **projective** provided for all **R**-modules **M** and **N** and all homomorphisms f from **M** onto **N**, if g is a homomorphism from **P** into **N**, then there is a homomorphism h from **P** into **M** so that $f \circ h = g$.

Prove that every free **R**-module is projective.

LECTURE 9

DIRECT DECOMPOSITION OF FINITELY GENERATED MODULES OVER A PID

9.1 THE FIRST STEP

The objective here is to show how to obtain a direct decomposition of a finitely generated module over a principal ideal domain. We would like that the direct factors admit no further nontrivial direct decomposition.

The operations of modules work in a direct product of modules in a coordinatewise manner. So knowing how to perform the operations in each direct factor leads immediately to a knowledge of how the operations work in the direct product. One point of inconvenience with direct products is that very few modules actually arise as direct products—simply because the elements of your favorite module are not tuples of any kind. So our direct decompositions make use of isomorphisms.

For the task at hand, our direct decompositions turn out to have only finitely many direct factors. In this situation, it is easy to replace the direct product with the notion of a direct sum. Suppose that we have the following direct decomposition of the **R**-module **M**:

$$\mathbf{M} \cong \mathbf{N} \times \mathbf{L}.$$

Then composing the isomorphism with the projection functions on the direct product, we find two homomorphisms $f : \mathbf{M} \twoheadrightarrow \mathbf{N}$ and $g : \mathbf{M} \twoheadrightarrow \mathbf{L}$ and these homomorphisms have the following properties:

(a) For every $v \in N$ and $w \in L$ there is some $u \in M$ so that $f(u) = v$ and $g(u) = w$.

(b) For every $u \in M$, if $f(u) = 0$ and $g(u) = 0$, then $u = 0$.

Another way to frame these two properties is in terms of the kernels of these homomorphisms. Let $\mathbf{N}'$ be the submodule that is the kernel of g and let $\mathbf{L}'$ be the submodule that is the kernel of f.

(a') For every $u \in M$ there are $v \in N'$ and $w \in L'$ so that $u = v + w$.

(b') The intersection of N' and L' is trivial.

Here is how to prove (a') from (a) and (b). Use (a) to get $w \in M$ such that $f(w) = 0$ and $g(w) = g(u)$. Then $g(u - w) = g(u) - g(w) = 0$. Observe that $u = (u - w) + w$ and $u - w \in \ker g = N'$ and $w \in \ker f = L'$ as desired.

I leave it to the hard-working graduate students to show that these two views (one from homomorphisms and one from kernels) are logically equivalent. The Homomorphism Theorem, after just a bit of work, yields that $\mathbf{N} \cong \mathbf{N}'$ and $\mathbf{L} \cong \mathbf{L}'$. This leads to the following definition. Let $\mathbf{N}'$ and $\mathbf{L}'$ be submodules of $\mathbf{M}$ that satisfy (a') and (b'). We say that $\mathbf{M}$ is a **direct sum** of $\mathbf{N}'$ and $\mathbf{L}'$ and we write $\mathbf{M} = \mathbf{N}' \oplus \mathbf{L}'$. Evidently, $\mathbf{N}' \oplus \mathbf{L}' \cong \mathbf{N}' \times \mathbf{L}'$.

We can extend this notion to three submodules $\mathbf{N}_0, \mathbf{N}_1$, and $\mathbf{N}_2$. Here is what works.

(a') For every $u \in M$ there are $v_0 \in N_0, v_1 \in N_1$, and $v_2 \in N_2$ so that $u = v_0 + v_1 + v_2$.

(b') The intersections $N_0 \cap (N_1 + N_2), N_1 \cap (N_0 + N_2)$, and $N_2 \cap (N_0 + N_1)$ are all trivial.

The hard-working graduate students should verify that this works and also that the obvious extension to any finite number of direct summands also succeeds.

Now let us turn to our task of decomposing modules. Here is a first step.

Fact. Let $\mathbf{R}$ be a nontrivial integral domain. As an $\mathbf{R}$-module $\mathbf{R}$ is directly indecomposable.

Proof. We know that $\mathbf{R}$ can be itself construed as an $\mathbf{R}$-module and as such it is a free $\mathbf{R}$-module of rank 1. To see that this module is directly indecomposable, suppose that $\mathbf{M}$ and $\mathbf{N}$ are $\mathbf{R}$-modules and that φ is an isomorphism from $\mathbf{R}$ onto $\mathbf{M} \times \mathbf{N}$. Let $M' = \{r \mid \varphi(r) = (u, 0) \text{ for some } u \in M\}$. Likewise, let $N' = \{r \mid \varphi(r) = (0, v) \text{ for some } v \in N\}$. Plainly, M' and N' are ideals in $\mathbf{R}$ and $M' \cap N' = \{0\}$ since φ is one-to-one. Since $\mathbf{R}$ is nontrivial, we see that $\mathbf{M}$ and $\mathbf{N}$ cannot both be trivial. Suppose, without loss of generality, that $\mathbf{M}$ is nontrivial. So M' is nontrivial. Pick $r \in M'$ with $r \neq 0$. We want to see that $\mathbf{N}$ must be a trivial module, or, what is the same, that N' is a trivial ideal. Let s be an arbitrary element of N'. Then $rs \in M' \cap N' = \{0\}$. That is, $rs = 0$. Since $r \neq 0$ and $\mathbf{R}$ is an integral domain, we conclude that $s = 0$. Since s was an arbitrary element of N', we have that N', and hence $\mathbf{N}$, is trivial. This means that $\mathbf{R}$ is a directly indecomposable $\mathbf{R}$-module, since $\mathbf{R}$ is itself nontrivial but in any direct decomposition we find that one of the direct factors must be trivial. □

This means that, over any nontrivial integral domain, any free module of finite rank directly decomposes into a direct product of finitely many copies of the ring; moreover, this direct decomposition is into directly indecomposable modules.

Here is another step we can take in directly decomposing a module.

Fact. Let $\mathbf{R}$ be a commutative ring. Suppose that $\mathbf{M}$ is an $\mathbf{R}$-module, that $\mathbf{F}$ is a free $\mathbf{R}$-module, and that f is a homomorphism from $\mathbf{M}$ onto $\mathbf{F}$ with kernel $\mathbf{N}$. Then there is a free $\mathbf{R}$-module $\mathbf{E}$ so that $\mathbf{M} \cong \mathbf{N} \times \mathbf{E}$.

Proof. Let B be a basis for $\mathbf{F}$. For each $u \in B$ pick $v_u \in M$ so that $f(v_u) = u$. The set $C = \{v_u \mid u \in B\}$ is a linearly independent subset of M. Here is how to see it:

> Let $w_0, \dots, w_{n-1}$ be finitely many distinct elements of C and let $r_0, \dots, r_{n-1} \in R$ with
>
> $$r_0 w_0 + \dots + r_{n-1} w_{n-1} = 0.$$
>
> Applying f to both sides we obtain
>
> $$r_0 f(w_0) + \dots + r_{n-1} f(w_{n-1}) = 0.$$
>
> But $f(w_0), \dots, f(w_{n-1})$ are distinct elements of B, which is linearly independent. So $r_0 = \dots = r_{n-1} = 0$, as desired.

Now let $\mathbf{E}$ be the submodule of $\mathbf{M}$ generated by C. So $\mathbf{E}$ is free since C is a basis. I contend that $\mathbf{M} \cong \mathbf{N} \times \mathbf{E}$. Here is how to define the isomorphism φ:

> Let w be an arbitrary element of M. Let $f(w) = r_0u_0 + \cdots + r_{n-1}u_{n-1}$ where $u_0, \ldots, u_{n-1}$ are distinct elements of B and all the r_i's are nonzero. Let $v_0, \ldots, v_{n-1}$ be elements of C so that $f(v_i) = u_i$ for all $i < n$. Put $x = r_0v_0 + \cdots + r_{n-1}v_{n-1}$. Then $x \in E$ and $f(x) = f(w)$. This means $w - x \in N$ since N is the kernel of f. So define $\varphi(w) = (w - x, x)$.

It is a straightforward piece of work (done by all hard working graduate students) to see that φ is an isomorphism. □

To invoke this last fact for a particular module **M** we essentially have to find a free submodule of **M**. Such a submodule would have a basis C. For $w \in C$ we would have to have the implication

$$rw = 0 \implies r = 0, \text{ for all } r \in R,$$

since this is just part of the definition of linear independence. Indeed, when **R** is an integral domain, all the elements of E, not just those in C would have to have this property. This suggests that N should consist of those elements that fail this property. That is elements $x \in M$ such that $rx = 0$ for some $r \neq 0$. Such elements are called **torsion** elements. The 0 of a module is always a torsion element, provided the ring is nontrivial. The module **M** is said to be **torsion free** provided 0 is its only torsion element. The step along the way is the following fact.

Fact. Let **R** be a nontrivial integral domain and let **M** be an **R**-module. Then the set T of torsion elements is a submodule of **M** and **M**/T is torsion free.

Proof. We have already noted that $0 \in T$. To see that T is closed under addition, let $u, v \in T$. Pick nonzero elements $r, s \in R$ so that $ru = 0 = sv$. Then $rs \neq 0$ since **R** is an integral domain. Now observe

$$(rs)(u+v) = (rs)u + (rs)v = (sr)u + (rs)v = s(ru) + r(sv) = 0 + 0 = 0$$

holds in **R**, since **R** is commutative. So $u + v \in T$. Finally, suppose that $t \in R$. Then $r(tu) = (rt)u = (tr)u = t(ru) = 0$, so $tu \in T$. In this way, we see that T is closed under the module operations and we can form the submodule **T**.

To see that **M**/T is torsion free, pick a nonzero element u/T of M/T and a scalar $r \in R$ so that $r(u/T) = 0/T$. Since u/T is nonzero we know that $u \notin T$. On the other hand $r(u/T) = (ru)/T = 0/T$ means that $ru \in T$. So pick a nonzero $s \in R$ so that $s(ru) = 0$. This means that $(sr)u = 0$. But, since $u \notin T$ we know that u is not a torsion element. So $sr = 0$. Since $s \neq 0$ and **R** is an integral domain, we see that $r = 0$. This means that u/T is not a torsion element. So **M**/T is a torsion free module. □

So when is a torsion free module actually free?

Fact. Let **R** be a nontrivial principal ideal domain. Every finitely generated torsion free **R**-module is free of finite rank.

Proof. Let **M** be a torsion free **R**-module generated by the finite set X. Let Y be a maximal linearly independent subset of X. Let **F** be the submodule of **M** generated by Y. Of course, **F** is free of finite rank. For each $x \in X$ pick $s_x \in R$ so that $s_x x \in F$. This is possible since if $x \in Y$ we can let $s_x = 1$, while if $x \notin Y$, then $Y \cup \{x\}$ is linearly dependent. This means that for some distinct $y_0, \ldots, y_{n-1} \in Y$ there are $s_x, r_0, \ldots, r_{n-1} \in R \setminus \{0\}$ so that

$$s_x x + r_0 y_0 + \cdots + r_{n-1} y_{n-1} = 0.$$

In this way, $s_x x \in F$. Now let s be the product of all the s_x's as x runs through X. Then $sx \in F$ for all $x \in X$. Since X generates **M**, we see that $sv \in F$ for all $v \in M$. Now let φ be the map from M into F defined via

$$\varphi(v) := sv \text{ for all } v \in M.$$

It is routine to check that φ is a homomorphism. The kernel of φ must be trivial since $\mathbf{M}$ is torsion free. So φ is one-to-one. This means that $\mathbf{M}$ is isomorphic with a submodule of the free module $\mathbf{F}$. Since $\mathbf{R}$ is a principal ideal domain, by the Freedom Theorem we conclude that $\mathbf{M}$ is free. Moreover, since $\mathbf{F}$ has finite rank, so must $\mathbf{M}$. □

The First Decomposition Theorem for Modules over an Integral Domain.
Let $\mathbf{R}$ *be a nontrivial principal ideal domain, let* $\mathbf{M}$ *be a finitely generated* $\mathbf{R}$*-module, and let* $\mathbf{T}$ *be the torsion submodule of* $\mathbf{M}$*. There is a free module* $\mathbf{F}$ *of finite rank such that*

$$\mathbf{M} \cong \mathbf{T} \times \mathbf{F}.$$

Moreover, the rank of $\mathbf{F}$ *is determined by* $\mathbf{M}$*.*

Proof. According to the Facts established above, we can take $\mathbf{F}$ to be $\mathbf{M}/T$. So only the "moreover" part of the theorem remains to be established. To this end, suppose that $\mathbf{F}'$ is some free module so that

$$\mathbf{M} \cong \mathbf{T} \times \mathbf{F}'.$$

The conclusion we want is $\mathbf{F} \cong \mathbf{F}'$.

What are the torsion elements of $\mathbf{T} \times \mathbf{F}'$? Suppose (u, v) is torsion. Pick $r \neq 0$ so that $r(u, v) = (0, 0)$. So $rv = 0$. But $v \in F'$, which being free is also torsion free. So $v = 0$. This means that the torsion elements of $\mathbf{T} \times \mathbf{F}'$ are exactly the elements of $T' := \{(u, 0) \mid u \in T\}$. In this way we see

$$\mathbf{F} = \mathbf{M}/T \cong (\mathbf{T} \times \mathbf{F}')/T' \cong \mathbf{F}'.$$

The rightmost isomorphism above comes from the Homomorphism Theorem since T' is the kernel of the project of the direct product onto its rightmost direct factor. □

Both the torsion module $\mathbf{T}$ and the free module $\mathbf{F}$ may admit further direct decomposition. As regards the free module, we know it can be decomposed as the direct product of n copies of the $\mathbf{R}$-module $\mathbf{R}$, which we have seen is directly indecomposable.

9.2 Problem Set 8

Algebra Homework, Edition 8
Ninth Week
Polynomials Again

Problem 38.
Prove that the polynomial $x^3y + x^2y - xy^2 + x^3 + y$ is irreducible in $\mathbb{Z}[x, y]$.

Problem 39.
Let $\mathbf{F}$ and $\mathbf{M}$ be modules over the same ring and let $\mathbf{F}$ be a free module. Let $h : \mathbf{M} \twoheadrightarrow \mathbf{F}$ be a homomorphism from $\mathbf{M}$ onto $\mathbf{F}$. Prove each of the following.

(a) There is an embedding $g : \mathbf{F} \rightarrowtail \mathbf{M}$ of $\mathbf{F}$ into $\mathbf{M}$ such that $h \circ g = \mathrm{id}_F$. (Here id_F denotes the identity map of the set F.)

(b) $\mathbf{M} = \ker h \oplus \mathbf{F}'$, where $\mathbf{F}'$ is the image of $\mathbf{F}$ with respect to g.

Problem 40.
Prove that there is a polynomial $f(x) \in \mathbb{R}[x]$ such that

(a) $f(x) - 1$ belongs to the ideal $(x^2 - 2x + 1)$;

(b) $f(x) - 2$ belongs to the ideal $(x + 1)$, and

(c) $f(x) - 3$ belongs to the ideal $(x^2 - 9)$.

9.3 The Second Step

Let $\mathbf{M}$ be any $\mathbf{R}$-module and X be any subset of M. Define

$$\operatorname{ann} X := \{r \mid r \in R \text{ and } rx = 0 \text{ for all } x \in X\}.$$

This set is called the **annihilator** of X. It is routine to check that $\operatorname{ann} X$ is always an ideal of $\mathbf{R}$, provided $\mathbf{R}$ is commutative. Running this game in the other direction, let $S \subseteq R$ and define

$$M[S] := \{u \mid u \in M \text{ and } ru = 0 \text{ for all } r \in S\}.$$

Again, it is routine to check that $M[S]$ is closed under the module operations, provided that $\mathbf{R}$ is commutative. So we obtain a submodule $\mathbf{M}[S]$ of the module $\mathbf{M}$. For a single element $r \in R$ we write $\mathbf{M}[r]$ for $\mathbf{M}[\{r\}]$.

Let $\mathbf{T}$ be a torsion module of a principal ideal domain $\mathbf{R}$. Suppose that the finite set X generates $\mathbf{T}$. As we did in one of the proofs in the preceding lecture, for each $x \in X$ we can pick a nonzero $s_x \in R$ so that $s_x x = 0$. Let s be the product of these finitely many s_x's as x runs through X. Since $\mathbf{R}$ is an integral domain we see that $s \neq 0$ and since $\mathbf{R}$ is commutative and X generates $\mathbf{T}$ we see that $su = 0$ for all $u \in T$. This means that $\operatorname{ann} T$ is a nontrivial ideal. Because $\mathbf{R}$ is a principal ideal domain we can pick $r \in R$ so that $(r) = \operatorname{ann} T$. This nonzero element r, which is unique up to associates, is called the **exponent of T**.

More generally, if u is a torsion element of an $\mathbf{R}$-module, where $\mathbf{R}$ is a principal ideal domain then there will be a nonzero element r so that $(r) = \operatorname{ann}\{u\}$. We call r the **order** of u and sometimes refer to $\operatorname{ann}\{u\}$ as the **order ideal** of u. If v is also a torsion element and s is the order of v, where r and s are relatively prime, then rs will be the order of $u + v$. (This could be proven by a hard working graduate student.)

We are ready to begin decomposing our torsion module.

Fact. Let $\mathbf{T}$ be a torsion $\mathbf{R}$-module with exponent r, where $\mathbf{R}$ is a principal ideal domain. Suppose that $r = sq$ where s and q are relatively prime. Then $\mathbf{T} \cong \mathbf{T}[s] \times \mathbf{T}[q]$.

Proof. Using the relative primeness of s and q select elements $a, b \in R$ so that $1 = as + bq$. So for any $u \in T$ we have

$$u = 1 \cdot u = (as + bq)u = q(bu) + s(au).$$

Observe that $q(bu) \in T[s]$ and $s(au) \in T[q]$. So every element of T can be expressed as a sum of an element of $T[s]$ and an element of $T[q]$. The expression is unique since if $u = v + w$ where $v \in T[s]$ and $w \in T[q]$, then $v - qbu = sau - w \in T[s] \cap T[q]$. But the order of any element of this intersection must divide both s and q, which are relatively prime. So the intersection is just $\{0\}$ and it follows that $v = qbu$ and $w = sau$. The map that sends u to (v, w) where $v \in T[s]$, $w \in T[q]$, and $u = v + w$ is easily seen to be an isomorphism. □

Suppose that r is the exponent of our torsion $\mathbf{R}$-module $\mathbf{T}$, where $\mathbf{R}$ is a principal ideal domain. Let $p_0, \ldots, p_{n-1} \in R$ be distinct primes and let $e_0, \ldots, e_{n-1}$ be positive integers so that

$$r = p_0^{e_0} \ldots p_{n-1}^{e_{n-1}}.$$

Then applying the Fact above over and over again we find

$$\mathbf{T} \cong \mathbf{T}[p_0^{e_0}] \times \cdots \times \mathbf{T}[p_{n-1}^{e_{n-1}}].$$

Modules of the form $\mathbf{T}[p^e]$ where $p \in R$ is prime and e is a positive integer are said to be **primary** or sometimes more specifically p**-primary**. These are just the torsion modules of exponent a power of p. So now we know that every finitely generated torsion module over a principal ideal domain can be decomposed as a product of primary modules. We state this as a theorem.

The Primary Decomposition Theorem.
Every finitely generated module over a principal ideal domain is isomorphic to a direct products of finitely many primary modules.

Still, these primary modules may admit further direct decomposition.
Which p-primary modules are directly indecomposable?

Fact. Let $\mathbf{R}$ be a principal ideal domain and $p \in R$ be prime. Every cyclic p-primary $\mathbf{R}$-module is directly indecomposable.

Proof. Let $\mathbf{T}$ be an $\mathbf{R}$-module of exponent p^e that is generated by w. Suppose that $\mathbf{T} \cong \mathbf{M} \times \mathbf{N}$. We need to argue that one of $\mathbf{M}$ and $\mathbf{N}$ is trivial. We know that $p^e w = 0$ but $p^{e-1} w \neq 0$. Pick $u \in M$ and $v \in N$ so that (u, v) generates $\mathbf{M} \times \mathbf{N}$. Then we have $p^e u = 0$ and $p^e v = 0$ and either $p^{e-1} u \neq 0$ or $p^{e-1} v \neq 0$. Without loss of generality, suppose $p^{e-1} u \neq 0$. This makes $\mathbf{M}$ nontrivial, so our ambition is to show that $\mathbf{N}$ is trivial. Now since the element (u, v) generates all elements of $\mathbf{M} \times \mathbf{N}$, we see that every element of $M \times N$ can be obtained by multiplying (u, v) by an appropriate scalar. So pick $r \in R$ so that $r(u, v) = (0, v)$. It follows that $ru = 0$ and $rv = v$. Since the order of u is p^e, we have that $p^e \mid r$. So $r = sp^e$ for some s. But then $v = rv = sp^e v = 0$. But this entails that $\mathbf{N}$ is trivial. □

Of course, we can get a cyclic submodule easily enough just by selecting a single element and using it to generate a submodule. Something more clever is possible.

Fact. Let $\mathbf{R}$ be a principal ideal domain and let $\mathbf{T}$ be a torsion $\mathbf{R}$-module of exponent r. Then $\mathbf{T}$ has an element of order r.

Proof. Pick distinct primes $p_0, \dots, p_{n-1} \in R$ and positive integers $e_0, \dots, e_{n-1}$ so that

$$r = p_0^{e_0} \dots p_{n-1}^{e_{n-1}}.$$

For each $j < n$, let $r_j = \frac{r}{p_j} = p_0^{e_0} \dots p_{j-1}^{e_{j-1}} p_j^{e_j - 1} p_{j+1}^{e_{j+1}} \dots p_{n-1}^{e_{n-1}}$. Notice that $r \nmid r_j$ for all $j < n$. This allows us, for each $j < n$, to pick $u_j \in T$ so that $r_j u_j \neq 0$. Now, for each $j < n$, put

$$v_j = \frac{r}{p_j^{e_j}} u_j.$$

Then, for each $j < n$, we have $p_j^{e_j} v_j = 0$ but $p_j^{e_j - 1} v_j \neq 0$. So v_j is an element of order $p_j^{e_j}$. It now follows that $v_0 + \dots + v_{n-1}$ is an element of order r, as desired. □

There is one additional fact that proves useful.

Fact. Let $\mathbf{T}$ be a torsion $\mathbf{R}$-module of exponent r, where $\mathbf{R}$ is a principal ideal domain. Let $\mathbf{M}$ and $\mathbf{N}$ be submodules of $\mathbf{T}$, where $\mathbf{M}$ is generated by an element of order r. Let f be a homomorphism from $\mathbf{N}$ into $\mathbf{M}$. Finally, let $v \in T$ with $v \notin N$. Then f can be extended to a homomorphism from the submodule generated by $N \cup \{v\}$ into $\mathbf{M}$.

Proof. Let u be the element of order r that generates $\mathbf{M}$. Let $\mathbf{N}'$ be the submodule generated by $N \cup \{v\}$. Evidently, $rv = 0 \in N$, so v/N has some nonzero order s in $\mathbf{N}'/N$ and $s \mid r$. So $sv \in N$. Pick $p \in R$ so that

$$r = ps.$$

Now $f(sv) \in M$ and $pf(sv) = f(psv) = f(rv) = f(0) = 0$. So the order of $f(sv)$ divides p.

Since $f(sv) \in M$ and $\mathbf{M}$ is generated by u pick $q \in R$ so that $f(sv) = qu$. Then we see $0 = pf(sv) = pqu$. This means

$$r \mid pq.$$

Hence, $ps \mid pq$. Since **R** is an integral domain, we have

$$s \mid q.$$

So pick $t \in R$ such that

$$q = st.$$

This entails $f(sv) = qu = stu$. Let $w = tu \in M$. So $f(sv) = sw$.

Now every element of N' can be written in the form $y + av$ for some choices of $y \in N$ and $a \in R$. There may be several ways to make these choices.

Here is how we define our extension f' of f:

$$f'(y + av) = f(y) + aw.$$

It is not clear that this definition is sound. Let us verify that. Suppose that $y + av = y' + a'v$ where $y, y' \in N$ and $a, a' \in R$. We need to see that

$$f(y) + aw = f(y') + a'w \text{ or written another way } f(y) - f(y') = (a' - a)w.$$

But notice that $y - y' = (a' - a)v$. So $(a' - a)v \in N$. This means $s \mid (a' - a)$. Pick $m \in R$ so that $a' - a = ms$. But this leads to

$$f(y) - f(y') = f(y - y') = f((a' - a)v) = f(msv) = mf(sv) = msw = (a' - a)w,$$

just as we desire. So our definition is sound. Since $f'(y) = f'(y + 0v) = f(y) + 0w = f(y)$, for all $y \in N$, we see that f' extends f. We must also check that f' is a homomorphism, a task we leave to hard working graduate students. □

So our scheme is to grab an element whose order is the exponent of **T**, let **M** be the submodule generated by that element, and hope to find another submodule **N** so that $\mathbf{T} \cong \mathbf{M} \times \mathbf{N}$. If we are lucky maybe the exponent of **N** will be smaller. Here is what we need.

Fact. Let **R** be a principal ideal domain and let **T** be a torsion **R**-module of exponent r. Then **T** has a cyclic submodule **M** of exponent r and a submodule **N** of exponent s so that $s \mid r$ and $\mathbf{T} \cong \mathbf{M} \times \mathbf{N}$.

Proof. Let $u \in T$ have order r and let **M** be the submodule of **T** generated by u. Let f be the identity map on **M**. Let

$$\mathcal{F} = \{g \mid g : \mathbf{N} \to \mathbf{M} \text{ is a homomorphism extending } f \text{ for some submodule } \mathbf{N} \text{ with } M \subseteq N \subseteq T\}.$$

We will apply Zorn's Lemma to $\mathcal{F}$, which is partially ordered by set-incluion. Notice that $f \in \mathcal{F}$, so $\mathcal{F}$ is not empty. Let $\mathcal{C}$ be a nonempty chain in $\mathcal{F}$. Certainly $\bigcup \mathcal{C}$ is an upper bound on $\mathcal{C}$. We must show it is in $\mathcal{F}$. We have noted before that the union of a chain of functions is again a function. The hard-working graduate students will see that the union of a chain of homomorphisms is itself a homomorphism. It is routine to see that the union of a chain of submodules (the domains of those homomorphisms) is again a submodule. So we see indeed that $\bigcup \mathcal{C}$ belongs to $\mathcal{F}$. Let g be a maximal element of $\mathcal{F}$. Since the fact just above would otherwise allow the extension of g to larger member of $\mathcal{F}$, a thing impossible by the maximality of g, we see that g is a homomorphism from **T** into **M** which extends the identity map on **M**. Let **N** be the kernel of g. Let s be the exponent of **N**. Evidently, $s \mid r$.

For any $w \in T$ we have $g(w) \in M$. But then $g(g(w)) = f(g(w))$ since g extends f. But f is the identity map. So we see that $g(g(w)) = g(w)$ for all $w \in T$. But notice

$$g(w - g(w)) = g(w) - g(g(w)) = g(w) - g(w) = 0.$$

This means that $w - g(w) \in N$ for all $w \in T$, since N is the kernel of g. So we find that

$$w = g(w) + (w - g(w)) \in M + N, \text{ for all } w \in T.$$

Another way to write this is

$$T = M + N.$$

Now suppose $w \in M \cap N$. Then, on the one hand, $g(w) = f(w) = w$ since $w \in M$ and g extends f (which is the identity function), while on the other hand $g(w) = 0$ since $w \in N = \ker g$. Taken together we find that $w = 0$. This means

$$M \cap N = \{0\}.$$

So we see that $\mathbf{T} = \mathbf{M} \oplus \mathbf{N}$. This yields our desired conclusion. □

The Invariant Factor Theorem.
Let $\mathbf{T}$ *be a nontrivial finitely generated torsion* $\mathbf{R}$*-module, where* $\mathbf{R}$ *is a principal ideal domain. Then for some natural number* n *there are* $r_0, r_1, \dots, r_n \in R$ *with*

$$r_n \mid r_{n-1} \mid \cdots \mid r_1 \mid r_0$$

and cyclic submodules $\mathbf{M}_0$ *of exponent* $r_0, \dots, \mathbf{M}_n$ *of exponent* r_n *so that*

$$\mathbf{T} \cong \mathbf{M}_0 \times \cdots \times \mathbf{M}_n.$$

Proof. Let the order of $\mathbf{T}$ be r_0. (Recall that we have already proven that a nontrivial finitely generated torsion module has an exponent.) Let $u_0 \in T$ have order r_0 and let $\mathbf{M}_0$ be the submodule generated by u_0. By the preceding Fact, there is a submodule $\mathbf{N}_0$ of order r_1 with $r_1 \mid r_0$ such that $\mathbf{T} \cong \mathbf{M}_0 \times \mathbf{N}_0$. If $\mathbf{N}_0$ is trivial, we can stop since $\mathbf{T} \cong \mathbf{M}_0$ in that case. Otherwise, pick $u_1 \in N_0$ with order r_1. Take $\mathbf{M}_1$ to be the submodule of $\mathbf{N}_0$ generated by u_1 and invoke the immediately preceding fact to get a proper submodule $\mathbf{N}_1$ of $\mathbf{N}_0$ of exponent r_2 so that $r_2 \mid r_1$ and $\mathbf{N}_0 \cong \mathbf{M}_1 \times \mathbf{N}_1$. At this stage we have

$$\mathbf{T} \cong \mathbf{M}_0 \times \mathbf{M}_1 \times \mathbf{N}_1 \text{ and } r_2 \mid r_1 \mid r_0 \text{ and } N_1 \subsetneq N_0,$$

where the exponent of $\mathbf{M}_0$ is r_0, the exponent of $\mathbf{M}_1$ is r_1, and the exponent of $\mathbf{N}_1$ is r_2. Again our process terminates in the event $\mathbf{N}_1$ is trivial, but otherwise the process can be continued. In this process two chains of submodules of $\mathbf{T}$ are constructed. One is the descending chain consisting of the submodules $\mathbf{N}_j$. The other is the ascending chain

$$\mathbf{M}_0 \subsetneq \mathbf{M}_0 \oplus \mathbf{M}_1 \subsetneq \dots.$$

Now we know, as a corollary of the Freedom Theorem that every submodule of $\mathbf{T}$ is finitely generated. We saw for rings that if every ideal was finitely generated then there could be no infinite ascending chains of ideals. The same reasoning applies here (as the hard working graduate student will establish) to see that there can be no infinite ascending chain of submodules of $\mathbf{T}$. This must mean the process described above terminates at some finite stage. This completes our proof. □

The $r_0, r_1, \dots, r_n$ mentioned in the theorem are called **invariant factors**.

Now we are ready for the chief existence theorem for direct decomposition of finitely generated modules.

The Elementary Divisor Theorem.
Let $\mathbf{T}$ *be a nontrivial finitely generated torsion* $\mathbf{R}$*-module, where* $\mathbf{R}$ *is a principal ideal domain. Then for some natural number* n*, there are cyclic primary submodules* $\mathbf{M}_0, \dots, \mathbf{M}_n$ *of* $\mathbf{T}$ *so that*

$$\mathbf{T} \cong \mathbf{M}_0 \times \cdots \times \mathbf{M}_n.$$

The exponents of the various cyclic primary submodules are referred to as the **elementary divisors** of $\mathbf{T}$.

The proof of the Elementary Divisor Theorem is obtained by applying the Invariant Factor Theorem to each of the direct factors arising from an application of the Primary Decomposition Theorem.

9.4 Problem Set 9

Algebra Homework, Edition 9
Tenth Week
A Grab Bag

Problem 41.
Let A be the 4×4 real matrix

$$A = \begin{pmatrix} 1 & 1 & 0 & 0 \\ -1 & -1 & 0 & 0 \\ -2 & -2 & 2 & 1 \\ 1 & 1 & -1 & 0 \end{pmatrix}$$

(a) Determine the rational canonical form of A.

(b) Determine the Jordan canonical form of A.

Problem 42.
Suppose that N is a 4×4 nilpotent matrix over a field $\mathbf{F}$ with minimal polynomial x^2. What are the possible rational canonical forms for N?

Problem 43.
Let $\mathbf{F}$ be the subring of the field of complex numbers consisting of those numbers of the form $a + ib$ where a and b are rationals. Let $\mathbb{G}$ be the subring of the field of complex numbers consisting of those numbers of the form $m + ni$ where m and n are integers.

(a) Describe all the units of $\mathbb{G}$.

(b) Prove that $\mathbf{F}$ is (isomorphic to) the field of fractions of $\mathbb{G}$.

(c) Prove that $\mathbb{G}$ is a principal ideal domain.

[Hint: In this problem it is helpful to consider the function that sends each complex number z to $z\bar{z} = |z|^2$.]

LECTURE 10

THE STRUCTURE OF FINITELY GENERATED MODULES OVER A PID

Here is one of the key results in our course.

The Structure Theorem for Finitely Generated Modules over a Principal Ideal Domain. *Let* $\mathbf{M}$ *be a finitely generated* $\mathbf{R}$*-module, where* $\mathbf{R}$ *is a principal ideal domain. There is a natural number n such that:*

(a) *there are n finitely generated directly indecomposable submodules* $\mathbf{M}_0,\ldots,\mathbf{M}_{n-1}$ *of* $\mathbf{M}$ *such that*

$$\mathbf{M} \cong \mathbf{M}_0 \times \cdots \times \mathbf{M}_{n-1}, \text{ and}$$

(b) *for any natural number m, if* $\mathbf{N}_0,\ldots,\mathbf{N}_{m-1}$ *are directly indecomposable* $\mathbf{R}$*-modules such that*

$$\mathbf{M} \cong \mathbf{N}_0 \times \cdots \times \mathbf{N}_{m-1},$$

then $n = m$ *and there is a permutation* σ *of* $\{0,\ldots,n-1\}$ *so that* $\mathbf{M}_i \cong \mathbf{N}_{\sigma(i)}$ *for all* $i < n$.

Moreover, the finitely generated directly indecomposable $\mathbf{R}$*-modules are, up to isomorphism, the* $\mathbf{R}$*-module* $\mathbf{R}$ *(that is the free* $\mathbf{R}$*-module of rank* 1*), and the* $\mathbf{R}$*-modules of the form* $\mathbf{R}/(r)$ *where* r *is a positive power of some prime element of the ring* $\mathbf{R}$ *(these are the cyclic primary* $\mathbf{R}$*-modules). Finally, the free* $\mathbf{R}$*-module of rank* 1 *is not primary and if* $r, s \in R$ *are prime powers and* $\mathbf{R}/(r) \cong \mathbf{R}/(s)$*, then* r *and* s *are associates in the ring* $\mathbf{R}$.

Before turning to the proof a few remarks are in order.

First, we have allowed $n = 0$. This results in the direct product of an empty system of $\mathbf{R}$-modules. A careful, but easy, examination of the definition of direct products reveals that such a direct product produces the trivial $\mathbf{R}$-module—that is the module whose only element is 0. Evidently, the trivial $\mathbf{R}$-module is the direct product of exactly one system of directly indecomposable $\mathbf{R}$-modules, namely of the empty system.

This theorem has three parts:

- the assertion of the existence of a decomposition into indecomposables,
- the assertion that such a decomposition is unique, and
- a description of the indecomposables.

These are the hallmarks of a good structure theorem. There are other theorems of this kind in mathematics. Perhaps the most familiar is the Fundamental Theorem of Arithmetic. To make the connection plain, consider the least complicated algebraic systems, namely just nonempty sets equipped with no operations. Then the finitely generated algebraic systems are just the finite nonempty sets and isomorphisms are just one-to-one correspondences. So two of these algebraic systems will be isomorphic if and only if they have the same number of elements. The Fundamental Theorem of Arithmetic says that every finite set is isomorphic to a direct product of directly indecomposable finite sets in a way that is unique up to isomorphism and rearranging the factorization, and that a set is directly indecomposable if and only if it has a prime number of elements.

This theorem also resonates with the notion of a unique factorization domain. We could reformulate our structure theorem to make this more apparent. Each finitely generated **R**-module is isomorphic to a lot of other **R**-modules (in fact, to a proper class of **R**-modules). Pick a representative from each of these isomorphism classes, taking care to include among these representatives the **R**-modules **R** and $\mathbf{R}/(r)$ where $r \in R$ is a positive power of some prime element of **R**. Let $\mathcal{M}$ be the set of all these representatives and let **1** be the representative of the trivial **R**-module. Then $\langle \mathcal{M}, \times, \mathbf{1}\rangle$ is an algebraic system (actually a monoid) with the unique factorization property.

Finally, the structure theorem above has a number of far-reaching consequences. Taking **R** to be $\mathbb{Z}$ we obtain a structure theorem for finitely generated Abelian group. Taking **R** to be $\mathbf{F}[x]$, where **F** is a field leads to the canonical form theorems of linear algebra.

Proof. Let us first dispose of the descriptions of the directly indecomposable **R**-modules that could arise in any factorization of **M**. These must be finitely generated because they will be isomorphic to submodules of **M** and, according to the Corollary of the Freedom Theorem every submodule of **M** must be finitely generated. We have already seen that the free **R**-module of rank 1 (namely the module **R**) is directly indecomposable and every other free **R**-module of finite rank $n > 1$ is a direct product of n-copies of **R**. We have also seen that the cyclic primary **R**-modules are the only finitely generated directly indecomposable torsion modules. Can there be any other finitely generated directly indecomposable **R**-module? By the First Decomposition Theorem every finitely generated **R**-module is isomorphic to a direct product of the form $\mathbf{T} \times \mathbf{F}$, where **T** is the torsion submodule and **F** is a submodule that is free. For a directly indecomposable module we must have either **F** trivial (and then our module would be torsion) or else **T** trivial (and then our module would be free). So the only finitely generated directly indecomposable **R**-modules are the ones already in hand, the **R**-module **R** and the cyclic primary **R**-modules.

We can say more about the cyclic primary **R**-modules. Let $r \in R$ be a positive power of a prime element of R. Then the ideal (r) is a submodule of the **R**-module **R**. The element $1/(r)$ of the quotient module $\mathbf{R}/(r)$ generates the quotient module and has order r. So the quotient module is cyclic and of exponent r. In this way we know cyclic **R**-modules of exponent r exist. Suppose that **N** is an **R**-module of exponent r which is generated by the single element u. Since $\{1\}$ is a basis for **R**, we know there is a homomorphism h from **R** onto **N** that takes 1 to u. Now for all $s \in R$ we have $h(s) = h(s \cdot 1) = sh(1) = su$. From this we see that

$$s \in \ker h \Leftrightarrow h(s) = 0 \Leftrightarrow su = 0 \Leftrightarrow r \mid s \Leftrightarrow s \in (r).$$

That is, $\ker h = (r)$. So by the Homomorphism Theorem $\mathbf{N} \cong \mathbf{R}/(r)$. So, up to isomorphism, the only cyclic **R**-module of exponent r (where r is a positive power of some prime) is $\mathbf{R}/(r)$.

Now observe that the free **R**-module **R** of rank 1 is not a torsion module since $s \cdot 1 = 0 \implies s = 0$. So the **R**-module **R** cannot be isomorphic with any of the modules $\mathbf{R}/(r)$ where r is a positive power of some prime. (One needs to observe here, as the hard-working graduate students will verify, that r cannot be a unit.) Now suppose that r and s are both positive powers of primes (we don't assume the primes are the same) and that $\mathbf{R}/(r) \cong \mathbf{R}/(s)$. Then r is the order of a generator of this cyclic module and so is s. This means that $r \mid s$ and $s \mid r$. Consequently, $(r) = (s)$ and r and s are associates.

Now consider part (a) of the theorem. This is an immediate consequence of the First Decomposition Theorem and the Elementary Divisor Theorem.

Finally, consider part (b). Some of the $\mathbf{N}_i$'s can be cyclic primary modules and others can be free of rank 1, according to our description of the directly indecomposable finitely generated modules. Without loss of generality, we assume that the primary modules come first. So pick $k \le m$ so that $\mathbf{N}_0, \ldots, \mathbf{N}_{k-1}$ are cyclic primary modules and $\mathbf{N}_k, \ldots, \mathbf{N}_{m-1}$ are free of rank 1. Let $\mathbf{T}$ be the direct product of the first group and $\mathbf{F}$ be the direct product of the second. So we find $\mathbf{M} \cong \mathbf{T} \times \mathbf{F}$. It is routine (according to hard-working graduate students) that $\mathbf{T}$ is a torsion module and also that $\mathbf{F}$ is free of rank $m-k$. Let $(v,u) \in T \times F$ be a torsion element of order r. Then $(0,0) = r(v,u) = (rv, ru)$. In particular, $ru = 0$. The element u can be written as a linear combination of the basis elements of $\mathbf{F}$. By distributing r through the linear combination and invoking both linear independence and the fact the r is a nonzero element of an integral domain, we see that $u = 0$. What we conclude is that the torsion elements of $\mathbf{T} \times \mathbf{F}$ are exactly those of the form $(v,0)$ where $v \in T$ is nonzero. Thus under the isomorphism $\mathbf{M} \cong \mathbf{T} \times \mathbf{F}$, the module $\mathbf{T}$ corresponds to the torsion submodule of $\mathbf{M}$. Then according to the First Decomposition Theorem the rank of $\mathbf{F}$ is determined by $\mathbf{M}$.

It remains to show that if $\mathbf{T} \cong \mathbf{M}_0 \times \cdots \times \mathbf{M}_{\ell-1} \cong \mathbf{N}_0 \times \cdots \times \mathbf{N}_{k-1}$, where all the $\mathbf{M}_i$'s and $\mathbf{N}_j$'s are cyclic primary $\mathbf{R}$-modules, then $\ell = k$ and, after a suitable reindexing, $\mathbf{M}_i \cong \mathbf{N}_i$ for all $i < \ell$.

Let $p \in R$ be prime. For any $\mathbf{R}$-module $\mathbf{Q}$ let

$$Q(p) = \{v \mid v \in Q \text{ and } p^e v = 0 \text{ for some positive integer } e\}.$$

It is routine to check that this set is closed under the module operations, so we have the submodule $\mathbf{Q}(p)$. It is also not hard to see (as hard-working graduate students will check) that

$$\mathbf{T}(p) \cong \mathbf{M}_0(p) \times \cdots \times \mathbf{M}_{\ell-1}(p) \cong \mathbf{N}_0(p) \times \cdots \times \mathbf{N}_{k-1}(p).$$

In this decomposition, if $\mathbf{M}_i$ (or $\mathbf{N}_j$) were primary with respect to a prime not associate to p, then the module $\mathbf{M}_i(p)$ (respectively $\mathbf{N}_j(p)$) would be trivial. On the other hand, if they were primary with respect to an associate of p, then $\mathbf{M}_i(p) = \mathbf{M}_i$ and $\mathbf{N}_j(p) = \mathbf{N}_j$. Since this holds for arbitrary primes p, we do not lose any generality by assuming the the primes underlying all the $\mathbf{M}_i$'s and $\mathbf{N}_j$'s are the same prime p.

Now suppose $\mathbf{Q}$ is a cyclic primary $\mathbf{R}$-module, where p^e is the exponent and u is a generator. Then $\mathbf{Q}[p]$ is generated by $p^{e-1}u$. So $\mathbf{Q}[p]$ is cyclic of exponent p. In this case, we know that $\mathbf{Q}[p] \cong \mathbf{R}/(p)$. Now (p) is a prime ideal of the ring $\mathbf{R}$. In a principal ideal domain, the maximal ideals and the prime ideals coincide. So the ring $\mathbf{R}/(p)$ is a field. This allows us to construe the $\mathbf{R}$-module $\mathbf{Q}[p]$ as a one-dimensional vector space over the field $\mathbf{R}/(p)$. In doing this, we are changing the scalar multiplication, but leaving the addition and the zero the same. Now we have

$$\mathbf{T}[p] \cong \mathbf{M}_0[p] \times \cdots \times \mathbf{M}_{\ell-1}[p] \cong \mathbf{N}_0[p] \times \cdots \times \mathbf{N}_{k-1}[p]$$

construed as vector spaces over the field $\mathbf{R}/(p)$, with each of the direct factors being a copy of the one-dimensional vector space. This means

$$\ell = \dim \mathbf{T}[p] = k.$$

So we have discovered that $\ell = k$, one of our desired conclusions.

So we are reduced to considering the following situation:

$$\begin{aligned} \mathbf{T} &\cong \mathbf{M}_0 \times \cdots \times \mathbf{M}_{\ell-1} \\ &\cong \mathbf{N}_0 \times \cdots \times \mathbf{N}_{\ell-1} \end{aligned}$$

where $\mathbf{M}_i$ is a cyclic module of exponent p^{e_i} and $\mathbf{N}_i$ is a cyclic module of exponent p^{f_i} for all $i < \ell$ and

$$\begin{aligned} e_0 &\ge e_1 \ge \cdots \ge e_{\ell-1} \\ f_0 &\ge f_1 \ge \cdots \ge f_{\ell-1}. \end{aligned}$$

It remains only to show that $e_i = f_i$ for all $i < \ell$. Suppose, for the sake of contradiction, that this were not so. Let i be as small as possible so that $e_i \neq f_i$. It is harmless to also suppose that $e_i > f_i$. Let $r = p^{f_i}$. Now multiplication by r is a homomorphism and it is easy to also see that

$$\begin{aligned} r\mathbf{T} &\cong r\mathbf{M}_0 \times \cdots \times r\mathbf{M}_{i-1} \times r\mathbf{M}_i \times \cdots \times r\mathbf{M}_{\ell-1} \\ &\cong r\mathbf{N}_0 \times \cdots \times r\mathbf{N}_{i-1} \times r\mathbf{N}_i \times \cdots \times r\mathbf{N}_{\ell-1}. \end{aligned}$$

Being homomorphic images of cyclic modules, each of the direct factors above is also cyclic. Because r is a positive power of the prime p, we see that the factor modules above are either primary (with prime p) or trivial. But exponents of all the $\mathbf{N}_j$'s where $i \leq j$ are factors of r, we see that these modules are all trivial. On the other hand, the exponent of $\mathbf{M}_i$ is p^{e_i} whereas $r = p^{f_i}$ with $e_i > f_i$. So $r\mathbf{M}_i$ is not trivial. This would mean

$$\begin{aligned} r\mathbf{T} &\cong r\mathbf{M}_0 \times \cdots \times r\mathbf{M}_{i-1} \times r\mathbf{M}_i \times \cdots \times r\mathbf{M}_{\ell-1} \\ &\cong r\mathbf{N}_0 \times \cdots \times r\mathbf{N}_{i-1}, \end{aligned}$$

where the top direct factorization has at least $i+1$ nontrivial cyclic primary factors but the bottom has only i. But we have just proven that the number of such factors must be the same no matter how the direct factorization is accomplished. This contradiction means our supposition must be rejected. So $e_i = f_i$ for all $i < \ell$. This establishes the uniqueness of our direct factorization into directly indecomposable modules. The proof of the last remaining part of our theorem, namely part (b), is complete. □

The Structure Theorem above is an extension of the Elementary Divisor Theorem formulated in the previous lecture. We can also extend the Invariant Factor Theorem.

The Extended Invariant Factor Theorem.
Let $\mathbf{T}$ *be a nontrivial finitely generated torsion* $\mathbf{R}$*-module, where* $\mathbf{R}$ *is a principal ideal domain. Then for some natural number* n *there are* $r_0, r_1, \ldots, r_n \in R$ *with*

$$r_n \mid r_{n-1} \mid \cdots \mid r_1 \mid r_0$$

and cyclic submodules $\mathbf{M}_0$ *of exponent* $r_0, \ldots, \mathbf{M}_n$ *of exponent* r_n *so that*

$$\mathbf{T} \cong \mathbf{M}_0 \times \cdots \times \mathbf{M}_n.$$

Moreover, the natural number n *is uniquely determined by* $\mathbf{T}$*, the sequence* $r_n \mid r_{n-1} \mid \cdots \mid r_1 \mid r_0$ *is uniquely determined up to associates, and cyclic submodules* $\mathbf{M}_0, \ldots, \mathbf{M}_n$ *are determined up to isomorphism.*

Only the various aspects of uniqueness require proof at this point. However, these proofs follow the lines of the uniqueness portion of the proof of the Structure Theorem. We leave the details in the hands of the hard working graduate students. It is useful to note that the cyclic modules which are the factors in this direct decomposition may not themselves be directly indecomposable.

Using the Structure Theorem, for each principal ideal domain $\mathbf{R}$ we can define a function d such that $d(p^e, \mathbf{M})$ is the number of direct factors isomorphic to the module $\mathbf{R}/(p^e)$ in any direct factorization of $\mathbf{M}$ into directly indecomposable modules, where $p \in R$ is prime, e is a positive natural number, and $\mathbf{M}$ is a finitely generated $\mathbf{R}$-module. In addition, we take $d(0, \mathbf{M})$ to be the number of direct factors isomorphic to the $\mathbf{R}$-module $\mathbf{R}$ (that is, the directly indecomposable free module).

Then we have the useful

Corollary. *Let* $\mathbf{R}$ *be a principal ideal domain and* $\mathbf{M}$ *and* $\mathbf{N}$ *be finitely generated* $\mathbf{R}$*-modules. Then* $\mathbf{M} \cong \mathbf{N}$ *if and only if* $d(q, \mathbf{M}) = d(q, \mathbf{N})$ *for all* q *such that either* $q = 0$ *or* q *is a positive power of a prime in* $\mathbf{R}$.

What this corollary asserts is that the system

$$\langle d(q,\mathbf{M}) \mid q = 0 \text{ or } q \text{ is the positive power of a prime of } \mathbf{R}\rangle$$

of natural numers is a complete system of invariants of **M**—that is this system of natural numbers determines **M** up to isomorphism.

As noted earlier, modules over the ring $\mathbb{Z}$ of integers are essentially the same as Abelian groups since, for instance, $3u = (1+1+1)u = u+u+u$ and $-7v = -(v+v+v+v+v+v+v)$. In this way, we see that for a $\mathbb{Z}$-module $\mathbf{M} = \langle M,+,-,0,a\cdot\rangle_{a\in\mathbb{Z}}$ the scalar multiplication is expressible by means of the additive structure $+,-$, and 0. In particular, any map between $\mathbb{Z}$-modules that respects $+,-$, and 0 must also respect the scalar multiplication, any subset of a $\mathbb{Z}$-module that is closed under $+,1$, and contains 0 will also be closed under all the scalar multiplications, and a similar remark holds for direct products—in any of these constructions one may ignore the scalar multiplications along the way, but impose them on the result (the homomorphic image, the subalgebra, or the direct product) by means of repeated addition.

With this in mind, noting that $\mathbb{Z}$ is a principal ideal domain, we obtain

The Fundamental Theorem for Finitely Generated Abelian Groups.
Let **A** *be a finitely generated Abelian group. There is a natural number n such that:*

(a) *there are n finitely generated directly indecomposable subgroups* $\mathbf{A}_0,\ldots,\mathbf{A}_{n-1}$ *of* **A** *such that*

$$\mathbf{A} \cong \mathbf{A}_0 \times \cdots \times \mathbf{A}_{n-1}, \text{ and}$$

(b) *for any natural number m, if* $\mathbf{B}_0,\ldots,\mathbf{B}_{m-1}$ *are directly indecomposable Abelian groups such that*

$$\mathbf{A} \cong \mathbf{B}_0 \times \cdots \times \mathbf{B}_{m-1},$$

then $n = m$ *and there is permutation* σ *of* $\{0,\ldots,n-1\}$ *so that* $\mathbf{A}_i \cong \mathbf{B}_{\sigma(i)}$ *for all* $i < n$.

Moreover, the finitely generated directly indecomposable Abelian groups are, up to isomorphism, the group $\langle\mathbb{Z},+,-,0\rangle$ *of integers with respect to addition (that is the free Abelian group of rank* 1*), and the cyclic groups of prime power order (these are the groups* $\mathbb{Z}_q$ *where q is a positive power of a prime number, the set of elements is* $\{0,1,\ldots,q-1\}$*, and addition works modulo q). Finally, the free Abelian group of rank* 1 *is not of prime power order and if* $r,s \in R$ *are prime powers and* $\mathbb{Z}_r \cong \mathbb{Z}_s$*, then* $r = s$.

This could be regarded as the elementary divisor version of the structure theorem for finitely generated Abelian groups. One could as easily formulate a structure theorem from the invariant factor point of view. To see how these two points of view compare consider a description, up to isomorphism, of all the Abelian groups of order 100. The prime factorization gives $100 = 2^2 5^2$. Using the elementary divisor perspective we see that the list, representative up to isomorphism and also pairwise nonisomorphic, of Abelian groups of order 100 is

$$\mathbb{Z}_4 \times \mathbb{Z}_{25} \qquad \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_{25} \qquad \mathbb{Z}_4 \times \mathbb{Z}_5 \times \mathbb{Z}_5 \qquad \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_5 \times \mathbb{Z}_5$$

while from the invariant factor perspective the list is

$$\mathbb{Z}_{100} \qquad \mathbb{Z}_2 \times \mathbb{Z}_{50} \qquad \mathbb{Z}_5 \times \mathbb{Z}_{20} \qquad \mathbb{Z}_{10} \times \mathbb{Z}_{10}.$$

At work here are the following direct decompositions:

$$\mathbb{Z}_{100} \cong \mathbb{Z}_4 \times \mathbb{Z}_{25} \qquad \mathbb{Z}_{50} \cong \mathbb{Z}_2 \times \mathbb{Z}_{25} \qquad \mathbb{Z}_{20} \cong \mathbb{Z}_4 \times \mathbb{Z}_5 \qquad \mathbb{Z}_{10} \cong \mathbb{Z}_2 \times \mathbb{Z}_5.$$

10.1 Problem Set 10

Algebra Homework, Edition 10
Eleventh Week
Decomposing Modules

Problem 44.
Let **R** be a nontrivial integral domain and **M** be an **R**-module. Prove the set T of torsion elements is a submodule of **M** and **M**/T is torsion free.

Problem 45.
Let **R** be a principal ideal domain and let **T** be a torsion **R**-module of exponent r. Prove that **T** has an element of order r.

Problem 46.
Prove that the sequence of invariant factors (i.e. the sequence $r_0, r_1, \ldots, r_n$) mentioned in the Invariant Factor Theorem is uniquely determined by the module.

Problem 47.
Let **M** be a finitely generated **R**-module, where **R** is a principal ideal domain. Prove each of the following.

(a) The direct decomposition using the Invariant Factor Theorem is the one using the smallest number of factors that are all cyclic.

(b) The direct decomposition using the Elementary Divisor Theorem is the one using the largest number of factors that are all cyclic.

LECTURE 11

Decomposition of Vector Spaces with a Designated Linear Operator

Let $\mathbf{V}$ be a finite dimensional vector space over a field $\mathbf{F}$ and let T be a linear operator on $\mathbf{V}$—that is, T is an endomorphism of the vector space $\mathbf{V}$. Our objective here is to decompose $\mathbf{V}$ as a direct product of subspaces that are invariant with respect to T. The most straightforward way to proceed with such a project is to adjoin T to the vector space as a new one place operation. This new algebraic system would have a binary operation $+$ (the old vector addition), a designated element 0 (the zero vector), a one-place operation $-$ of forming negations, a one-place operation aI for each $a \in F$ (the scalar muliplications), and the new one-place operation T. The program would then become the direct decomposition of this new algebraic system into directly indecomposable factors. It is possible to carry out this program, to prove the corresponding structure theorem (which would prove the existence and uniqueness of such decompositions and describe the directly indecomposable algebras, much as in the last section).

However, there is an alternate route to the same result that allows us to take advantage of the work we have done with modules. The idea is to regard $\mathbf{V}$ as a module over the principal ideal domain $\mathbf{F}[x]$ instead of over the field $\mathbf{F}$. This means we have to define what $f(x) \cdot v$ means for every vector $v \in V$ and every polynomial $f(x) \in \mathbf{F}[x]$. Here is the definition:

$$f(x) \cdot v := f(T)(v).$$

Here $f(T) = a_0 I + a_1 T + a_2 T^2 + \cdots + a_n T^n$ where $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ and $T^2 = T \circ T$, $T^3 = T \circ T \circ T$, and so on. It is easy to see that each $f(T)$ is a linear operator (that is, an endomorphism of the vector space $\mathbf{V}$). The polynomials of degree 0 provide the ordinary scalar multiplications of the vector space. So construing $\mathbf{V}$ as a module over $\mathbf{F}[x]$ in effect adjoins many more one-place operations than our first approach, but they are all built up from the $a \cdot I$ and T by addition and composition. This is why the two approaches are equivalent.

Recall from linear algebra that the linear operators on a finite-dimensional vector space $\mathbf{V}$ constitute a finite dimensional vector space themselves. So for any linear operator T the set $\{I, T, T^2, T^3, \dots\}$ is linearly dependent. This means that for some natural number m there are $a_0, a_1, \dots, a_m \in F$ with $a_m \neq 0$ so that $a_0 I + a_1 T + \cdots + a_m T^m = 0$. In other words, there is a nonzero polynomial $f(x) \in \mathbf{F}[x]$ so that $f(T)$ is the zero linear operator (the map taking all vectors to the zero vector). Evidently, $\{f(x) \mid f(T)$ is the zero operator$\}$ is an ideal of $\mathbf{F}[x]$. Since $\mathbf{F}[x]$ is a principal ideal domain first ideal is generated by a single polynomial.

In fact we can take this polynomial to be the monic polynomial $m_T(x)$ of least degree in this ideal. This polynomial is called the **minimal polynomial** of T.

Now fix a linear operator T on the finite dimensional vector space $\mathbf{V}$. We use $\mathbf{V}_T$ to denote the module over $\mathbf{F}[x]$ described above. We know that this module can be decomposed into a direct product of cyclic submodules. What do these cyclic submodules look like? Well, suppose that $v \in V$ is a generator. Then the submodule consists of all the vectors of the form $f(T)(v)$ as $f(x)$ runs through the ring of polynomials. Hence the linear span of the set $\{v, Tv, T^2v, \ldots\}$ is the whole submodule. Since this submodule is, among other things, a subspace of $\mathbf{V}$ (with additional operations), we know that some finite subset must span the submodule. Let m be as small as possible so that $\{v, Tv, \ldots, T^m v\}$ spans the submodule. Then there are $a_0, \ldots, a_m \in F$ so that

$$T^{m+1}v = a_0 v + \cdots + a_m T^m v.$$

This leads to

$$T^{m+2}v = a_0 Tv + \cdots + a_m T^{m+1} v = a_0 Tv + \cdots + a_m(a_0 v + \cdots + a_m T^m v).$$

In this way we see that m is also the smallest natural number so that $T^{m+1}v$ is a linear combination of $\{v, Tv, \ldots, T^m v\}$. I contend that this set is linearly independent. Suppose

$$b_0 v + b_1 Tv + \cdots + b_m T^m v = 0.$$

Now b_m must be 0, otherwise $T^m v = -\frac{b_0}{b_m} v - \cdots - \frac{b_{m-1}}{b_m} T^{m-1} v$. Once the term $b_m T^m v$ has been eliminated (because it is 0), we can apply the same reasoning to see that $b_{m-1} = 0$, and then that $b_{m-2} = 0$, and so on. In this way we establish the linear independence of $\{v, Tv, \ldots, T^m v\}$. Thus we see that our cyclic submodule, construed as an ordinary vector space over the field $\mathbf{F}$ has a very nice basis. We call this kind of basis a T**-cyclic** basis.

Here is what happens if we represent T with respect to this basis. Put

$$v_0 = v, v_1 = Tv, \ldots, v_m = T^m v.$$

Then

$$\begin{aligned}
Tv_0 &= v_1 = 0v_0 + 1v_1 + 0v_2 + \cdots + 0v_m \\
Tv_1 &= v_2 = 0v_0 + 0v_1 + 1v_2 + \cdots + 0v_m \\
&\vdots \\
Tv_{m-1} &= v_m = 0v_0 + 0v_1 + 0v_2 + \cdots + 1v_m \\
Tv_m &= a_0 v_0 + a_1 v_1 + \cdots + a_m v_m
\end{aligned}$$

This produces the matrix

$$\begin{pmatrix}
0 & 0 & 0 & \ldots & 0 & a_0 \\
1 & 0 & 0 & \ldots & 0 & a_1 \\
0 & 1 & 0 & \ldots & 0 & a_2 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & \ldots & 0 & a_{m-1} \\
0 & 0 & 0 & \ldots & 1 & a_m
\end{pmatrix}$$

This is a very pleasant matrix with lots of entries 0 and the nonzero entries located in rather restrained positions. Let us rewrite that last equation:

$$\begin{aligned}
Tv_m &= a_0v_0 + a_1v_1 + \cdots + a_mv_m \\
T^{m+1}v &= a_0v + a_1Tv + \cdots + a_mT^mv \\
0 &= -T^{m+1}v + a_mT^mv + \cdots + a_1Tv + a_0v \\
0 &= T^{m+1}v - a_mT^mv - \cdots - a_1Tv - a_0v \\
0 &= (T^{m+1} - a_mT^m - \cdots - a_1T - a_0)v \\
0 &= m_T(T)v
\end{aligned}$$

where $m_T(x) = x^{m+1} - a_mx^m - \cdots - a_0 \in \mathbf{F}[x]$. Notice that this is a monic polynomial of least degree which belongs to the annihilator of $\mathbf{V}_T$. So it is an exponent of the module $\mathbf{V}_T$.

We could start with any monic polynomial $f(x) = b_0 + b_1x + \cdots + b_mx^m + x^{m+1}$ of positive degree. If this polynomial happened to be the minimal polynomial of some linear operator T so that $\mathbf{V}_T$ was a cyclic module, then the associated matrix, as above, would be

$$C_f = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 & -b_0 \\ 1 & 0 & 0 & \dots & 0 & -b_1 \\ 0 & 1 & 0 & \dots & 0 & -b_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & -b_{m-1} \\ 0 & 0 & 0 & \dots & 1 & -b_m \end{pmatrix}.$$

This matrix is called the **companion matrix** of the polynomial $f(x)$. Observe that this is a $(m+1) \times (m+1)$ matrix, where $m+1$ is the degree of $f(x)$. It is easy to write down the companion matrix given the monic polynomial and, vice versa, given the companion matrix to write down the monic polynomial. A routine calculation also reveals that for any monic polynomial $f(x)$ we have

$$f(x) = \det(xI - C_f).$$

This means that $f(x)$ is the **characteristic polynomial** of its companion matrix.

We summarize these findings in the following Fact.

Fact. Let $\mathbf{V}$ be a finite dimensional vector space over a field $\mathbf{F}$ and T be a linear operator on $\mathbf{V}$ such that $\mathbf{V}_T$ is a cyclic module over $\mathbf{F}[x]$ with generator v and with minimal polynomial $m_T(x)$ of degree $n+1$. Then $\{v, Tv, T^2v, \dots, T^nv\}$ is a basis for $\mathbf{V}_T$ and, with respect to this basis, the matrix of T is the companion matrix of $m_T(x)$ and the characteristic polynomial of T is the same as the minimal polynomial of T.

Now let's apply the Invariant Factor Theorem:

The Rational Canonical Form Theorem.

Let $\mathbf{V}$ be a finite dimensional vector space over the field $\mathbf{F}$. Let T be a linear operator of $\mathbf{V}$. Then for some natural number n there are monic polynomials $f_0(x), f_1(x), \dots, f_n(x) \in \mathbf{F}[x]$ with $f_0(x) = m_T(x)$, the minimal polynomial of T, such that

$$f_n(x) \mid f_{n-1}(x) \mid \cdots \mid f_1(x) \mid f_0(x)$$

and cyclic submodules of $\mathbf{V}_T$ (sometimes called T-cyclic subspaces of $\mathbf{V}$) $\mathbf{V}_0$ of exponent $f_0(x)$,..., $\mathbf{V}_n$ of exponent $f_n(x)$ so that

$$\mathbf{V}_T \cong \mathbf{V}_0 \times \cdots \times \mathbf{V}_n.$$

Moreover, the natural number n is uniquely determined by T, the sequence $f_n(x) \mid f_{n-1}(x) \mid \cdots \mid f_1(x) \mid f_0(x)$ of monic polynomials is uniquely determined, and cyclic submodules $\mathbf{V}_0, \ldots, \mathbf{V}_n$ are determined up to isomorphism. Furthermore, each of the submodules $\mathbf{V}_k$ for $k \le n$ has a T-cyclic basis B_k, and $B_n \cup \cdots \cup B_0$ is a basis for $\mathbf{V}$. The linear operator T is represented with respect to this basis by

$$\begin{pmatrix} C_{f_n} & 0 & 0 & \ldots & 0 \\ 0 & C_{f_{n-1}} & 0 & \ldots & 0 \\ 0 & 0 & C_{f_{n-2}} & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & C_{f_0} \end{pmatrix}$$

where the companion matrices of the $f_k(x)$'s are placed in order as square blocks along the diagonal, with all remaining entries of the matrix 0.

The matrix representing T in this theorem is said to be in **rational canonical form**. The uniqueness assertions of the Extended Invariant Factor Theorem ensure that a linear operator has exactly one rational canonical form. The following important theorem is now a corollary.

The Cayley-Hamilton Theorem.
Let T be a linear operator on a finite dimensional vector space over a field. The minimal polynomial of T divides the characteristic polynomial of T. Hence, $f(T)$ is the constantly zero linear operator, when $f(x)$ is the characteristic polynomial of T.

Actually, it is easy to see that the characteristic polynomial is just the product of the invariant factors.

Recall from linear algebra that if A and B are $m \times m$ matrices with entries in the field $\mathbf{F}$, then we say that A and B are **similar** provided there is a linear operator T on the m-dimensional vector space over $\mathbf{F}$ such that T can be represented by both A and B (using appropriated bases). So we find

Rational Canonical Form Theorem: Matrix Version.
Let $\mathbf{F}$ be a field and m be a positive natural number. Every $m \times m$ matrix with entries from $\mathbf{F}$ is similar to exactly one matrix in rational canonical form.

Now let's turn to the elementary divisor perspective. Consider the case when $\mathbf{V}_T$ is a cyclic primary $\mathbf{F}[x]$ module. In this case, there is a vector $v \in V$, an irreducible monic polynomial $f(x) \in \mathbf{F}[x]$, and a positive natural number e so that $(f(x))^e$ is the order of v. As long as the field $\mathbf{F}$ is arbitrary, the polynomial $f(x)$ could be quite complicated—for instance it might have arbitrarily large degree. In such a situation, it would be difficult to improve on the process we used above to obtain the rational canonical form. However, two fields immediately come to mind where the situation is much more restrained. The field $\mathbb{C}$ of complex numbers has the property that all irreducible polynomials in $\mathbb{C}[x]$ have degree 1, while over the field $\mathbb{R}$ of real numbers there can also be irreducible polynomials of degree 2 but of no higher degrees. Both of these facts will be proved in the next semester. So let us consider that $f(x) = x - a$ for some $a \in F$. Then put

$$\begin{aligned} v_0 &= v \\ v_1 &= (T - aI)v_0 = Tv_0 - av_0 \\ v_2 &= (T - aI)v_1 = Tv_1 - av_1 \\ &\vdots \\ v_{e-1} &= (T - aI)v_{e-2} = Tv_{e-2} - av_{e-2} \\ 0 &= (T - aI)v_{e-1} = Tv_{e-1} - av_{e-1} \end{aligned}$$

Rearranging this just a bit we get

$$\begin{aligned}
Tv_0 &= av_0 + v_1\\
Tv_1 &= av_1 + v_2\\
&\vdots\\
Tv_{e-2} &= av_{e-2} + v_{e-1}\\
Tv_{e-1} &= av_{e-1}
\end{aligned}$$

Now by an argument similar (hard working graduate students will provide the variations needed) to the ones used above, we can see that the e distinct vectors $v_0, v_1, \ldots, v_{e-1}$ form a basis for the vector space $\mathbf{V}$. With respect to this basis, the linear operator T has the following matrix

$$\begin{pmatrix}
a & 0 & 0 & 0 & \ldots & 0 & 0\\
1 & a & 0 & 0 & \ldots & 0 & 0\\
0 & 1 & a & 0 & \ldots & 0 & 0\\
0 & 0 & 1 & a & \ldots & 0 & 0\\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots\\
0 & 0 & 0 & 0 & \ldots & a & 0\\
0 & 0 & 0 & 0 & \ldots & 1 & a
\end{pmatrix}.$$

A matrix of this form is called a **Jordan block** and the basis underlying is called a **Jordan basis**. Observe that it is easy given the polynomial $(x-a)^e$ to write down its Jordan block and, vice versa, given the Jordan block, the polynomial can be recovered at once. Moreover, $(x-a)^e$ is the characteristic polynomial $\det(xI-J)$ of the Jordan block J.

This time, appealing to the Structure Theorem we get

The Jordan Canonical Form Theorem.

Let $\mathbf{V}$ be a finite dimensional vector space over the field $\mathbf{F}$. Let T be a linear operator of $\mathbf{V}$ such that the irreducible factors of the minimal polynomial of T are all of degree 1. Then for some natural number n there are polynomials $(x-a_0)^{e_0}, (x-a_1)^{e_1}, \ldots, (x-a_{n-1})^{e_{n-1}} \in \mathbf{F}[x]$, namely the elementary divisors of $\mathbf{V}_T$, and cyclic primary submodules of $\mathbf{V}_T$: $\mathbf{V}_0$ of exponent $(x-a_0)^{e_0}$,$\ldots$, $\mathbf{V}_n$ of exponent $(x-a_{n-1})^{e_{n-1}}$ so that

$$\mathbf{V}_T \cong \mathbf{V}_0 \times \cdots \times \mathbf{V}_{n-1}.$$

Moreover, the natural number n is uniquely determined by T, the polynomials $(x-a_k)^{e_k}$ for $k<n$ are uniquely determined, and cyclic submodules $\mathbf{V}_0,\ldots,\mathbf{V}_n$ are determined up to isomorphism. Furthermore, each of the submodules $\mathbf{V}_k$ for $k \le n$ has a Jordan basis B_k, and $B_n \cup \cdots \cup B_0$ is a basis for $\mathbf{V}$. The linear operator T is represented with respect to this basis by

$$\begin{pmatrix}
J_0 & 0 & 0 & \ldots & 0\\
0 & J_1 & 0 & \ldots & 0\\
0 & 0 & J_2 & \ldots & 0\\
\vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 0 & 0 & \ldots & J_{n-1}
\end{pmatrix}$$

where the Jordan blocks of the $(x-a_k)^{e_k}$'s are placed in some order as square blocks along the diagonal, with all remaining entries of the matrix 0.

The matrix mentioned at the conclusion of this theorem is said to be in **Jordan canonical form**. It should be noted here, that there may be several matrices in Jordan form associated with the linear operator T

according to this theorem. This happens because the order in which the Jordan blocks appear along the diagonal is arbitrary. The permutation mentioned in the statement of the Structure Theorem reflects the same point. It is clear that if A and B are two matrices in Jordan form that can be obtained from each other by rearranging the Jordan blocks along the diagonal, the A and B are similar. So the Structure Theorem gives us

Jordan Canonical Form Theorem: Matrix Version.
Let $\mathbf{F}$ *be a field and* m *be a positive natural number. Let* A *be an* $m \times m$ *matrix with entries from* $\mathbf{F}$ *with the additional property that the irreducible factors of the minimal polynomial of the matrix are all of degree* 1*. Then* A *is similar to a matrix in Jordan canonical form and any matrices in Jordan canonical form that are similar to* A *can be obtained from each other by rearranging the Jordan blocks.*

11.1 Problem Set 11

Algebra Homework, Edition 11
Twelfth Week
Modules

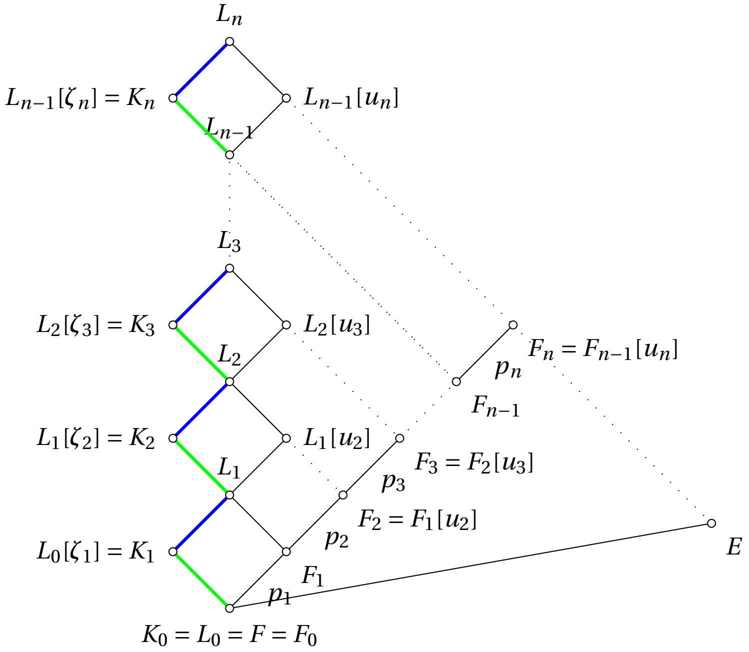
Problem 48.
Let **R** and **S** be commutative Noetherian rings. Prove that $\mathbf{R} \times \mathbf{S}$ is also Noetherian.

Problem 49.
Let **R** be a commutative ring such that every submodule of a free **R**-module is also a free **R**-module. Prove that **R** is a principal ideal domain.

Problem 50.
Let **R** be a principal ideal domain and let **M** and **N** be finitely generated **R**-modules such that $\mathbf{M} \times \mathbf{M} \cong \mathbf{N} \times \mathbf{N}$. Prove $\mathbf{M} \cong \mathbf{N}$.

Problem 51.
Give an example of two 4×4 matrices with real entries that have the same minimal polynomial and the same characteristic polynomial but are not similar.

Groups and Fields
Part II



Preface for Part II

The first part of this account of first-year graduate algebra was devoted to the essentials of rings and modules. There the central ambition was to give an account of the theory of unique factorization and the fundemental structure theorem for finitely generated modules over a principal ideal domain.

This second part begins with a presentation of the basics of the theory of groups, a presentation which is rapid in view of the sophistication gained by the graduate students in the first part of the course. But the real focus here will be an effort to lay our hands on the roots of polynomials whose coefficients all lie in some given field. In short, our most important goal will be a development of Galois theory. Along the way, we will see how to devise algebraically closed fields, we will have in hand a proof of the Fundamental Theorem of Algebra, a proof of the transcendence of the numbers π and e, as well as a proof of Hilbert's Nullstellensatz, which illuminates the Galois connection between algebraic geometry and the ring of polynomials in several variables over an algebraically closed field.

Amongst this rich array of material are the solutions, devised by mathematicians in the 19th century, to open problems of long standing—some for thousands of years. You will see why it is impossible to trisect arbitrary angles, duplicate the cube, or square the circle with only straightedge and compass—and why there is no formula similar to the quadratic formula for describing roots of polynomials of degree 5 or higher.

Once more, as you proceed through these pages you will find many places where the details and sometimes whole proofs of theorems will be left in your hands. The way to get the most from this presentation is to take it on with paper and pencil in hand and do this work as you go. There are also weekly problem sets. Most of the problems have appeared on Ph.D. examinations at various universities. In a real sense, the problem sets are the real heart of this presentation.

George F. McNulty
Columbia, SC
2016

Lecture 12

Concrete Groups

Consider the Euclidean plane $\mathbb{P}$. As a mathematical system we might construe $\mathbb{P}$ as $\langle P, B, E\rangle$ where P is the set of points on the plane, B is the three-place betweenness relation among points (we want $B(a,b,c)$ to mean that the point b is on the line segment joining the points a and c) and E is the four-place equidistance relation among points (we want $E(a,b,c,d)$ to mean that the distance between points a and b is the same as the distance between the points c and d; that is the line segment joining a and b is congruent to the line segment joining c and d). Were this a course in Euclidean geometry we would consider in detail the maps from plane into the plane that preserved the relations B and E. A bit of thought should lead hard-working graduate students to the conclusion that among these maps are the maps that preserve distance. That is σ is such a map provided for all points a and b, the distance from a to b is the same as the distance from $\sigma(a)$ to $\sigma(b)$. The fancy word for distance-preserving maps is *isometry*. More plain spoken folks call these rigid motions.

There are lots of isometries. For example, translating every point 2 units of distance to the northwest is an isometry. You could pick an arbitrary point as a center, and rotate the whole plane about that point by some angle θ. You could pick an arbitrary line and reflect the plane across the line.

The rigid motion easiest to understand is the one that does nothing: the identity map. Rigid motions can be composed and the result is again a rigid motion—one might first perform a translation and follow that by a reflection, for example. Each rigid motion is plainly one-to-one. It takes a bit of thought to see that they must map P onto P. This means that each rigid motion can be inverted. The hard-working graduate students can see that the inverse is again a rigid motion. In this way, more complex rigid motions can be devised by repeatedly composing and inverting the translations, rotations, and reflections. An interesting exercise, well within the grasp of hard-working graduate students, is to determine all of the isometries of the plane. Let $\mathbb{I}$ denote the set of all isometries of the plane.

There are some other maps that preserve the relations B and E. Here is one example. Fix a particular point $a \in \mathbb{P}$. Let σ be the map that sends any $q \in \mathbb{P}$ to the midpoint of the line segment joining p and q (and sending p to itself). This map is a contraction toward the point p. There are, of course, other contractions, to say nothing of expansions. Let $\operatorname{Aut}\mathbb{P}$ be the collection of all automorphisms of the plane—that is all the one-to-one maps from $\mathbb{P}$ onto $\mathbb{P}$ that preserve both the relations B and E. This collection includes all the isometries but is larger since it also includes all the expansions and contractions, as well as all the maps that arise from them by way of composition. These maps are sometimes called similarities. The hardy graduate student may try to classify all the maps that belong to $\operatorname{Aut}\mathbb{P}$.

Here is a similar situation. Let $\mathbb{R}_2$ denote the two-dimensional vector space over the field of real numbers. The automorphisms of $\mathbb{R}_2$ are just the invertible linear operators on this vector space. While we may identify the vectors with points on the plane, the vector space $\mathbb{R}_2$ and the Euclidean plane $\mathbb{P}$ are not the same. For example, $\mathbb{R}_2$ gives a special role to the origin whereas any point of $\mathbb{P}$ is like any other point. Also $\operatorname{Aut}\mathbb{R}_2$ and $\operatorname{Aut}\mathbb{P}$ are different as well. Each automorphism of $\mathbb{R}_2$ fixes the origin. So there are no nontrivial translations in $\operatorname{Aut}\mathbb{R}_2$, the only rotations must use the origin as their centers, and the only reflections must be reflections across lines through the origin. So a lot of maps in $\operatorname{Aut}\mathbb{P}$ seem to be missing from $\operatorname{Aut}\mathbb{R}_2$. On the other hand, $\operatorname{Aut}\mathbb{R}_2$ has maps that are not rigid motions. For example, in $\operatorname{Aut}\mathbb{R}_2$ one can find scalings, that is maps which stretch or shrink vectors. This is effected by multiplying by a fixed nonzero scalar. At any rate, $\operatorname{Aut}\mathbb{R}_2$ contains the identity map, it is closed under composition of linear operators, and it is also closed under the formation of inverse of linear operators.

Here is a related situation. Let us consider just a part of $\operatorname{Aut}\mathbb{R}_2$, namely those linear operators with determinant 1. (It may help to think of each linear operator as a 2×2 matrix.) Let S denote this collection of more specialized invertible linear operators. The only scalings that remain in S are multiplication by 1 (namely, the identity map) and multiplication by -1. However, S is still pretty rich. For all intents and purposes, it is the collection of 2×2 matrices with real entries that have determinant 1. As with the other cases, the identity map belongs to S and S is closed under composition of operators and under the formation of inverses of operators.

We have four examples: $\mathbb{I}$, $\operatorname{Aut}\mathbb{P}$, $\operatorname{Aut}\mathbb{R}_2$, and S. Each of these is a collection of one-to-one functions from some set onto itself. Each of these collections includes the identity map and is closed under composition of functions and under the formation of inverse functions. In a sense, each of these are collections of second order objects that are functions on some (first-order) mathematical system. Evidently, we could derive such collections of the second-order from a wide assortment of mathematical systems. We can convert these three sets, and any others that arise in a similar way, into algebraic systems (algebras, for short) as

$$\langle \operatorname{Aut}\mathbb{P}, \circ, ^{-1}, \mathbf{1}\rangle \qquad \langle \operatorname{Aut}\mathbb{R}_2, \circ, ^{-1}, \mathbf{1}\rangle \qquad \langle S, \circ, ^{-1}, \mathbf{1}\rangle.$$

Here we use $\mathbf{1}$ to denote the identity map, $\circ$ to denote the composition of functions, and $^{-1}$ to denote the formation of inverses. These are algebras whose signature provides one two-place operation symbol to designate the composition, a single one-place operation symbol to designate the formation of inverses, and an operation symbol of rank 0 to designate the identity.

The general situation, of which these are three special cases, starts with a set X. In our first example X is the Euclidean plane. We consider a set G of one-to-one maps from X onto X that includes the identity map $\mathbf{1}_X$ on X and that is closed with respect to both the composition of functions and the formation of inverse functions. The resulting algebra

$$\langle G, \circ, ^{-1}, \mathbf{1}_X\rangle$$

is called a **concrete group**. Since X can be any set and the selection of a particular G given X is unrestrained, apart from the closure conditions, there is quite a rich assortment of concrete groups. Even so, we see that for concrete groups we know very well how the operations of functional composition and the formation of inverse functions work and we have a firm grip on the identity map.

A **group** is any algebra that is isomorphic to a concrete group. It is interesting to note that the concept of a group arises by the process of abstraction from its concrete instances. Loosely speaking, an (abstract) group is a mathematical system that shares all its "algebraic" properties with some concrete group. This process differs from the process of generalization which prompted the concept of a ring. There, the idea was to extract from many particular instances, like the integers or the system of 2×2 matrices, a set of common properties. A ring was any algebra that had the selected common properties. One should notice that the properties we selected in coming to the notion of a ring were conventional and practical—that is, they were convenient properties like the distributive law, which arose again and again in practice. The theory of rings, in some sense, is the working out of the logical consequences of these selected properties. While these properties of plus and times are fairly natural in that they arose in the course of millennia of mathematical practice, there does not appear to be anything absolutely inevitable about them. The notion of a group, on the other hand, did not arise through the selection of a set of properties but rather through the selection of a class of concrete instances.

Groups of Permutations

Before turning to the theory of abstract groups, we will develop the first facts about concrete groups.

Let X be any set. A **permutation** on X is just a one-to-one function from X onto X. We use $\operatorname{Sym} X$ to the denote the set of all permutations on X and $\mathbf{Sym}\, X$ to denote the (concrete) group $\langle \operatorname{Sym} X, \circ, {}^{-1}, \mathbf{1}_X\rangle$. We refer to this group as the **symmetric group** on X or sometimes as the groups of symmetries of X.

If Y is a set such that $|Y| = |X|$, that is if X and Y have the same cardinality, then $\mathbf{Sym}\, X$ and $\mathbf{Sym}\, Y$ will be isomorphic. Indeed, suppose that f is a one-to-one correspondence from X to Y. Then the map

$$\sigma \mapsto f \circ \sigma \circ f^{-1}$$

for all $\sigma \in \operatorname{Sym} X$, turns out to be an isomorphism from $\mathbf{Sym}\, X$ to $\mathbf{Sym}\, Y$, as the hard-working graduate can check.

For many purposes the thing about X that really matters for $\mathbf{Sym}\, X$ is the cardinality of X. This being the case, for a cardinal κ we use $\mathbf{S}_\kappa$ to denote the concrete group of all permutations of κ. Here we take κ to be the set of all ordinals strictly smaller than κ. When κ is finite, this means that $\kappa = \{0, 1, 2, \ldots, \kappa - 1\}$. For example, $6 = \{0, 1, 2, 3, 4, 5\}$. So $\mathbf{S}_6 = \mathbf{Sym}\{0, 1, 2, 3, 4, 5\}$.

Let $x \in X$ and $\sigma \in \operatorname{Sym} X$. The set

$$\{\sigma^k(x) \mid k \in \mathbb{Z}\}$$

is called the **orbit** of x under the action of σ. For a fixed permutation σ, the set X is actually partitioned into orbits. The equivalence relation that lies behind this partition makes elements $x, y \in X$ equivalent if and only if $\sigma^k(x) = y$ for some integer k. Notice that an orbit is either countably infinite or finite. The countably infinite orbits arise when $\sigma^k(x) \neq \sigma^j(x)$ whenever $k \neq j$. These orbits can the arranged like the integers:

$$(\ldots, \sigma^{-3}(x), \sigma^{-2}(x), \sigma^{-1}(x), \sigma^0(x), \sigma^1(x), \sigma^2(x), \sigma^3(x), \ldots)$$

where, of course, $\sigma^0(x) = x$. Suppose, on the other hand, that $\sigma^k(x) = \sigma^j(x)$ where $j < k$. Then some fiddling reveals that $\sigma^{k-j}(x) = x$. Let n be the smallest positive integer such that $\sigma^n(x) = x$. Then the orbit of x under the action of σ turns out to be $\{x, \sigma(x), \ldots, \sigma^{n-1}(x)\}$, as checked by every one of the hard-working graduate students. We can also regard this orbit as a kind of arranged list:

$$(x, \sigma(x), \sigma^2(x), \sigma^3(x), \ldots, \sigma^{n-1}(x))$$

so long as we think of it as a linear representation of a circular arrangement—that is we think of $\sigma^{n-1}(x)$ as the (unrepresented) predecessor of x. We could represent the permutation σ by simply listing all these arranged orbits. For instance, here is such a representation of one member of S_6:

$$\sigma = (0, 2, 4)(1)(5, 3)$$

This is a compact way that writing

$$\sigma(0) = 2$$
$$\sigma(2) = 4$$
$$\sigma(4) = 0$$
$$\sigma(1) = 1$$
$$\sigma(5) = 3$$
$$\sigma(3) = 5.$$

The natural number 1 is a fixed point of σ. By convention, fixed points are omitted from the representation. So we arrive at

$$\sigma = (0,2,4)(5,3).$$

The two parts in this representation are called **cycles**. They are cyclic representations of the two nontrivial (here that means have at least two elements) orbits into which σ partitions the set $\{0,1,2,3,4,5\}$. These cycles have lengths: $(0,2,4)$ is a three-cycle, while $(5,3)$ is a two-cycle. The orbits are, of course, disjoint. So we have decomposed σ into disjoint cycles.

Something interesting emerges here. Let τ be the permutation in S_6 represented by $(0,2,4)$ and ρ be the permutation represented by $(5,3)$. Then, as the hard-working graduate student can check, $\sigma = \tau \circ \rho$. This suggests that we can capture composition of permutations by juxtaposing a bunch of cycles. Suppose μ is the permutation on $\{0,1,2,3,4,5\}$ represented by $(0,1)(2,3)(4,5)$. We would like to represent $\sigma \circ \mu$ by

$$(0,2,4)(5,3)(0,1)(2,3)(4,5).$$

Observe that the listed cycles are no longer disjoint, so we should not think of this as a list of orbits. Rather, let us see what happens to the natural number 2 when we apply $\sigma \circ \mu$ to it. We know $\mu(2) = 3$ and $\sigma(3) = 5$ so that $\sigma \circ \mu(2) = 5$. Now consider the following

$$\begin{aligned}
\sigma \circ \mu(2) &= (0,2,4)(5,3)(0,1)(2,3)(4,5)2 \\
&= (0,2,4)(5,3)(0,1)(2,3)2 \quad \text{since } (4,5) \text{ fixes } 2 \\
&= (0,2,4)(5,3)(0,1)3 \\
&= (0,2,4)(5,3)3 \\
&= (0,2,4)5 \\
&= 5.
\end{aligned}$$

A more compact way to display the same information is

$$\begin{array}{ccccccccccc}
 & (0,2,4) & & (5,3) & & (0,1) & & (2,3) & & (4,5) & \\
5 & \longleftarrow & 5 & \longleftarrow & 3 & \longleftarrow & 3 & \longleftarrow & 2 & \longleftarrow & 2
\end{array}$$

In fact, we could extend this to display the whole effect of the composite permutation.

$$\begin{array}{ccccccccccc}
 & (0,2,4) & & (5,3) & & (0,1) & & (2,3) & & (4,5) & \\
1 & \longleftarrow & 1 & \longleftarrow & 1 & \longleftarrow & 0 & \longleftarrow & 0 & \longleftarrow & 0 \\
2 & \longleftarrow & 0 & \longleftarrow & 0 & \longleftarrow & 1 & \longleftarrow & 1 & \longleftarrow & 1 \\
5 & \longleftarrow & 5 & \longleftarrow & 3 & \longleftarrow & 3 & \longleftarrow & 2 & \longleftarrow & 2 \\
4 & \longleftarrow & 2 & \longleftarrow & 2 & \longleftarrow & 2 & \longleftarrow & 3 & \longleftarrow & 3 \\
3 & \longleftarrow & 3 & \longleftarrow & 5 & \longleftarrow & 5 & \longleftarrow & 5 & \longleftarrow & 4 \\
0 & \longleftarrow & 4 & \longleftarrow & 4 & \longleftarrow & 4 & \longleftarrow & 4 & \longleftarrow & 5
\end{array}$$

Inspecting this array we find

$$\sigma \circ \mu = (0,1,2,5)(3,4),$$

which is the decomposition of $\sigma \circ \mu$ into a product of disjoint cycles.

Representing the inverse of a permutation is easy. For example, the inverse of $(0,1,2,5)(3,4)$ is just $(4,3)(5,2,1,0)$. We just write everything in reverse order.

We could pursue the same strategy of notation for representing the permutations and their products and inverses for any set X. For infinite sets, this becomes tricky, but when X is finite the strategy can be carried through without trouble.

Our first Fact is clear.

Fact. Every permutation on a finite set can be decomposed as a product of disjoint cycles. This decomposition is unique, up to rearranging the cycles. Also, any two disjoint cycles commute.

Permutations, even or odd

Let X be a set. A permutation σ of X is a **transposition** provided there are distinct elements $x, y \in X$ such that σ exchanges x and y and leaves every other element of X fixed—that is $\sigma(x) = y$ and $\sigma(y) = x$ and $\sigma(w) = w$ for all $w \in X \setminus \{x, y\}$. In the notation above, this means $\sigma = (x, y)$. Evidently, σ is its own inverse: $\sigma \circ \sigma = \mathbf{1}_X$.

Fact. Every permutation on a finite set can be decomposed as a product of transpositions.

Proof. The identity permutation is the product of the empty system of transpositions. Since every permutation is a product of cycles, we need only prove that every cycle is a product of transpositions. Let $a_0, a_1, \ldots, a_{k-1}$ be distinct elements of our finite set. Just check

$$(a_0, a_1, \ldots, a_{k-1}) = (a_1, a_2)(a_0, a_{k_1})\ldots(a_0, a_3)(a_0, a_2).$$

□

The decomposition of a permutation into transpositions is not unique. For example, $(0,1,2) = (1,2)(0,2) = (1,0)(1,2) = (0,2)(2,1)(0,1)(0,2)$. However, a shred of uniqueness remains.

Fact. Let $\sigma_0, \ldots, \sigma_{k-1}$ and $\tau_0, \ldots, \tau_{\ell-1}$ be sequences of transpositions such that

$$\sigma_0 \circ \sigma_1 \circ \cdots \circ \sigma_{k-1} = \tau_0 \circ \tau_1 \circ \cdots \circ \tau_{\ell-1}.$$

Then k and ℓ have the same parity, that is either both k and ℓ are even or both k and ℓ are odd.

Proof. Let X be the underlying set. Let us assume to the contrary that k is even and ℓ is odd. Since every transposition is its own inverse, we are led to

$$\tau_{\ell-1} \circ \cdots \circ \tau_0 \circ \sigma_0 \circ \sigma_1 \circ \cdots \circ \sigma_{k-1} = \mathbf{1}_X.$$

So we see that the identity $\mathbf{1}_X$ can be written as the product of a sequence of transpositions of length $k + \ell$, which is odd. So our proof will be completed by the following contention, since the identity permutation fixes every element of X.

Contention. Suppose σ is a permutation of X so that

$$\sigma = \tau_0 \circ \tau_1 \circ \cdots \circ \tau_{m-1}$$

where m is odd and each τ_j is a transposition. Then σ moves some element of X.

We assume, without loss of generality, that m is the smallest odd number so that σ is the product of a sequence of length m of transpositions. Let $\tau_{m-1} = (a, e)$ where a and e are distinct elements of X. (The hard working graduate students must figure out what to do when X has fewer than two elements.) We will actually prove that σ moves a. We could do the same for any element of X that is moved by any of the transpositions $\tau_0, \tau_1, \ldots, \tau_{m-1}$. We achieve this by rewriting the factorization of σ in $m - 1$ steps. We start by letting $\rho_{m-1} = \tau_{m-1}$. So our initial factorization is

$$\sigma = \tau_0 \circ \ldots \tau_k \circ \tau_{k+1} \circ \tau_{k+2} \circ \cdots \circ \tau_{m-2} \circ \rho_{m-1}.$$

It has the property that no transpostion to the *right* of ρ_{m-1} moves a. After a number of steps, we will have the factorization

$$\sigma = \tau_0 \circ \ldots \tau_k \circ \rho_{k+1} \circ \tau'_{k+2} \circ \cdots \circ \tau'_{m-2} \circ \tau'_{m-1},$$

where ρ_{k+1} and τ'_j for $k + 1 < j < m$ are transpositions and a is moved by ρ_{k+1} but fixed by all the transpositions to its right in the factorization.

The rewriting happens in this way. We will replace $\tau_k \circ \rho_{k+1}$ by $\rho_k \circ \tau'_{k+1}$ so that a is moved by ρ_k but fixed by τ'_{k+1}. First we observe that $\tau_k \neq \rho_{k+1}$ since otherwise $\tau_k \circ \rho_{k+1} = \mathbf{1}_X$ and we could delete these two factors resulting in a factorization of σ of smaller odd length—a violation of the minimality of m. Let us say that $\rho_{k+1} = (a,b)$. Then there are only three alternatives for τ_k: it is disjoint from (a,b) or it moves b but not a or it moves a but not b. So $\tau_k \circ \rho_{k+1}$ has one the the following forms

$$\begin{gathered}(c,d)(a,b)\\(b,c)(a,b)\\(a,c)(a,b)\end{gathered}$$

where all the letters in each line stand for distinct elements of X. It is easy to check that each of the equations below holds in $\mathbf{Sym}\,X$.

$$\begin{gathered}(c,d)(a,b) = (a,b)(c,d)\\(b,c)(a,b) = (a,c)(b,c)\\(a,c)(a,b) = (a,b)(b,c)\end{gathered}$$

So to obtain the next factorization of σ we simply replace the left side of the appropriate equation by its right side. For example, if $\tau_k = (b,c)$, then we take $\rho_k = (a,c)$ and $\tau'_{k+1} = (b,c)$. Here is what happens in detail:

$$\begin{aligned}\sigma &= \tau_0 \circ \cdots \circ \;\; \tau_k \;\; \circ \; \rho_{k+1} \circ \tau'_{k+2} \circ \cdots \circ \tau'_{m-1}\\\sigma &= \tau_0 \circ \cdots \circ (b,c) \circ (a,b) \circ \tau'_{k+2} \circ \cdots \circ \tau'_{m-1}\\\sigma &= \tau_0 \circ \cdots \circ (a,c) \circ (b,c) \circ \tau'_{k+2} \circ \cdots \circ \tau'_{m-1}\\\sigma &= \tau_0 \circ \cdots \circ \;\; \rho_k \;\; \circ \; \tau'_{k+1} \circ \tau'_{k+2} \circ \cdots \circ \tau'_{m-1}\end{aligned}$$

After $m-1$ rewrite steps of this kind we obtain

$$\sigma = \rho_0 \circ \tau'_1 \circ \cdots \circ \tau'_{m-1},$$

a factorization of σ into transpositions. But now observe that a is moved by ρ_0 but fixed by all the τ'_k's. Hence, σ moves a, as desired. □

We will call a permutation σ of a set X **even** if it can be decomposed as a product of a sequence of transpositions, the sequence being of even length. We call σ **odd** if it can be decomposed as a product of a sequence of transpositions, the sequence being of odd length. If the number of elements of X moved by σ is finite, we see from the two facts above, that these are mutually exclusive and exhaustive alternatives. If the number of elements of X moved by σ is infinite, then σ cannot be written as the product of an finite sequence of transpositions, since any permutation that can be written as a product of a finite sequence of transpositions must fix all members of X that do not arise in the transpositions. Any permutation which moves infinitely many members of X cannot be either even or odd. For example, the permutation of $\mathbb{Z}$ such that $k \mapsto k+1$ for all $k \in \mathbb{Z}$ is of this kind.

In any case, the set $\mathrm{Alt}\,X$ of even permutations of X, contains $\mathbf{1}_X$ and is closed under composition and the formation of inverses. So

$$\mathbf{Alt}\,X := \langle \mathrm{Alt}\,X, \circ, {}^{-1}, \mathbf{1}_X \rangle$$

is another example of a concrete group. It is called the **alternating group** on X. When $X = n = \{0,1,\ldots,n-1\}$ we adopt the notation $\mathbf{A}_n$.

12.1 Problem Set 12

Algebra Homework, Edition 12
Thirteenth Week
Dealing with oncrete groups and a few other matters

Problem 52.

(a) How many elements in $\mathbf{S}_8$ commute with the permutation $(012)(34567)$?

(b) How many elements are there in S_8 with order 15?

Problem 53.

(a) Let $\mathbf{G}$ be a subgroup of the symmetric group $\mathbf{S}_n$. Show that if G contains an odd permutation, then $\mathbf{G} \cap \mathbf{A}_n$ is of index 2 in $\mathbf{G}$.

(b) Let $\mathbf{G}$ be a group of order $2t$, where t is odd, and consider the representation $\tau\colon \mathbf{G} \to \mathbf{S}_{2t}$, given by $\tau(g)(h) = gh$, for $h \in G$. Show that if $g^2 = 1$, then $\tau(g)$ is an odd permutation.

(c) Deduce that a group of order $2t$, where t is odd, cannot be simple.

Problem 54.

(a) Let $\mathbf{S}_n$ denote the group of permutations of the set $\{0, 1, \ldots, n-1\}$. How many different subgroups of order 4 does $\mathbf{S}_4$ have? Justify your calculation. (Two subgroups are considered different if they are different as sets.)

(b) There is a homomorphism of $\mathbf{S}_4$ onto $\mathbf{S}_3$. (You do not need to prove that there exists such a homomorphism.) Show that there is no homomorphism of $\mathbf{S}_5$ onto S_4.

Problem 55.
Let $\mathbf{M}$ and $\mathbf{N}$ be finitely generated modules over the same principal ideal domain. Prove that if $\mathbf{M} \times \mathbf{N} \times \mathbf{M} \cong \mathbf{N} \times \mathbf{M} \times \mathbf{N}$, then $\mathbf{M} \cong \mathbf{N}$.

Problem 56.
Give an example of two **dissimilar** matrices A and B with real entries that have all the following properties:

(a) A and B have the same minimal polynomial,

(b) A and B have the same characteristic polynomial, and

(c) The common minimal polynomial has no real roots.

LECTURE 13

The Theory of Abstract Groups: Getting Off the Ground

13.1 Defining the class of groups by equations

In a concrete group the operations are easy to understand: composition of functions, the formation of inverse functions, and the identity function as distinguished function. In an abstract group we lose this tight grip on the basic operations. Nevertheless, since an isomorphism ties each abstract group to a concrete group many properties of the basic operations of the concrete group must also hold in the abstract case. In the midst of the 19th century Arthur Cayley made the breakthrough to the class of abstract groups with the following theorem.

Cayley's Equational Axiomatization Theorem. *Let* $\mathbf{G} = \langle G, \cdot, ^{-1}, 1\rangle$ *be an algebra of the signature of groups.* $\mathbf{G}$ *is a group if and only if the following equations are true in* $\mathbf{G}$*:*

$$\begin{array}{lll} x^{-1} \cdot x = 1 & x \cdot (y \cdot z) = (x \cdot y) \cdot z & 1 \cdot x = x \\ x \cdot x^{-1} = 1 & & x \cdot 1 = x \end{array}$$

Proof. Suppose first that $\mathbf{G}$ is a group. This means that it is isomorphic to a concrete group. The five equations above hold about composition of functions, the formation of inverse functions, and the identity function. So they must also hold in $\mathbf{G}$.

Now let us suppose that $\mathbf{G}$ is an algebra in which these five equations happen to be true. To prove that $\mathbf{G}$ is a group we will devise a concrete group and show that $\mathbf{G}$ can be embedded into it. Having nothing else at hand we will take $\mathbf{Sym}\, G$ as our concrete group. Our embedding $\Phi : \mathbf{G} \to \mathbf{Sym}\, G$ can be given explicitly. We see that for each $g \in G$ that $\Phi(g)$ is supposed to be a permutation of G. Since writing things like $\Phi(g)(h)$ is clumsy, we use Φ_g in place of $\Phi(g)$. For each $g \in G$ we define $\Phi_g : G \to G$ in the following way. For any $h \in G$, put

$$\Phi_g(h) = g \cdot h.$$

We need to verify that each Φ_g is indeed a permutation of G, that the map Φ is one-to-one, and that Φ is a homomorphism. The five equations above allow us to succeed.

Contention. Φ_g is a permutation of G, for each $g \in G$.

It is evident from the definition of Φ_g that it is a function from G to G. Since the permutations of G are just the invertible functions from G to G, we will prove here that Φ_g and $\Phi_{g^{-1}}$ are inverses of each other. We will need this anyway to see that Φ is a homomorphism. That Φ_g and $\Phi_{g^{-1}}$ are inverses is equivalent to the assertion that for all $h, k \in G$

$$\Phi_g(h) = k \text{ if and only if } \Phi_{g^{-1}}(k) = h.$$

Here is the argument:

$$\begin{aligned}
\Phi_g(h) = k &\Rightarrow g \cdot h = k \\
&\Rightarrow g^{-1} \cdot (g \cdot h) = g^{-1} \cdot k \\
&\Rightarrow (g^{-1} \cdot g) \cdot h = \Phi_{g^{-1}}(k) \quad \text{by associativity} \\
&\Rightarrow 1 \cdot h = \Phi_{g^{-1}}(k) \quad \text{by } x^{-1} \cdot x = 1 \\
&\Rightarrow h = \Phi_{g^{-1}}(k) \quad \text{by } 1 \cdot x = x \\
&\Rightarrow h = g^{-1} \cdot k \\
&\Rightarrow g \cdot h = g \cdot (g^{-1} \cdot k) \\
&\Rightarrow \Phi_g(h) = (g \cdot g^{-1}) \cdot k \quad \text{by associativity} \\
&\Rightarrow \Phi_g(h) = 1 \cdot k \quad \text{by } x \cdot x^{-1} = 1 \\
&\Rightarrow \Phi_g(h) = k \quad \text{by } 1 \cdot x = x
\end{aligned}$$

This argument is a bit pedantic but it does show where four of our five equations come into play.

Contention. Φ is one-to-one.

Let us suppose that $\Phi_g = \Phi_h$. Then in particular, $\Phi_g(1) = \Phi_h(1)$. So we find

$$g = g \cdot 1 = \Phi_g(1) = \Phi_h(1) = h \cdot 1 = h.$$

Here we have appealed to our fifth equation $x \cdot 1 = x$, to conclude that Φ is one-to-one.

Here is our last contention:

Contention. Φ is a homomorphism.

So we must establish the following:

$$\begin{aligned}
\Phi_{g \cdot h} &= \Phi_g \circ \Phi_h \\
\Phi_{g^{-1}} &= \left(\Phi_g\right)^{-1} \\
\Phi_1 &= \mathbf{1}_G
\end{aligned}$$

Let k be any element of G. Here is the demonstration of the first piece.

$$\Phi_{g \cdot h}(k) = (g \cdot h) \cdot k = g \cdot (h \cdot k) = g \cdot \Phi_h(k) = \Phi_g\left(\Phi_h(k)\right) = \Phi_g \circ \Phi_h)(k).$$

We already established the second piece on the way to showing that Φ_g is a permutation. The last is easiest of all:

$$\Phi_1(k) = 1 \cdot k = k, \text{ for all } k \in G.$$

So Φ_1 is the identity function, as desired.

Notice that what we showed is that Φ is a one-to-one homomorphism from **G** into **Sym** G. We did not claim that Φ is onto Sym G. Rather that concrete group isomorphic to **G** is a subalgebra of **Sym** G. □

Corollary 13.1.1. *Every subalgebra of a group is again a group. Every homomorphic image of a group is again a group. The direct product of any system of groups is again a group.*

These facts all follow since the truth equations is preserved under all these constructions.

In our proof of Cayley's Theorem each of the five equations came into play, suggesting that perhaps all of them are needed. On the other hand, maybe a slicker proof would avoid the use of some of those equations. Actually, two of the equations can be omitted, as indicated in one of the problem sets.

Notably absent from the five equations listed in Cayley's Theorem is the commutative law: $x \cdot y = y \cdot x$. Of course the reason is that composition of functions (even permutations) is not generally commutative. However, there are many familiar groups, like $\langle \mathbb{Z}, +, -, 0\rangle$, that satisfy the commutative law. They are called **Abelian** groups in honor of Niels Hendrick Abel.

While starting with concrete groups and then forming the class of all abstract groups and then seeing that this latter class can be described by a set of simple equations reflects the actual historical development, most expositors of group theory have chosen to start with the set of five equations (or something like them) as a way to define the notion of a group. From this perspective Cayley's Theorem is called the Cayley Representation Theorem because it shows that every group (i.e. an algebra satisfying those equations) is isomorphic to (can be represented as) a group of permutations.

A large number of authors conceive a group as an algebra $\langle G, \cdot\rangle$ with the follow properties:

(a) $\cdot$ is an associative operation on G.

(b) There is an element $1 \in G$ such that $1 \cdot a = a \cdot 1 = a$ for all $a \in G$.

(c) For every $a \in G$ there is some $b \in G$ so that $a \cdot b = b \cdot a = 1$.

They then go on to show that only one element of a group can play the role of 1 and that every element of a group has exactly one inverse. This approach has the advantage that the algebras involved have only one operation. But it carries with it some annoyance as well. To see why, here is a more formalized presentation of this axiomatic approach.

(a) $\forall x, y, z\left[x \cdot (y \cdot z) \approx (x \cdot y) \cdot z\right]$.

(b) $\exists u \forall x\,[u \cdot x \approx x \cdot u \approx x]$.

(c) $\forall x \exists y \forall z\left[(x \cdot y) \cdot z \approx z \approx z \cdot (x \cdot y) \& (y \cdot x) \cdot z \approx z \approx z \cdot (y \cdot x)\right]$.

By formalizing these statements we can see easily that they have more involved logical forms than equations. While it takes more work, one can prove that the truth of sentences of these forms are preserved under the formation of homomorphic images and direct products. The same does not apply to subalgebras. For example $\langle \mathbb{R}^+, \cdot\rangle$ the positive reals under multiplication, is a group in this sense. It is easy to see that the positive reals strictly less than 1 constitute a subalgebra. But this subalgebra has neither a multiplicative unit (1 is just not in there) nor does any element have a multiplicative inverse. So this sublagebra is not a group. This annoyance in minor, of course. When group theory is developed from this starting point it very soon (say within five pages) gets to the point where one has named the multlipicative unit and begins using some notation for the formation of inverses. From that point on the development is carried forward just as if these operations were given at the very beginning.

One might wonder if group theory could be developed using a different choice of basic operations. Indeed, it can. Since $x \cdot x^{-1} = 1$ we see that the distinguished element can be defined by an equation. We could regard 1 as an abbreviation for $x \cdot x^{-1}$ provided we add the equation $x \cdot x^{-1} = y \cdot y^{-1}$ to our list. The we could dispense with 1 and make do with just $\cdot$ and $^{-1}$. A more radical step is to use the operation

$$x \mid y := x^{-1} \cdot y.$$

It takes some doing, but the hard working graduate students filled up with interest about this point should be able to write down a short list of equations just in terms of the one two-place operation symbol $\mid$ which entirely captures the notion of group. One way to start this project is to try to devise a term in x and y and | to recapture $\cdot$. A peculiar feature of this approach to group theory is the discovery of a single (rather long) equation in | that defines the class of all groups. From the perspective of group theory, such a single equation would be a very cumbersome basis upon which to develop the subject. On the other hand, from the perspective of mathematical logic, it is a remarkable and intriguing property of the notion of a group.

13.2 Homomorphisms and their kernels—and an extraordinary property of subgroups

Just as they did in the case of rings, homomorphisms and their kernels will play a central role in the development of group theory.

Fact. Let **G** and **H** be groups and let $f : G \to H$. Then f is a homomorphism if and only if $f(a \cdot b) = f(a) \cdot f(b)$ for all $a, b \in G$.

Proof. Of course the direction from left to right is immediate. For the converse, we need to show that f preserves the other operations. The equalities

$$f(1) \cdot 1 = f(1) = f(1 \cdot 1) = f(1) \cdot f(1)$$

hold because both **G** and **H** are groups and because f preserves the product. So we see $f(1) \cdot 1 = f(1) \cdot f(1)$ holds in **H**. Because **H** is a group we can cancel $f(1)$ from both sides, leaving $1 = f(1)$, which we need if f is going to be a homomorphism. So what about inverses? Well

$$1 = f(1) = f(a \cdot a^{-1}) = f(a) \cdot f(a^{-1}).$$

So $1 = f(a) \cdot f(a^{-1})$ holds in **H**. Since **H** is a group, we can multiply both sides of this equation by $\left(f(a)\right)^{-1}$ to arrive at $\left(f(a)\right)^{-1} = f(a^{-1})$, as desired. So the Fact is established. □

In using this fact to prove that some map is a homomorphism it is important to prove in advance that both **G** and **H** are groups.

Now suppose **G** is a group and h is a homomorphism from **G** onto **H**. Let

$$\theta := \{(a, b) \mid a, b \in G \text{ and } h(a) = h(b)\}.$$

So θ is the functional kernel of h. It was convenient in the theory of rings to replace θ with the congruence class of 0. We can do something along these lines in groups as well. Observe for all $a, b \in G$ we have

$$a\,\theta\, b \Leftrightarrow (b^{-1} \cdot a)\,\theta\,(b^{-1} \cdot b) \Leftrightarrow (b^{1} \cdot a)\,\theta\, 1.$$

Another way to formulate this is $b \in a/\theta \Leftrightarrow (b^{-1} \cdot a) \in 1/\theta$. The upshot is that the congruence class $1/\theta$ completely determines the whole congruence relation. (Just as in ring theory congruence class of 0—an ideal—completely determined the congruence relation.) What properties of $1/\theta$ make it so special?

- $1/\theta$ is a subgroup of **G**.
- If $a \in 1/\theta$ and $b \in G$, then $(b^{-1} \cdot a \cdot b) \in 1/\theta$.

The hard-working graduate student will have no trouble verifying these points. For instance, the last follows since if $h(a) = 1$, then $h(b^{-1} \cdot a \cdot b) = h(b)^{-1} \cdot h(a) \cdot h(b) = h(b)^{-1} \cdot 1 \cdot h(b) = 1$.

A subgroup **N** of a group **G** is said to be **normal** provided for all a and b, if $a \in N$ and $b \in G$, then $(b^{-1} \cdot a \cdot b) \in N$. We use $\mathbf{N} \lhd \mathbf{G}$ to symbolize that **N** is a normal subgroup of **G**.

Theorem on Congruences and Normal Subgroups. *Let* **G** *be a group. The following are equivalent.*

(a) **N** *is a normal subgroup of* **G**.

(b) $N = 1/\theta$ *for some congruence* θ *of* **G**.

(c) $N = \{a \mid a \in G \text{ and } h(a) = 1\}$ *for some homomorphism from* **G**.

(d) $\{(a,b) \mid a,b \in G \text{ and } (b^{-1}\cdot a) \in N\}$ *is a congruence of* **G**.

Proof. Based on the discussion above and on our general understanding of the connection between homomorphisms and congruence relations, the only implication that really calls for further attention is $(a) \Rightarrow (b)$.

To establish this implication, let **N** be a normal subgroup of **G** and put

$$\theta := \{(a,b) \mid a,b \in G \text{ and } (b^{-1}\cdot a) \in N\}.$$

Contention. θ is an equivalence relation on G.

For every $a \in G$ we see that $a^{-1}\cdot a = 1 \in N$ since **N** is a subgroup of **G**. This means that $(a,a) \in \theta$, so θ is reflexive.

To see symmetry, suppose $(a,b) \in \theta$. This means $(b^{-1}\cdot a) \in N$. Since **N** is a subgroup of **G** we know that $(b^{-1}\cdot a)^{-1} \in N$. But because **N** is a group we know that $a^{-1}\cdot b = (b^{-1}\cdot a)^{-1}$. Hence $(a^{-1}\cdot b) \in N$. But this entails $(b,a) \in \theta$ and the symmetry of θ is proved.

Finally, for transitivity suppose $(a,b),(b,c) \in \theta$. This means

$$(b^{-1}\cdot a) \in N$$
$$(c^{-1}\cdot b) \in N.$$

Since **N** is a subgroup of **G** we get

$$(c^{-1}\cdot b)\cdot(b^{-1}\cdot a) \in N.$$

Just a bit of fiddling gives $(c^{-1}\cdot a) \in N$. This entails $(a,c) \in \theta$ and establishes the transitivity of θ, concluding the proof that θ is an equivalence relation on G. It is useful to point out that only the fact that **N** is a *subgroup* of **G** was used here.

Contention. θ is a congruence relation of **G**.

We need to establish two things:

$$a\,\theta\,b \text{ and } c\,\theta\,d \Rightarrow (a\cdot c)\,\theta\,(b\cdot d)$$
$$a\,\theta\,b \Rightarrow a^{-1}\,\theta\,b^{-1}.$$

Using the definition of θ, these become

$$(b^{-1}\cdot a) \in N \text{ and } (d^{-1}\cdot c) \in N \Rightarrow \big((b\cdot d)^{-1}\cdot(a\cdot c)\big) \in N$$
$$(b^{-1}\cdot a) \in N \Rightarrow \big((b^{-1})^{-1}\cdot a^{-1}\big) \in N.$$

After some minor fiddling this gives

$$(b^{-1}\cdot a) \in N \text{ and } (d^{-1}\cdot c) \in N \Rightarrow \big(d^{-1}\cdot b^{-1}\cdot a\cdot c\big) \in N$$
$$(b^{-1}\cdot a) \in N \Rightarrow \big(b\cdot a^{-1}\big) \in N.$$

Let's tackle the top implication. Suppose $(b^{-1}\cdot a),(d^{-1}\cdot c) \in N$. Because **N** is a normal subgroup we see that $(d^{-1}\cdot b^{-1}\cdot a\cdot d) \in N$. Because **N** is a subgroup $d^{-1}\cdot b^{-1}\cdot a\cdot c = (d^{-1}\cdot b^{-1}\cdot a\cdot d\cdot d^{-1}\cdot c) \in N$ as desired. For the remaining implication, suppose $(b^{-1}\cdot a) \in N$. Since **N** is a subgroup we can apply the inverse to get $(a^{-1}\cdot b) \in N$. Now invoking the normality of **N** we see $b\cdot a^{-1} = b\cdot(a^{-1}\cdot b)\cdot b^{-1} \in N$ as desired. □

The import of this theorem is that we can use normal subgroups and congruences interchangeably, just as we used ideals and congruences interchangeably in ring theory. When h is a homomorphism from the group **G** we will call $\{a \mid a \in G \text{ and } h(a) = 1\}$ the **kernel** of h. Of course, it is a normal subgroup and in fact, the normal subgroups of **G** are exactly the kernels of homomorphisms from **G**.

Let **N** be a normal subgroup of the group **G**. What are the congruence classes associated with **N**? We know that N itself is the congruence class containing the element 1. What about the others? Let $a \in G$. The congruence class containing a is evidently

$$\{b \mid b \in G \text{ and } (a^{-1} \cdot b) \in N\}.$$

Let $aN = \{a{\cdot}c \mid c \in N\}$. Then it is clear that $(a^{-1}{\cdot}b) \in N$ if and only if $b \in aN$. This means that the congruence class containing a is

$$\{b \mid b \in G \text{ and } (a^{-1} \cdot b) \in N\} = \{b \mid b \in G \text{ and } b \in aN\} = aN.$$

We observe that this little line of reasoning remains true even if **N** is only a subgroup of **G** (of course we should only say "equivalence class" in this case). Sets of the form aN are called **(left) cosets** of **N**. Right cosets I leave to your imagination. Here is the remarkable thing discovered by Lagrange in 1771. (Actually, Euler may have made the same discovery a decade earlier. One difficulty with these attributions is that the concept of a group itself did not emerge in a fully explicit form until almost a century later.)

Lagrange's Theorem. *Let* **H** *be a subgroup of the group* **G**. *All the cosets of* **H** *have the same cardinality. In particular,* $|H|$ *divides* $|G|$. *Moreover, if* **K** *is a homomorphic image of* **G**, *then* $|K|$ *also divides* $|G|$.

Proof. Let a be an arbitrary element of G. We need only exhibit a one-to-one correspondence from H onto aH. Just define $\Phi : H \to aH$ via

$$\Phi(b) = ab \text{ for all } b \in H.$$

This map is onto aH by the very definition of aH. It is one-to-one since $a \cdot b = a \cdot c \Rightarrow b = c$, since the cancellation law works in all groups. The first divisibility statement works because we know that G is partitioned into the cosets of H—that is, G is the disjoint union of some number of sets all of size $|H|$. The second divisibility statement follows since if **H** is the kernel of a homomorphism from **G** onto **K**, then by the Homomorphism Theorem $|K|$ must be the number of distinct cosets of **H**. □

Lagrange's Theorem (well this is only one of his theorems...) holds for all groups, finite or infinite, but it was the first key tool for dealing with finite groups. So a group of size 21 cannot have a subgroup of size 6 and the sizes of its homomorphic images can be found only among $1, 3, 7$, and 21.

Let **G** be a group and let **H** be a subgroup. We put

$$[\mathbf{G} : \mathbf{H}] = |\{aH \mid a \in G\}|.$$

That is $[\mathbf{G} : \mathbf{H}]$ is the number of left cosets of **H** in **G**. It is called the **index** of **H** in **G**. Another way to frame Lagrange's Theorem is to assert, for **H** a subgroup of **G** and $h : \mathbf{G} \twoheadrightarrow \mathbf{L}$, that

$$|G| = [\mathbf{G} : \mathbf{H}]|H| = |L||\ker h|.$$

It is evident that everywhere in the above discussion, if we were to replace *left* coset by *right* coset the result would be entirely similar. In particular, The number of left cosets of **H** is the same as the number of right cosets of **H**. This does not mean that every left coset is a right coset (and vice versa). The demonstration of the fact below is left to the hard-working graduate students.

Fact. Let **H** be a subgroup of the group **G**. Then **H** is a normal subgroup of **G** if and only if every left coset of **H** is **G** is a right coset of **H** in **G**.

13.3 Problems Set 13

Algebra Homework, Edition 13
Fourteenth Week
Some little problems about groups

Problem 57.
Derive a list of equations that follow from the equations axiomatizing the theory of groups. This is rather open ended, but see if you can get a handful of useful looking equations.

Problem 58.
The five equations used to axiomatize groups are not all needed. Find a simpler set of equations that will serve.

Problem 59.
Prove that the additive group of all polynomials in x with integer coefficients is isomorphic to the multiplicative group of all postive rational numbers.

Problem 60.
Let $\mathbf{A}$ and $\mathbf{B}$ be groups and let $f : A \to B$. Prove that f is a homomoprhism if and only if $f(aa') = f(a)f(a')$ for all $a, a' \in A$.

Problem 61.
Let $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$ be groups. Let h be a homomorphism from $\mathbf{A}$ onto $\mathbf{B}$ and let g be a homomorphism from $\mathbf{A}$ onto $\mathbf{C}$ such the $\ker h = \ker g$.

Prove that there is an isomorphism f from $\mathbf{B}$ onto $\mathbf{C}$.

Problem 62.

(a) Let $\mathbf{G}$ be a subgroup of the symmetric group $\mathbf{S}_n$. Show that if G contains an odd permutation, then $\mathbf{G} \cap \mathbf{A}_n$ is of index 2 in $\mathbf{G}$.

(b) Let $\mathbf{G}$ be a group of order $2t$, where t is odd, and consider the homomorphism $\tau : \mathbf{G} \to \mathbf{Sym}(G)$, given by $\tau(g)(h) = gh$, for $h \in G$. Show that if $g^2 = 1$, then $\tau(g)$ is an odd permutation.

(c) Deduce that a group of order $2t$, where t is odd, cannot be simple.

LECTURE 14

ISOMORPHISM THEOREMS: THE GROUP VERSIONS

Last semester we learned a collection of theorems in the general context. Seeing in the previous lecture that, in group theory, we can replace congruences by normal subgroups, in this lecture we simply present the corresponding specializations without further proof.

The Homomorphism Theorem (Group Version). *Let* $\mathbf{A}$ *be a group, let* $f : \mathbf{A} \twoheadrightarrow \mathbf{B}$ *be a homomorphism from* $\mathbf{A}$ *onto* $\mathbf{B}$*, and let* $\mathbf{N}$ *be a normal subgroup of* $\mathbf{A}$*. All of the following hold.*

(a) *The kernel of f is a normal subgroup of* $\mathbf{A}$.

(b) $\mathbf{A}/\mathbf{N}$ *is a group.*

(c) *The map η that assigns to each $a \in A$ the left coset aN is a homomorphism from* $\mathbf{A}$ *onto* $\mathbf{A}/\mathbf{N}$ *and its kernel is* $\mathbf{N}$.

(d) *If* $\mathbf{N}$ *is the kernel of f, then there is an isomorphism g from* $\mathbf{A}/\mathbf{N}$ *to* $\mathbf{B}$ *such that $f = g \circ \eta$.*

The Second Isomorphism Theorem (Group Version).

theorem!Second Isomorphism Theorem (Group Version) Let $\mathbf{A}$ *be a group, let* $\mathbf{N}$ *be a normal subgroup of* $\mathbf{A}$*, and let* $\mathbf{B}$ *be a subgroup of* $\mathbf{A}$*. Then each of the following holds.*

(a) $\mathbf{N} \cap \mathbf{B}$ *is a normal subgroup of* $\mathbf{B}$.

(b) $\mathbf{NB}$ *is a subgroup of* $\mathbf{A}$ *and* $\mathbf{N}$ *is a normal subgroup of* $\mathbf{NB}$.

(c) $\mathbf{NB}/\mathbf{N} \cong \mathbf{B}/\mathbf{N} \cap \mathbf{B}$.

The Third Isomorphism Theorem (Group Version). *Let* $\mathbf{A}$ *be a group and let* $\mathbf{N}$ *and* $\mathbf{K}$ *be normal subgroups of* $\mathbf{A}$ *with $N \subseteq K$. Then*

(a) $\mathbf{K}/\mathbf{N}$ *is a normal subgroup of* $\mathbf{A}/\mathbf{N}$*, and*

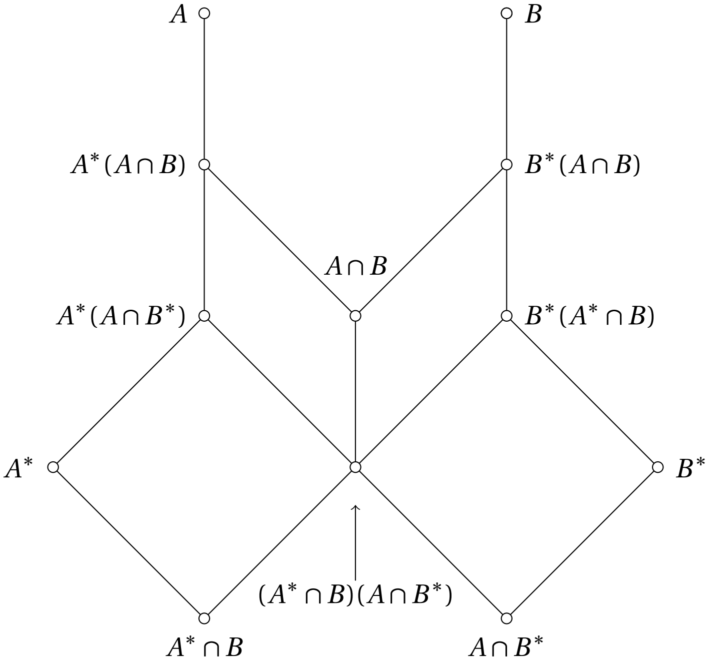
(b) $\mathbf{A}/\mathbf{N} \big/ \mathbf{K}/\mathbf{N} \cong \mathbf{A}/\mathbf{K}$.

The Correspondence Theorem (Group Version). *Let* $\mathbf{A}$ *be a group and let* $\mathbf{N}$ *be a normal subgroup of* $\mathbf{A}$. *Let* $P = \{\mathbf{K} \mid \mathbf{K} \lhd \mathbf{A}$ *and* $N \subseteq K\}$. *Then the map from* P *to set of normal subgroups of* $\mathbf{A}/\mathbf{N}$ *that sends each* $\mathbf{K} \in P$ *to* $\mathbf{K}/\mathbf{N}$ *is an isomorphism between the ordered set* $\langle P, \subseteq\rangle$ *and the set of normal subgroups of* $\mathbf{A}/\mathbf{N}$ *ordered by set inclusion.*

Just as the corresponding special cases for rings were handy in the development of the theory of rings, these special cases for groups will play a similar role. Here is an interesting conclusion that comes from putting Lagrange's Theorem and the Second Isomorphism Theorem together.

Corollary 14.0.1. *Let* $\mathbf{A}$ *be a finite group and let* $\mathbf{B}$ *be a subgroup of* $\mathbf{A}$ *and let* $\mathbf{N}$ *be a normal subgroup of* $\mathbf{A}$. *Then* $|NB||N \cap B| = |N||B|$. *In particular if* $\mathbf{N}$ *and* $\mathbf{B}$ *have only the identity element in common, that* $|NB| = |N||B|$.

We conclude this lecture with a more involved theorem for groups that is due to Hans Zassenhaus. We will use this theorem later. This theorem is called the Zassenhuas's Butterfly Lemma. The name is suggested by the following picture, which displays part of the lattice of subgroups of some group $\mathbf{G}$.



The Butterfly of Hans Zassenhaus

The Butterfly Lemma of Hans Zasssenhaus. *Let* $\mathbf{G}$ *be a group with subgroups* $\mathbf{A}, \mathbf{A}^*, \mathbf{B}$ *and* $\mathbf{B}^*$, *where* $\mathbf{A}^*$ *is a normal subgroup of* $\mathbf{A}$ *and* $\mathbf{B}^*$ *is a normal subgroup of* $\mathbf{B}$. *Then all of the following hold.*

(a) $\mathbf{A}^*(\mathbf{A}\cap\mathbf{B}^*) \lhd \mathbf{A}^*(\mathbf{A}\cap\mathbf{B})$.

(b) $\mathbf{B}^*(\mathbf{A}^*\cap\mathbf{B}) \lhd \mathbf{B}^*(\mathbf{A}\cap\mathbf{B})$.

(c) $\mathbf{A}^*(\mathbf{A}\cap\mathbf{B})/\mathbf{A}^*(\mathbf{A}\cap\mathbf{B}^*) \cong \mathbf{B}^*(\mathbf{A}\cap\mathbf{B})/\mathbf{B}^*(\mathbf{A}^*\cap\mathbf{B})$.

Proof. Because $\mathbf{B}^*$ is a normal subgroup of $\mathbf{B}$ it is easy to see that $\mathbf{A}\cap\mathbf{B}^* \lhd \mathbf{A}\cap\mathbf{B}$. Likewise, $\mathbf{A}^*\cap\mathbf{B}\lhd\mathbf{A}\cap\mathbf{B}$. So we can also conclude that $(\mathbf{A}^*\cap\mathbf{B})(\mathbf{A}\cap\mathbf{B}^*)\lhd\mathbf{A}\cap\mathbf{B}$. Here is why. First we know that the product of a normal subgroup and a subgroup is again a subgroup. (This point came up in the Second Isomorphism Theorem—hard working graduate students checked it then....) So we find that $(\mathbf{A}^*\cap\mathbf{B})(\mathbf{A}\cap\mathbf{B}^*)$ is a subgroup of $\mathbf{A}\cap\mathbf{B}$. For normality, pick $c \in A^*\cap B$, $d \in A\cap B^*$, and $a \in A\cap B$. Then notice

$$a(cd)a^{-1} = aca^{-1}ada^{-1} = (aca^{-1})(ada^{-1}).$$

Since $aca^{-1} \in A^*\cap B$, by normality of $\mathbf{A}^*\cap\mathbf{B}$ in $\mathbf{A}\cap\mathbf{B}$, and likewise $ada^{-1}\in A\cap B^*$, we see that

$$a(cd)a^{-1} \in (\mathbf{A}^*\cap\mathbf{B})(\mathbf{A}\cap\mathbf{B}^*).$$

Let $\mathbf{D}$ denote $(\mathbf{A}^*\cap\mathbf{B})(\mathbf{A}\cap\mathbf{B}^*)$. So we have $\mathbf{D}\lhd\mathbf{A}\cap\mathbf{B}$.

Define the map $f : A^*(A\cap B) \to (A\cap B)/D$ in the following way. Let $a\in A^*$ and $c \in A\cap B$. Put $f(ac) = cD$. First we need to see that we can get away with this. Suppose $a_o \in A^*$ and $c_o \in A\cap B$ so that $ac = a_oc_o$. We need $cD = c_oD$ or what is the same $c_oc^{-1}\in D$. We certainly get $a_o^{-1}a = c_0c^{-1}$. The left side is in A^* and the right in $A\cap B$. Since they are equal, we see that $c_0c^{-1} \in A^*\cap A\cap B = A^*\cap B \subseteq (A^*\cap B)(A\cap B^*) = D$. So our definition of the map f is sound.

It is evident from the definition that f maps onto $(A\cap B)/D$. We aim to show that f is a homomorphism with kernel $\mathbf{A}^*(\mathbf{A}\cap\mathbf{B}^*)$. Then we can appeal to the Homomorphism Theorem to obtain an isomorphism. Reversing the roles of A^* and B^* we can obtain a second isomorphism. Composing one with the inverse of the other gets us the isomorphism we desired in the statement of the lemma.

We need to see that f preserves products. Let $a, a_o \in A^*$ and $c, c_o \in A\cap B$. Then

$$f((ac)(a_oc_o)) = f(aca_oc^{-1}cc_o) = f(aa_o'cc_o) = cc_oD = cDc_oD = f(ac)f(a_oc_o).$$

Observe the appeal to normality of $\mathbf{A}^*$ in $\mathbf{A}$. So we find that f is a homomorphism.

Last, we need to understand the kernel of f.

$$ac \in \ker f \Leftrightarrow f(ac) = 1D \Leftrightarrow cD = D \Leftrightarrow c\in D \Leftrightarrow c = de \text{ for some } d\in A^*\cap B \text{ and some } e\in A\cap B^*.$$

This means

$$ac\in\ker f \Leftrightarrow ac = (ad)e \text{ for some } d\in A^*\cap B \text{ and some } e \in A\cap B^* \Leftrightarrow ac\in A^*(A\cap B^*).$$

Hence, we find $\ker f = \mathbf{A}^*(\mathbf{A}\cap\mathbf{B}^*)$ and our proof of the Butterfly Lemma is complete. □

14.1 Problem Set 14

Algebra Homework, Edition 14

Fifteenth Week

Subgroups and homomorphic images of groups

Problem 63.
Prove that every group that has a proper subgroup of finite index must have a proper normal subgroup of finite index.

Problem 64.
Let **G** be a group. Prove that **G** cannot have four distinct proper normal subgroups $\mathbf{N}_0, \mathbf{N}_1, \mathbf{N}_2$, and $\mathbf{N}_3$ so that $\mathbf{N}_0 \leq \mathbf{N}_1 \leq \mathbf{N}_2 \leq \mathbf{G}$ and so that $N_1 N_3 = G$ and $N_2 \cap N_3 = N_0$.

Problem 65.
Let **H** and **K** be subgroups of the group **G** each of finite index in **G**. Prove that $\mathbf{H} \cap \mathbf{K}$ is also a subgroup of finite index in **G**.

Problem 66.
Let p be a prime number. Prove that if a and b are elements of the symmetric group S_p, where a has order p and b is a transposition, then $\{a, b\}$ generates S_p.

Problem 67.
Show that there is no group that has exactly one subgroup that is not a normal subgroup.

LECTURE 15

USEFUL FACTS ABOUT CYCLIC GROUPS

Let **G** be a group and let $X \subseteq G$. We use $\langle X\rangle$ to denote, ambiguously, both the subgroup of **G** generated by X and the underlying universe of that subgroup. So we can construe $\langle X\rangle$ has the intersection of all the subgroups of **G** that include X (the shrinkwrap viewpoint) or as the set of all elements of G that can be built from the elements of X by iteratively applying the basic operations of the group **G**.

The group **G** is **cyclic** provided there is some $a \in G$ so that $G = \langle\{a\}\rangle$. That is, **G** is generated by some single element. To save notation, we write $\langle a\rangle$ for $\langle\{a\}\rangle$. Taking $a^{-n} := (a^{-1})^n$ for every natural number, we see that

$$\langle a\rangle = \{a^r \mid r \in \mathbb{Z}\}.$$

It is easy to see that in any group, the equation $x^r x^s = x^{r+s}$ must be true, where r and s are any integers. From this we get

Fact. Every cyclic group is Abelian.

For any group **G** and any $a \in G$, we let **order** of a, denoted by $o(a)$, be $|\langle a\rangle|$.

Fact. The order of any element of a group is either countably infinite or it is finite and not 0.

Fact. Let **G** be a group and $a \in G$. The element a has finite order n if and only if n is the smallest positive natural number such that $a^n = 1$ in **G**.

Proof. Let us first consider the case when $a = 1$. Then $\langle a\rangle = \{1\}$, a set with 1 element and $n = 1$ is also the least positive natural number so that $1^n = 1$. So in the remainder of this proof we consider the case when $a \neq 1$.

First suppose that a has finite order n. Then $\{a^r \mid r \in \mathbb{Z}\}$ is finite. So pick integers $k < \ell$ so that $a^k = a^\ell$. It follows that $a^{\ell-k} = 1$ and $\ell - k > 0$. Pick m to be the least positive natural number so that $a^m = 1$. So the elements $1 = a^0, a^1, \dots, a^{m-1}$ must all be distinct. This set evidently contains 1 and it is closed under inverses since $a^k a^{m-k} = a^{k+m-k} = a^m = 1$, for all $k < m$. It is also closed under products since for $k, \ell < m$ we have $a^k a^\ell = a^{k+\ell} = a^{mq+r} = (a^m)^q a^r = 1^q a^r = a^r$, where q and r are the unique integers such that

$$k + \ell = mq + r \qquad \text{where } 0 \leq r < m.$$

So $\langle a\rangle = \{1, a, a^2, \dots, a^{m-1}\}$. In this way we see that $n = m$, as desired.

For the converse, suppose that n is the least positive integer so that $a^n = 1$. Then we have just shown that $\langle a\rangle = \{1, a, a^2, \dots, a^{n-1}\}$ and that the elements listed are distinct. So n is the order of a. □

The proof above carries a bit more information.

Fact. A finite cyclic group of order n is isomorphic to the group $\mathbb{Z}_n = \langle\{0,1,\ldots,n-1\},+_n,-_n.0\rangle$ where the operations work modulo n.

Proof. Let $\langle a\rangle$ be the cyclic group of order n. The elements of this group are $a^0,a^1,a^2,\ldots,a^{n-1}$. As shown above, the operations work like this for all natural numbers $k,\ell < n$

$$\begin{aligned}(a^k)^{-1} &= a^{n-k}\\ a^k a^\ell &= a^r \qquad \text{where } 0 \le r < n \text{ and } k+\ell \equiv r \pmod{n}.\end{aligned}$$

Now just observe that the "logarithm" that sends $a^k \mapsto k$ for all natural numbers $k < n$ is an isomorphism from $\langle a\rangle$ onto $\mathbb{Z}_n$. □

The same sort of logarithm function applies to infinite cyclic groups, giving the following

Fact. Every infinite cyclic group is isomorphic to the additive group of integers, that is to $\mathbb{Z} := \langle\mathbb{Z},+,-,0\rangle$.

Fact. Let $\mathbf{G}$ be a group and let $a \in G$ have finite order n. If $a^m = 1$ in $\mathbf{G}$ then $n \mid m$.

Proof. Let q and r be the unique integers such that

$$m = nq + r \qquad \text{where } 0 \le r < n.$$

Then $1 = a^m = a^{nq+r} = (a^n)^q a^r = 1^q a^r = a^r$. Since n is the order of a and $0 \le r < n$, we must have $r = 0$. Thus $m = nq$ and $n \mid m$. □

Fact. Every subgroup of a cyclic group is cyclic.

Proof. Let $\mathbf{H}$ be a subgroup of the cyclic group $\mathbf{G}$ and let a be an element of G which generates $\mathbf{G}$. As the trivial group is cyclic, we will consider the remaining case that $\mathbf{H}$ is nontrivial. Let k be the least positive natural number so that $a^k \in H$. We see that k must be strictly smaller than $o(a)$. Our contention is that a^k generates $\mathbf{H}$. So let a^ℓ be an arbitrary element of H. Let q and r be the unique integers so that

$$\ell = kq + r \qquad \text{where } 0 \le r < k.$$

Then $a^\ell = (a^k)^q a^r$. Now since $a^k \in H$, then so is $(a^k)^{-q}$. But $a^\ell \in H$. Hence $a^r = a^\ell (a^k)^{-q} \in H$. Since $0 \le r < k$, we see that $r = 0$ by the minimality of k. Hence $a^\ell = (a^k)^q$ and so a^ℓ is in the subgroup generated by a^k. Since a^ℓ was an arbitrary element of H, we see that $\mathbf{H}$ is generated by a^k and therefore that $\mathbf{H}$ is cyclic. □

The next fact provides a remarkable property that cyclic groups possess that is not common even among Abelian groups.

Fact. Let $\mathbf{G}$ be a cyclic group of finite order n and let k be a natural number so that $k \mid n$. Then $\mathbf{G}$ has exactly one subgroup of order k. Moreover, $\mathbf{G}$ has no other subgroups.

Proof. Let a generate $\mathbf{G}$ and let m be the natural number so that $km = n$. Then the order of a^m is the least ℓ so that $a^{m\ell} = 1$. Since $mk = n$, we see that $\ell \le k$. But also $mk = n \le m\ell$. Hence $k \le \ell$. Thus $k = \ell$ is the order of a^m. This means that $\mathbf{G}$ has at least one subgroup of order k, namely $\langle a^m\rangle$. Since every subgroup of $\mathbf{G}$ is cyclic, let us suppose that a^j generates a subgroup of order k. That is, a^j has order k. Pick the integers q and r so that

$$j = mq + r \qquad \text{where } 0 \le r < m.$$

Now we know that k is the least positive integer so that $a^{jk} = 1$, so we see

$$1 = a^{jk} = a^{mkq+rk} = (a^{mk})^q a^{rk} = (a^n)^q a^{rk} = 1^q a^{rk} = a^{rk}.$$

But since $0 \le r < m$ we have $0 \le rk < mk = n$. Since n is the order of a, we see that $r = 0$. But then $j = mq$. Hence $a^j = (a^m)^q$. This means that $a^j \in \langle a^m \rangle$. Hence $\langle a^j \rangle \subseteq \langle a^m \rangle$. But both of these sets are finite and have the same cardinality. So they must be equal, as desired.

That **G** has no other subgroups is immediate by Lagrange. □

The next couple of facts deal with Abelian groups and will help us distinguish which Abelian groups are actually cyclic.

Fact. Let **G** be an Abelian group and let $a, b \in G$. If the orders of a and b are finite and relatively prime, then $o(ab) = o(a)o(b)$.

Proof. Suppose $(ab)^k = 1$. Then $a^k = b^{-k} \in \langle a \rangle \cap \langle b \rangle$. So according to Lagrange, the order of a^k must divide the order of a and also the order of b. These orders are relatively prime, so the order of a^k must be 1. That is $a^k = (a^k)^1 = 1$. A similar argument gives that $b^k = 1$. So we have both $o(a) \mid k$ and $o(b) \mid k$. Since $o(a)$ and $o(b)$ are relatively prime we see $o(a)o(b) \mid k$. But it is easy to verify (as hard working graduate students will) that $(ab)^{o(a)o(b)} = 1$. So we see $o(ab) = o(a)o(b)$. □

Let **G** be a group. The **exponent of G** is the least positive integer e so that $a^e = 1$ for all $a \in G$. Every finite group has an exponent. Certain infinite groups also have exponents, but most do not.

Fact. Let **G** be an Abelian group and suppose that a is an element of largest order and that order is finite. Then the exponent of **G** is the order of a.

Proof. Let b be an arbitrary element of G and n be the order of a. We need only show that $b^n = 1$. Now we know that the order of b is bounded above by n, so in particular it is finite. Let it be m. Now factor n and m:

$$\begin{aligned} n &= p_0^{e_0} p_1^{e_1} \dots p_k^{e_k} \\ m &= p_0^{f_0} p_1^{f_1} \dots p_k^{f_k} \end{aligned}$$

where $p_0, \dots, p_k$ are distinct primes and the e_j's and f_j's are natural numbers. In the event that $m|n$ we have $b^n = 1$ as desired.

So consider the case when $m \nmid n$. Then there is some $j \le k$ so that $f_j > e_j$. Without loss of generality (and to simplify notation) let $j = 0$.

Now put $c = a^{p_0^{e_0}}$ and $d = b^{p_1^{f_1} \dots p_k^{f_k}}$. Then

$$\begin{aligned} o(c) &= p_1^{e_1} \dots p_k^{e_k} \\ o(d) &= p_0^{f_0} \end{aligned}$$

This means that the orders of c and d are relatively prime. It follows that the order of cd is $p_0^{f_0} p_1^{e_1} \dots p_k^{e_k}$. But this is larger than n contrary to the maximality of the order of a. So we must reject this case. □

Fact. Let **G** be a finite Abelian group. **G** is cyclic if and only if $|G|$ is the exponent of **G**.

The proof is immediate from the two preceding facts.

Let $\varphi(n)$ be Euler's totient function. That is, $\varphi(n)$ is the number of natural numbers less than n that are relatively prime to n.

Fact. Let **G** be a finite cyclic group of order n. Then **G** has precisely $\varphi(n)$ elements of order n and they are those of the form a^m where a generates **G** and m is relatively prime to n with $0 \leq r < n$.

Proof. First, suppose that m satisfies the listed conditions. Let k be the order of a^m. Then k is the least positive natural number such that $1 = (a^m)^k = a^{mk}$. So $n \mid mk$. But since m and n are relatively prime, we find that $n \mid k$. On the other hand, Lagrange tells us that $k \mid n$. Thus $k = n$. So a^m indeed has order n.

Now suppose a^m has order n and $0 \leq m < n$. Let d be the greatest common divisor of m and n. Let s be the natural number so that $ds = n$ and let t be the natural number so that $dt = m$. Then $(a^m)^s = a^{dts} = (a^{ds})^t = (a^n)^t = 1^t = 1$. Since n is the order of a^m we find that $n \mid s$. On the other hand, $n = ds$. So $n = s$ and $d = 1$. Since $d = 1$ we conclude that m and n are relatively prime. □

Fact. Let **G** and **H** be finite cyclic groups of order n. Then $\varphi(n)$ is the number of isomorphisms from **G** onto **H**.

Proof. We already observed that each of these groups is isomorphic to $\mathbb{Z}_n$, so there are certainly isomorphisms between them. Let a generate **G**. Now any isomorphism must preserve the order of elements and so it must take a to a generator of **H**. Suppose b is a generator of **H**. Now in our proof that these cyclic groups were isomorphic to $\mathbb{Z}_n$ we use logarithm maps. Composing the logarithm map from **G** to $\mathbb{Z}_n$ with the inverse of the logarithm map from **H** to $\mathbb{Z}_n$ we obtain an isomorphism from **G** onto **H** that sends $a^k \mapsto b^k$ for all natural numbers $k < n$. Since there are $\varphi(n)$ choices for b, we have found $\varphi(n)$ distinct isomorphisms from **G** onto **H**. Are there anymore?

Suppose Φ is an isomorphism from **G** onto **H**. Then $\Phi(a) = b$ is a generator of **H**. Moreover, we have $\Phi(a^k) = (\Phi(a))^k = b^k$ for all natural numbers $k < n$. So Φ is one of the isomorphisms counted in the previous paragraph. □

Let $U_n = \{m \mid m \text{ is a natural number relatively prime to } n \text{ and } m < n\}$. Notice that $1 \in U_n$. By imposing multiplication modulo n and the correct inversion we make a group $\mathbf{U}_n$. (In finding the inverse, the hard-working graduate students should consider that the relative primeness of m and n leads to $1 = ms + nq$ for appropriate integers s and t.) We leave the following, which is a corollary of the fact above as a challenge to the graduate students. We use Aut **G** to denote the set of all automorpisms of the group **G**. This set contains the identity map and is closed under composition of functions and the formation of inverse functions. So we can turn it into a group, denoted, of course, by **Aut G**.

Fact. Let **G** be a finite cyclic group of order n. Then $\mathbf{Aut\,G} \cong \mathbf{U}_n$.

Finally, here is a theorem of Euler.

Fact. For all postive natural numbers m and n that are relatively prime, we have $m^{\varphi(n)} \equiv 1 \pmod{n}$.

Proof. First of all, we may insist that $m < n$. The reason is that the modulo n residue map is a homomorphism from the ring of integers onto the ring of integers modulo n. So $m \in U_n$. By Lagrange, the order of m divides the cardinality of U_n which is $\varphi(n)$. So the desired conclusion follows. □

Lecture 16

Group Representation: Groups Acting on Sets

Let $\mathbf{G}$ be a group and let X be a set. By an **action** of $\mathbf{G}$ on X we just mean a homomorphism $\Phi : \mathbf{G} \to \mathbf{Sym}\, X$.

Cayley's Theorem gave us one example of $\mathbf{G}$ acting on G. Recall that there $\Phi_a(b) := ab$ for all $a, b \in G$. This action is sometimes called the action of $\mathbf{G}$ on G by translation (on the left). We showed this Φ is one-to-one. One-to-one actions are said to be **faithful**.

Lagrange's Theorem also suggests an action. Let $\mathbf{G}$ be a group and let $\mathbf{H}$ be a subgroup of $\mathbf{G}$. Let X be the collection of left cosets of $\mathbf{H}$ in $\mathbf{G}$. We can have the action Φ defined so that $\Phi_a(bH) := (ab)H$ for all $a, b \in G$.

Here is another action. Let $\mathbf{G}$ be a group. Let X be G and define Φ so that $\Phi_a(b) := a^{-1}ba$ for all $a, b \in G$. Of course, one must actually show that this Φ is a homomorphism from $\mathbf{G}$ into $\mathbf{Sym}\, G$. Of course, this will be verified by the hard-working. . . . This is the action of $\mathbf{G}$ on G by conjugation.

Roughly speaking, the idea of representations is that by exploring (a number of different) concrete homomorphic images of a group we might find out more about the group itself. By analogy, think of the homomorphic images has shadows onto a two-dimensional screen of some three-dimensional object. By understanding enough of the shadows we might be able to reconstruct what the object looks like.

The language using the homomorphism Φ can be streamlined. This streamlining can sometimes be ambiguous, but usually it is not. What Φ does is associate with each $g \in G$ a permutation Φ_g of X. The first step in the streamlining is to just regard g as a name for this permutation—in other words, drop the Φ and raise the g. This means we get things like the following for all $x \in X$ and all $g, h \in G$

$$\begin{aligned} 1(x) &= \mathbf{1}_X(x) = x \\ (gh)(x) &= (g \circ h)(x) = g(h(x)) \end{aligned}$$

The last step in the streamlining process is to drop a set of parentheses—that is, to write gx in place of $g(x)$. Then the equations above become, for all $x \in X$ and all $g, h \in G$,

$$\begin{aligned} 1x &= x \\ (gh)x &= g(hx) \end{aligned}$$

You might notice that the formation of inverses is not mentioned above. This is legitimate since we know that both $\mathbf{G}$ and $\mathbf{Sym}\, X$ are groups. Sometimes authors say that an action of $\mathbf{G}$ on X is a kind of "scalar" multiplication of group elements by elements of X that satisfies the two equations above. This amounts to the same thing since the map Φ can be recovered from this information and it can be shown to be a homomorphism.

Let Φ be an action of **G** on a set X and let $x \in X$. The **orbit** of x under the action is the set

$$\mathcal{O}_x := \{\Phi_g(x) \mid g \in G\}.$$

In the streamlined notation this becomes

$$\mathcal{O}_x := \{gx \mid g \in G\}.$$

Observe that $x \in \mathcal{O}_x$ since $1 \in G$. Also notice that If $y \in \mathcal{O}_x$, then $\mathcal{O}_x = \mathcal{O}_y$. Here is why. Let $g \in G$ so that $gx = y$. Now observe

$$\begin{aligned} z \in \mathcal{O}_y &\Leftrightarrow hy = z \text{ for some } h \in G \\ &\Leftrightarrow h(gx) = z \text{ for some } h \in G \\ &\Leftrightarrow (hg)x = z \text{ for some } h \in G \\ &\Leftrightarrow kx = z \text{ for some } k \in G (\text{ careful!}) \\ &\Leftrightarrow z \in \mathcal{O}_x. \end{aligned}$$

Hence $\mathcal{O}_y = \mathcal{O}_x$. Thus the orbits of any elements $x, y \in X$ either coincide or they are disjoint. This means that X is partitioned into orbits by the action of **G**.

Now let $x \in X$. The **stablizer** of x with respect to the action Φ is the following set

$$\operatorname{Stab} x := \{g \mid g \in G \text{ and } gx = x\}.$$

That is, the stablizer of x consists of all the elements of G that leave x fixed under the action. It is easy to see that $\operatorname{Stab} x$ is closed under the group operations:

$$\begin{gathered} 1x = x \text{ so } 1 \in \operatorname{Stab} x \\ gx = x \text{ and } hx = x \Rightarrow (gh)x = x \text{ so } \operatorname{Stab} x \text{ is closed under products} \\ gx = x \Rightarrow x = g^{-1}x \text{ so } \operatorname{Stab} x \text{ is closed under inverses.} \end{gathered}$$

In this way we arrive at the group **Stab** x, which is a subgroup of **G**.

Here is the

Key Fact About Group Actions. *Let the group* **G** *act on the set* X*. Then we have* $|\mathcal{O}_x| = [\mathbf{G} : \mathbf{Stab}\, x]$ *for all* $x \in X$.

Proof. Notice that $[\mathbf{G} : \mathbf{Stab}\, x]$ is the number of left cosets of $\operatorname{Stab} x$ in **G**. To prove the Key Fact we present a one-to-one correspondence between $\mathcal{O}_x$ and the collection of left cosets of $\operatorname{Stab} x$. As preparation, suppose $y \in \mathcal{O}_x$. Then there is at least one $g \in G$ so that $gx = y$. Suppose also $h \in G$ and $hx = y$. Then, of course $gx = hx$ and so $(h^{-1}g)x = x$. This means that $h^{-1}g \in \operatorname{Stab} x$ or, what is the same, g and h are in the same left coset of $\operatorname{Stab} x$. This allows us to define Ψ from $\mathcal{O}_x$ to the collection of left cosets of $\operatorname{Stab} x$ as follows:

$$\Psi(y) := g \operatorname{Stab} x \text{ where } gx = y.$$

This definition works for all $y \in \mathcal{O}_x$. It remains to show that Ψ is the desired one-to-one correspondence.

To see one-to-oneness, let $y, z \in \mathcal{O}_x$ with $\Psi(y) = \Psi(z)$. Pick $g, h \in G$ so that $gx = y$ and $hx = z$. So we get $g \operatorname{Stab} x = h \operatorname{Stab} x$. This means $h^{-1}g \in \operatorname{Stab} x$. So $(h^{-1}g)x = x$. But then $gx = hx$. So we find $y = gx = hx = z$, and conclude that Ψ is one-to-one.

To see that Ψ maps $\mathcal{O}_x$ onto the collection of cosets of $\operatorname{Stab} x$, let $g \in G$. We must find $y \in \mathcal{O}_x$ so that $\Psi(y) = g \operatorname{Stab} x$. Let us try $y = gx$. It works, enough said. □

Let the group **G** act on the set X. By a **transversal** for this action we mean a set $T \subseteq X$ so that $|T \cap \mathcal{O}_x| = 1$ for each $x \in X$. This means that T is constituted by picking one "representative" element from each orbit. The next fact is a corollary of the Key Fact.

Fact. Let the group **G** act of the set X and let T be a transversal for this action. Then

$$|X| = \sum_{t \in T} [\mathbf{G} : \mathbf{Stab}\, t].$$

There are some interesting consequences when these notions are applied to the action of **G** on G by conjugation. Under this action

$$\Phi_g(h) := g^{-1}hg$$

for all $g, h \in G$. (This is one instance where our streamlining is unreasonable.) Our first remark is that conjugation by g is not only a permutation of G, but is, in fact, an automorphism of **G**. Just observe that $g^{-1}(hk)g = g^{-1}hgg^{-1}kg = (g^{-1}hg)(g^{-1}kg)$. This means that $\Phi : \mathbf{G} \to \mathrm{Aut}\,\mathbf{G}$. Automorphisms of **G** that are conjugations by some fixed element g are called **inner automorphisms**. We see that since they constitute the image of **G** under the homomorphism Φ, the inner automorphisms of **G** form a subgroup of **Aut G**, which is in turn a subgroup of **Sym** G. The group of inner automorphisms of **G** is denoted by **Inn G**. What is the kernel of Φ? Well, $g \in \ker \Phi$ if and only if Φ_g is $\mathbf{1}_G$ if and only if $g^{-1}hg = h$ for all $h \in G$ if and only if $hg = gh$ for all $h \in G$. This means

$$\ker \Phi = \{g \mid hg = gh \text{ for all } h \in G\}.$$

This group is call the **center** of **G** and consist of all elements of G that commute with every element of G. We use $\mathbf{Z}(\mathbf{G})$ to denote the center of the group **G**. The next fact merely gathers together these findings.

Fact. Let **G** be a group. Then the center $\mathbf{Z}(\mathbf{G})$ is a normal subgroup of **G** and $\mathbf{G}/\mathbf{Z}(\mathbf{G}) \cong \mathbf{Inn}(\mathbf{G})$.

Now consider the corollary of the Key Fact, applied to the action by conjugation. We get

$$|G| = \sum_{t \in T} [\mathbf{G} : \mathbf{Stab}\, t].$$

To understand this a little better, look at $\mathbf{Stab}\, t = \{g \mid g \in G \text{ and } g^{-1}tg = t\} = \{g \mid g \in G \text{ and } tg = gt\}$. So under this action Stab t turns out to be the set of those elements of G which commute with t. This set is called the **centralizer** of the element t and is denoted by $C(t)$. So we have

$$C(t) := \{g \mid g \in G \text{ and } tg = gt\}.$$

We know it is a subgroup of **G**, as all stablizers must be. This subgroup is denoted by $\mathbf{C}(t)$. Notice that $C(t) = G$ is equivalent to $t \in Z(\mathbf{G})$ and also to $[\mathbf{G} : \mathbf{C}(t)] = 1$. Now break the transversal T into two disjoint pieces T_0 and T_1, where $t \in T_0$ if and only if $C(t) = G$. Then we get

$$|G| = \sum_{t \in T_0} [\mathbf{G} : \mathbf{C}(t)] + \sum_{t \in T_1} [\mathbf{G} : \mathbf{C}(t)].$$

Now apply the Key Fact to the first sum.

$$|G| = \sum_{t \in T_0} |\mathcal{O}_t| + \sum_{t \in T_1} [\mathbf{G} : \mathbf{C}(t)].$$

Since the orbits are disjoint we get

$$|G| = |\bigcup_{t \in T_0} \mathcal{O}_t| + \sum_{t \in T_1} [\mathbf{G} : \mathbf{C}(t)].$$

But actually $T_0 = Z(\mathbf{G})$ and each $\mathcal{O}_t = \{t\}$ for $t \in T_0$. This means

$$|G| = |Z(\mathbf{G})| + \sum_{t\in T_1} [\mathbf{G} : \mathbf{C}(t)].$$

This equation is called the **Conjugacy Class Equation** or sometimes just the **Class Equation**.

Here is a useful consequence of the Conjugacy Class Equation.

Fact. Every nontrivial finite group of prime power order has a nontrivial center.

Proof. Let p be a prime number and suppose $\mathbf{G}$ is a group of order p^n where n is a positive natural number. The indices $[\mathbf{G} : \mathbf{C}(t)]$ where $t \in T_1$ that occur in the Conjugacy Class Equation are larger than 1 and so by Lagrange each of them is some positive power of p. Thus $p \mid \sum_{t\in T_1} [\mathbf{G} : \mathbf{C}(t)]$ and $p \mid |G|$. Therefore $p \mid |Z(\mathbf{G})|$. So the center of $\mathbf{G}$ must have at least p elements. It is nontrivial. □

The Key Fact tells us something about the sizes of the orbits induced by the action of a group on a set. What about the number of orbits? To get at this, we need a companion to the notion of stablizer. Let $\mathbf{G}$ act on X and let $g \in G$. The **fixed set** of g is

$$\operatorname{Fix} g := \{x \mid x \in X \text{ and } gx = x\}.$$

The Cauchy-Frobenius Formula. *Let the group* $\mathbf{G}$ *act on the set* X *and let* κ *be the number of orbits of this action. Then*

$$\kappa|G| = \sum_{g\in G} |\operatorname{Fix} g|.$$

Proof. Let

$$P := \{(g, x) \mid g \in G \text{ and } x \in X \text{ with } gx = x\}.$$

Observe that for each $x \in X$ we have $(g, x) \in P$ if and only if $g \in \operatorname{Stab} x$. This gives us

$$|P| = \sum_{x\in X} |\operatorname{Stab} x|.$$

Now do the same with the other coordinate. For each $g \in G$ we have $(g, x) \in P$ if and only if $x \in \operatorname{Fix} g$. So we find

$$|P| = \sum_{g\in G} |\operatorname{Fix} g|.$$

So we find $\sum_{g\in G} |\operatorname{Fix} g| = \sum_{x\in X} |\operatorname{Stab} x|$. Now let T be a transversal of the orbits of this action. So $|T| = \kappa$, the number of orbits. But X is the disjoint union of the orbits. So we can rearrange the right-hand sum as follows:

$$\sum_{x\in X} |\operatorname{Stab} x| = \sum_{t\in T} \sum_{x\in \mathcal{O}_t} |\operatorname{Stab} x|.$$

Let $x \in \mathcal{O}_t$. Pick $h \in G$ so that $ht = x$. Then observe that for all $g \in G$ we get $gx = x \Leftrightarrow ght = ht \Leftrightarrow h^{-1}ght = t$. This means that $g \in \operatorname{Stab} x \Leftrightarrow h^{-1}gh \in \operatorname{Stab} t$. But conjugation by h is an automorphism of $\mathbf{G}$, so in particular it follows the subgroups $\mathbf{Stab}\, x$ and $\mathbf{Stab}\, t$ are isomorphic. But we only want that if $x \in \mathcal{O}_t$ then

$|\operatorname{Stab} x| = |\operatorname{Stab} t|$. In this way we arrive at

$$\begin{aligned}
\sum_{g\in G} |\operatorname{Fix} g| &= \sum_{x\in X} |\operatorname{Stab} x| \\
&= \sum_{t\in T} \sum_{x\in \mathcal{O}_t} |\operatorname{Stab} x| \\
&= \sum_{t\in T} \sum_{x\in \mathcal{O}_t} |\operatorname{Stab} t| \\
&= \sum_{t\in T} |\operatorname{Stab} t| \sum_{x\in \mathcal{O}_t} 1 \\
&= \sum_{t\in T} |\operatorname{Stab} t||\mathcal{O}_t| \\
&= \sum_{t\in T} |G| \\
&= |G| \sum_{t\in T} 1 \\
&= |G||T| = \kappa |G|.
\end{aligned}$$

In the above chain of reasoning we use $|G| = |\operatorname{Stab} t||\mathcal{O}_t|$. This is just another way to state the Key Fact. □

We see above (and in the next Lecture) that it is informative to consider homomorphisms $\Phi\colon \mathbf{G} \to \mathbf{Sym}\, X$. This is actually just the first—or maybe second step, counting Cayley first—in a direction that leads more deeply into group theory and its many applications. The idea is to replace $\mathbf{Sym}\, X$ with some other well understood group. Finite dimensional vector spaces (over such familiar fields as $\mathbb{Q}, \mathbb{R}$, and $\mathbb{C}$) are among the most thoroughly understood algebraic systems. Why not replace $\mathbf{Sym}\, X$ by the group of all invertible linear operators on some finite dimensional vector space? Such a group might as well be regarded as a group of invertible square matrices. The notation is various but $\mathbf{GL}(n, \mathbf{F})$ is mostly used to denote the group of all $n \times n$ invertible matrices with entries from the field $\mathbf{F}$. This is called the **general linear group**. A related subgroup might be denoted by $\mathbf{SL}(n, \mathbf{F})$, called the special linear group, which consists of all the $n \times n$ matrices of determinant 1. Representation theory proper sets off from this beginning—it is a topic worth its own course.

16.1 PROBLEM SET 15

ALGEBRA HOMEWORK, EDITION 15
SIXTEENTH WEEK
AUTOMORPHISMS OF GROUPS

PROBLEM 68.
Prove that $\mathbf{Aut}(\mathbf{S}_n) \cong \mathbf{S}_n$, for ever natural number n, except when $n = 2$ or $n = 6$. You can use, without proof, that if $n \neq 6$ then, in $\mathbf{S}_n$, the image, under any automorphism, of any transposition is again a transposition.

PROBLEM 69.
Let $\mathbf{H} \leq \mathbf{G}$. Prove that $\mathbf{N}_G(\mathbf{H})/\mathbf{C}_G(\mathbf{H})$ is embeddable into $\mathbf{Aut}(\mathbf{H})$.

PROBLEM 70.
Prove that there is no group $\mathbf{G}$ such that $\mathbf{G}/\mathbf{Z}(\mathbf{G}) \cong \mathbb{Z}$, where $\mathbb{Z}$ denotes the group of integers under addition.

PROBLEM 71.
Let $\mathbf{G}$ be a finite group.

(a) If $\mathbf{H}$ is a proper subgroup of $\mathbf{G}$, show that there is some element $x \in G$ which is not contained in any subgroup conjugate to $\mathbf{H}$.

(b) A maximal subgroup of $\mathbf{G}$ is a proper subgroup which is not contained in any other proper subgroup. Derive from the first part of the problem that if all maximal subgroups of $\mathbf{G}$ are conjugate, $\mathbf{G}$ must be cyclic.

PROBLEM 72.
Let $\mathbf{G}$ be a group of order n. Define $\varphi : G \to G$ by $\varphi(a) = a^{n^2+3n+1}$ for all $a \in G$. Prove that φ is an automorphism of $\mathbf{G}$.

PROBLEM 73.
Prove that the group of automorphisms of a finite cyclic group is Abelian.

PROBLEM 74.
Let $\mathbf{G}$ be a finite group of order $|G|$ and let $\mathbf{Z}(\mathbf{G})$ denote the center of $\mathbf{G}$. Prove the following.

(a) If $\mathbf{G}/\mathbf{Z}(\mathbf{G})$ is cyclic, then $\mathbf{G}$ in Abelian.

(b) if $|G| = pq$, where p and q are primes, then either $\mathbf{Z}(\mathbf{G}) = \{1\}$ or $\mathbf{G}$ is Abelian.

LECTURE 17

When Does a Finite Group Have a Subgroup of Size n?

Let **G** be a finite group and **H** be a subgroup of **G**. Lagrange tells us that $|H|$ must divide $|G|$. What about the converse? If n divides $|G|$ must **G** have a subgroup of order n? How many such subgroups? If not for all such n then for which?

Example. The alternating group $\mathbf{A}_4$, which has cardinality 12, has no subgroup of order 6.

Proof. By writing down the disjoint cycle decompositions one can see that in addition to the identity permutation, $\mathbf{A}_4$ has 3 elements of order 2 and 8 elements of order 3 making altogether 12 elements. The elements of order 3 are the 3-cycles and the elements of order 2 are the product of disjoint transpositions.

Let us try to make a subgroup **H** of order 6. There not being enough elements of order 2, we see that H must have an element of order 3. It does not harm to suppose that $(0,1,2) \in H$. Then the square of this element (which is also its inverse) $(0,2,1)$ also belong to H. With the identity permutation this gives us 3 of the 6 elements. We must also leave out of H the permutations $(0,2,3)$, $(0,3,1)$, and $(1,3,2)$, since putting them in would also force in their inverses as well as their products with $(0,1,2)$. After a bit of computation we find that H would then have to have more than 6 elements. Next we see that we cannot put any of the element of order 2 into H since the product of $(0,1,2)$ with any element of order 2 yields one of the three 3-cycles we just threw out of H. This leaves 3 other 3-cycles to consider. But the product of $(0,1,2)$ with any of them yields an element of order 2. So we cannot put together six of the element of A_4 to form a subgroup. □

We cannot have the full-blown converse to Lagrange's Theorem. In the example above, while we didn't get a subgroup of order 6, we certainly saw subgroups of order 2 and 3, the primes that divide 12. Of course that is for just the one group $\mathbf{A}_4$. But it suggests a starting point. Cauchy noticed the following fact.

Fact. Let **G** be a finite group and let p be a prime number. If p divides $|G|$, then **G** has a subgroup of order p.

Proof. Since p is prime, having a subgroup of order p is the same as having an element of order p.

For the sake of contradiction, suppose it were not so. Then let **G** be a smallest finite group witnessing this failure. Consider the Conjugacy Class Equation

$$|G| = |Z(\mathbf{G})| + \sum_{t \in T_1} [\mathbf{G} : \mathbf{C}(t)]$$

recalling that for $t \in T_1$ we have that centralizer $\mathbf{C}(t)$ is a proper subgroup of $\mathbf{G}$. By the minimality of $\mathbf{G}$ we find that $p \nmid |C(t)|$. Since p does divide $|G|$ we have by Lagrange that p must divide $[\mathbf{G} : \mathbf{C}(t)]$ for each $t \in T_1$. This forces the conclusion that p divides $|Z(\mathbf{G})|$. By the minimality of $|G|$ this entails that $G = Z(\mathbf{G})$. So $\mathbf{G}$ is Abelian. Now let $\mathbf{H}$ be any subgroup of $\mathbf{G}$. Because $\mathbf{G}$ is Abelian $\mathbf{H}$ is a normal subgroup of $\mathbf{G}$. So $\mathbf{G}/\mathbf{H}$ is again a group and Lagrange tells us that $|G| = |G/H||H|$. So p divides $|G/H|$ or p divides $|H|$. Suppose $\mathbf{H}$ is a proper nontrivial subgroup of $\mathbf{G}$. Then both $\mathbf{G}/\mathbf{H}$ and $\mathbf{H}$ are smaller groups. So if p divides one of these order then it must have an element of order p. This is impossible in the case of $\mathbf{H}$ since then the same element would be an element of order p in $\mathbf{G}$. What about if $\mathbf{G}/\mathbf{H}$ has an element of order p? This is the same as asserting that there is some $a \in G$ with $a \notin H$ but $a^p \in H$. Let r be the order of a^p in $\mathbf{H}$. So p and r are relatively prime since p does not divide $|H|$ and $a^{pr} = 1$. But then $(a^r)^p = 1$. Since $\mathbf{G}$ has no elements of order p, it must be that $a^r = 1$. Since p and r are relatively prime, pick integers u and v so that $1 = pu + rv$. Then $a = a^1 = (a^p)^u (a^r)^v = (a^p)^u 1^v = (a^p)^u$. But $a^p \in H$. Therefore, $a \in H$ contrary to the way a was chosen.

What all that means is that $\mathbf{G}$ is a finite Abelian group with no proper nontrivial subgroups. That is, it is a finite cyclic group. But we already know that every finite cyclic group has elements of every order that divide the order of the group. That is the desired final contradiction. □

Here is a second slicker (but perhaps not as revealing a) proof published by J. H. McKay in 1959.

Proof. Let

$$X := \{(g_0, \dots, g_{p-1}) \mid g_i \in G \text{ for all } i < p \text{ and } g_0 g_1 \dots g_{p-1} = 1\} - \{(1, \dots, 1)\}.$$

Let $\mathbf{A}$ be a cyclic group of order p and let a be a generator of $\mathbf{A}$. Let $\mathbf{A}$ act on X in such a way that $a(g_0, g_1, \dots, g_{p-1}) = (g_{p-1}, g_0, g_1, \dots, g_{p-2})$ for any $(g_0, \dots, g_{p-1}) \in X$. So the action of the generator of $\mathbf{A}$ is to rotate the coordinates of members of X by one notch. The hard-working graduate students should verify that the resulting rotated tuple belongs again to X and that this really does describe a group action. Now Lagrange and the Key Fact tell us that the size of each orbit must divide $|A| = p$. Because p is prime this means that an orbit must have size 1 or size p. An orbit of size one is just what we are looking for: a p-tuple that multiplies to 1 but which is the same tuple under all those rotation (i.e. every entry in the tuple is identical with every other entry).

So what if all the orbits were of size p? Since X is the disjoint union of the orbits, this would mean that $|X|$ would be divisible by p. But we can figure out the size of X. To make a p-tuple belonging to X we can make free and independent choices for the first $p-1$ entries. But then the last entry is forced to be the inverse of the product of the first $p-1$ entries. This means

$$|\{(g_0, \dots, g_{p-1}) \mid g_i \in G \text{ for all } i < p \text{ and } g_0 g_1 \dots g_{p-1} = 1\}| = |G|^{p-1}$$

So $|X| = |G|^{p-1} - 1$. But p divides $|G|$ so p cannot divide the cardinality of X. This means that not all the orbits can have cardinality p. So one of them must have cardinality 1. □

Let p be a prime number. A group is said to be a p-**group** provided each of its elements has an order which is a power of p. We get the following fact from Lagrange and Cauchy.

Fact. Let $\mathbf{G}$ be a finite group and p be a prime number. $\mathbf{G}$ is a p-group if and only if $|G|$ is a power of p.

Proof. First, suppose that $\mathbf{G}$ is a p-group and that the order of $\mathbf{G}$ is n. By Cauchy's Theorem the only prime that can divide n is p. So n must be a power of p.

Conversely, if $|G|$ is a power of p, then by Lagrange the order of any subgroup must also be a power of p. This applies to the cyclic subgroups, so any element must have order a power of p. This means $\mathbf{G}$ is a p-group. □

Using the Correspondence Theorem, the Theorem of Lagrange, and the Theorem of Cauchy as a base, we can even get a considerable extension of Cauchy's Theorem in the case of p-groups.

Fact. Let $\mathbf{G}$ be a group, let p be a prime number, and let k be a natural number such that $|G| = p^k$. Then there is a sequence $\mathbf{G}_0 \lhd \mathbf{G}_1 \lhd \cdots \lhd \mathbf{G}_k$ of normal subgroups of $\mathbf{G}$ such that $|G_j| = p^j$ for all $j \leq k$.

Proof. Of course $\mathbf{G}_0$ will be the trivial subgroup of $\mathbf{G}$ and $\mathbf{G}_k = \mathbf{G}$.

We will prove our fact by induction on k. In the base step of the induction we have $k = 0$ and the sequence we desire has just one group in it: $\mathbf{G}$ itself, which is a trivial group. So consider the inductive step. Here we take as an hypothesis that our fact is true for k and prove it for $k+1$. So let $\mathbf{G}$ be a group of order p^{k+1}. We know that the center $\mathbf{Z}(\mathbf{G})$ is nontrivial. Cauchy tells us it must have an element of order p. Let $\mathbf{G}_1$ be the subgroup generated by such an element. Since $\mathbf{G}_1$ is a subgroup of the center, each of the elements of G_1 commutes with all the elements of G. This ensures that $\mathbf{G}_1 \lhd \mathbf{G}$. Now according to Lagrange, $\mathbf{G}/\mathbf{G}_1$ of order p^k. By our inductive hypothesis it has a sequence

$$\mathbf{H}_0 \lhd \mathbf{H}_1 \lhd \cdots \lhd \mathbf{H}_k = \mathbf{G}/G_1$$

of normal subgroups so that $|H_j| = p^j$ for all $j \leq k$. According to the Correspondence Theorem there is a corresponding sequence

$$\mathbf{G}_1 \lhd \mathbf{G}_2 \lhd \cdots \lhd \mathbf{G}_{k+1} = \mathbf{G}$$

of normal subgroups of $\mathbf{G}$ so that $\mathbf{H}_j = \mathbf{G}_{j+1}/G_1$ for all $j \leq k$. So according to Lagrange, $|G_{j+1}| = p^{j+1}$ for all $j \leq k$. So the sequence

$$\mathbf{G}_0 \lhd \mathbf{G}_1 \lhd \mathbf{G}_2 \lhd \cdots \lhd \mathbf{G}_{k+1} = \mathbf{G}$$

is just what we desire. □

A generation after Cauchy, the great Norwegian mathematician Peter L. M. Sylow (a high school teacher for most of his life) made the breakthrough that launched a century and more of vigorous development of the rich theory of finite groups.

Let $\mathbf{G}$ be a finite group and p be a prime number. Then there must be a natural number m such that p^m divides $|G|$ but p^{m+1} does not divide $|G|$. Any subgroup of $\mathbf{G}$ of order p^m is said to be a **Sylow p-subgroup** of $\mathbf{G}$. Of course, if p does not divide $|G|$ then $m = 0$ and the Sylow p-subgroup is the trivial group (and not of much interest).

The First Sylow Theorem. *Let $\mathbf{G}$ be a finite group and p be a prime number. If p^k divides $|G|$, then $\mathbf{G}$ has a subgroup of order p^k. In particular, $\mathbf{G}$ has a Sylow p-subgroup.*

Proof. There is really nothing to prove unless p divides $|G|$. So we take that to be the case.

Once we prove that $\mathbf{G}$ has a Sylow p-subgroup we can appeal to a fact above to get the other subgroups we desire.

Our proof is by induction of $|G|$. As the base step is trivial, we consider just the inductive step.

Suppose $\mathbf{H}$ is a proper subgroup so that p does not divide $[\mathbf{G}:\mathbf{H}]$. Lagrange tells us that $|G| = [\mathbf{G}:\mathbf{H}]|H|$. So we see that p divides $|H|$ and that any Sylow p-subgroup of $\mathbf{H}$ is a Sylow p-subgroup of $\mathbf{G}$. So we could appeal to the induction hypothesis to get a Sylow p-subgroup of $\mathbf{G}$.

So it remains to consider the case that for every proper subgroup $\mathbf{H}$ of $\mathbf{G}$ we have that p divides $[\mathbf{G}:\mathbf{H}]$. Recall the Conjugacy Class Equation:

$$|G| = |Z(\mathbf{G})| + \sum_{t \in T_1} [\mathbf{G}:\mathbf{C}(t)].$$

In the sum each of the centralizers $\mathbf{C}(t)$ is a proper subgroup of $\mathbf{G}$. So it follows from the Conjugacy Class Equation that p divides $|Z(\mathbf{G})|$. According to Cauchy, $\mathbf{Z}(\mathbf{G})$ has a subgroup $\mathbf{N}$ of order p. Since $N \subseteq Z(\mathbf{G})$

we know that $\mathbf{N}$ is a normal subgroup of $\mathbf{G}$. But $\mathbf{G}/N$ is smaller than $\mathbf{G}$, so by the inductive hypothesis it must have a Sylow p-subgroup $\mathbf{P}_o$. Let m be the natural number so that p^m divides $|G|$ but p^{m+1} does not divide $|G|$. Because $|N| = p$, in view of Lagrange's Theorem, we see that $|P_o| = p^{m-1}$. Now let $P = \{a \mid a \in G \text{ and } a/N \in P_o\}$. It is easy to check that P is closed under the group operations (the inverse image under any homomorphism of a subgroup of the range is a subgroup of the domain). So we have a subgroup $\mathbf{P}$ of $\mathbf{G}$ and Lagrange tells us that $|P| = |P_o||N| = p^{m-1}p = p^m$. This means that $\mathbf{P}$ is a Sylow p-subgroup of $\mathbf{G}$, as desired. □

Our proof of the Second Sylow Theorem uses the notion of the normalizer of a subgroup. Suppose $\mathbf{H}$ is a subgroup of the group $\mathbf{G}$. Let $N_\mathbf{G}\mathbf{H} := \{a \mid a \in G \text{ and } aH = Ha\}$. We see and $H \subseteq N_\mathbf{G}\mathbf{H} \subseteq G$. Hard-working graduates can check that $N_\mathbf{G}\mathbf{H}$ is closed under the group operations. So we have the subgroup $\mathbf{N_G H}$. It is called the **normalizer** of $\mathbf{H}$ in $\mathbf{G}$. Evidently, $\mathbf{H}$ is a normal subgroup of $\mathbf{N_G H}$, and indeed the normalizer is the largest subgroup of $\mathbf{G}$ in which $\mathbf{H}$ is normal.

There is another way to get at the normalizer. Let $\mathbf{G}$ be a group and let X be the collection of all subgroups of $\mathbf{G}$. Let $\mathbf{G}$ act on X by conjugation. Then a little work shows, for any subgroup $\mathbf{H}$, that $\mathbf{Stab}\,\mathbf{H} = \mathbf{N_G H}$. The orbit of $\mathbf{H}$ under this action is just all the subgroups of $\mathbf{G}$ that are conjugate to $\mathbf{H}$. The Key Fact tells us, in this setting, that the number of subgroups conjugate with $\mathbf{H}$ is $[\mathbf{G} : \mathbf{N_G H}]$.

The Second Sylow Theorem. *Let* $\mathbf{G}$ *be a finite group and let p be a prime number. Let* $\mathbf{P}$ *be a Sylow p-subgroup of* $\mathbf{G}$ *and let* $\mathbf{H}$ *be a p-subgroup of* $\mathbf{G}$. *Then* $\mathbf{H}$ *is a subgroup of some conjugate of* $\mathbf{P}$. *In particular, any two Sylow p-subgroups of* $\mathbf{G}$ *are conjugates.*

Proof. Pick m so that $|P| = p^m$.

Let X be the collection of subgroups of $\mathbf{G}$ that are conjugates of $\mathbf{P}$. Let $\mathbf{H}$ act on X by conjugation. Consider one of the orbits $\mathcal{O}^H$ and let $\mathbf{P}_o$ be a member of $\mathcal{O}^H$. We have superscripted this orbit with H since later in this proof we will use a second action and consider one of its orbits. The Key Fact tells us

$$|\mathcal{O}^H| = [\mathbf{H} : \mathbf{Stab}\, P_o].$$

Now $\operatorname{Stab} P_o = \{h \mid h \in H \text{ and } h^{-1}P_o h = P_o\} = H \cap N_\mathbf{G} P_o$. Let $H_1 = H \cap N_\mathbf{G} P_o$. Now $\mathbf{H}_1$ is a subgroup of $\mathbf{N_G P}_o$ and $\mathbf{P}_o$ is a normal subgroup of $\mathbf{N_G P}_o$. Working inside $\mathbf{N_G P}_o$ we apply the Second Isomorphism Theorem:

$$\mathbf{H}_1\mathbf{P}_o/P_o \cong \mathbf{H}_1/H_1 \cap P_o.$$

We have $[\mathbf{H}_1\mathbf{P}_o : \mathbf{P}_o] = [\mathbf{H}_1 : \mathbf{H}_1 \cap \mathbf{P}_o]$. Since $\mathbf{H}$ is a p-group so is $\mathbf{H}_1$. So pick ℓ so that $[\mathbf{H}_1\mathbf{P}_o : \mathbf{P}_o] = p^\ell$. By Lagrange $|H_1 P_o| = [\mathbf{H}_1\mathbf{P}_o : \mathbf{P}_o]|P_o| = p^\ell p^m = p^{\ell+m}$. Since $\mathbf{P}_o$ is a Sylow p-subgroup of $\mathbf{G}$, it must be that $\ell = 0$ and so $|H_1 P_o| = |P_o|$. But $P_o \subseteq H_1 P_o$ and these sets are finite. That means $P_o = H_1 P_o$. In turn we have $H_1 \subseteq H_1 P_o = P_o$. Recalling the definition of H_1, we get $H \cap N_\mathbf{G} P_o \subseteq P_o$. So intersecting H on both sides of this inclusion we get $H \cap N_\mathbf{G} P_o \subseteq H \cap P_o$. But since $P_o \subseteq N_\mathbf{G} P_o$ we find

$$\operatorname{Stab} P_o = H \cap N_\mathbf{G} P_o = H \cap P_o.$$

So the size of our arbitrary orbit $\mathcal{O}^H$ is $[\mathbf{H} : \mathbf{H} \cap \mathbf{P}_o]$. Notice that this must be a power of p.

On the other hand, if we let $\mathbf{G}$ act on X by conjugation then, as noted above the statement of the theorem, $|\mathcal{O}^G_P| = [\mathbf{G} : \mathbf{N_G P}]$. Since $\mathbf{P}$ is a Sylow p-subgroup of $\mathbf{G}$ we see that p cannot divide $|\mathcal{O}^G_P|$. Since $\mathcal{O}^G_P$ is a disjoint union of H-orbits and each H-orbit has cardinality a power of p, there must be at least one orbit whose size is $p^0 = 1$. Let $P_o = a^{-1}Pa$ be the element of this orbit. Then $[\mathbf{H} : \mathbf{H} \cap \mathbf{P}_o] = 1$, the size of the orbit. This means $H = H \cap P_o$. Hence $H \subseteq P_o = a^{-1}Pa$, as desired. □

Here is a useful corollary of the Second Sylow Theorem.

Fact. If $\mathbf{G}$ is a finite group and p is a prime number. A Sylow p-subgroup of $\mathbf{G}$ is normal if and only if $\mathbf{G}$ has exactly one Sylow p-subgroup.

So can we get a handle on the number of Sylow p-subgroups a finite group might have?

The Third Sylow Theorem. *Let* $\mathbf{G}$ *be a finite group and let* p *be a prime number. Then the number of distinct Sylow* p*-subgroups of* $\mathbf{G}$ *is congruent to* 1 *modulo* p *and divides* $|G|$.

Proof. Let X be the collection of all Sylow p-subgroups of $\mathbf{G}$. Let $\mathbf{P}$ be a Sylow p-subgroup of $\mathbf{G}$ and let $\mathbf{P}$ act on X by conjugation. Consider any orbit $\mathcal{O}$ of this action and let $\mathbf{P}_o \in \mathcal{O}$. Now just as in the proof for the Second Sylow Theorem (letting $\mathbf{P}$ play the role here that $\mathbf{H}$ played there), we find

$$|\mathcal{O}| = [\mathbf{P} : \mathbf{P} \cap \mathbf{P}_o].$$

Again we find that each orbit has cardinality a power of p. Observe that $\{\mathbf{P}\}$ is an orbit of this action and it is of size 1. For any other orbit $\mathcal{O}$ we have for $\mathbf{P}_o \in \mathcal{O}$ that $P \neq P_o$ so that $P \cap P_o$ is strictly smaller than P. This entails, by the equation displayed above, that p divides the size of every orbit different from $\{\mathbf{P}\}$. But X is a disjoint union of the orbits, so $|X|$ is the sum of the size of the orbits. So we get that $|X|$, the number of Sylow p-subgroups of $\mathbf{G}$, is congruent to 1 modulo p.

On the other hand, by letting $\mathbf{G}$ act on X by conjugation we get only one orbit, according to the Second Sylow Theorem. So letting $\mathbf{P}$ be any Sylow p-subgroup (we have one by the First Sylow Theorem), the Key Fact tells us

$$|X| = [\mathbf{G} : \mathbf{Stab}\,\mathbf{P}].$$

By Lagrange we have $|G| = [\mathbf{G} : \mathbf{Stab}\,\mathbf{P}]|\operatorname{Stab}\mathbf{P}| = |X||\operatorname{Stab}\mathbf{P}|$. So the number $|X|$ of Sylow p-subgeroups of $\mathbf{G}$ divides the order of $\mathbf{G}$. □

17.1 Problems Set 16

Algebra Homework, Edition 16
Seventeenth Week
Ask Sylow

Problem 75.
Let p be the smallest prime that divides the cardinality of the finite group $\mathbf{G}$. Prove that any subgroup of $\mathbf{G}$ of index p must be normal.

Problem 76.
How many elements of order 7 are there in a simple group of order 168?

Problem 77.
Let $\mathbf{N}$ be a normal subgroup of the finite group $\mathbf{G}$ and let $\mathbf{K}$ be a p-Sylow subgroup of $\mathbf{N}$ for some prime p. Prove that $\mathbf{G} = \mathbf{N_G}(\mathbf{K})\mathbf{N}$.

Problem 78.
Prove that there is no simple group of order 56.

Problem 79.
Let $\mathbf{G}$ be a finite group, let $\mathbf{P}$ be a Sylow p-subgroup of $\mathbf{G}$ and let $\mathbf{H} = \mathbf{N_G}(\mathbf{P})$ be the normalizer of $\mathbf{P}$ in $\mathbf{G}$. Show that, for all $g \in G$ we have $gHg^{-1} = H$ if and only if $g \in H$.

Problem 80.
Let $\mathbf{G}$ be a finite group, and $\mathbf{P}$ be a Sylow p-subgroup of $\mathbf{G}$. Let $\mathbf{H}$ be a subgroup of $\mathbf{G}$ containing $\mathbf{P}$:

$$\mathbf{P} \leq \mathbf{H} \leq \mathbf{G}.$$

Suppose that $\mathbf{P}$ is normal in $\mathbf{H}$ and $\mathbf{H}$ is normal in $\mathbf{G}$. Show that $\mathbf{P}$ is normal in $\mathbf{G}$.

Lecture 18

Decomposing Finite Groups

We have seen the Structure Theorem for Finitely Generated Modules over a Principal Ideal Domain. That theorem said there was a way to assemble each such module from indecomposable pieces in a way that was essentially unique. Recall that it had three aspects: an existence statement, a uniqueness statement, and a description of the indecomposable modules. Roughly speaking, such a theorem opens a way to tackle many problems: first figure out what happens to the indecomposable pieces and then figure out what goes on when you put the indecomposable pieces together to form more complicate modules. Another very useful consequence was the association with each such finitely generated module a sequence of numbers that determines the module up to isomorphism.

Of course, that theorem gave us an excellent structure theorem for finitely generated Abelian groups. Here we want to address the question of whether there is a similar result that applies to all finite groups or at least some way to pull a complicated finite group apart into less intricate pieces.

18.1 Direct products of groups

Since we know how to form direct products of any system of algebras all of the same signature, we know how to form direct products of any system of groups and, as we observed after Cayley's Theorem, such direct products will again be groups.

Just as for modules, so for groups we can give a nice internal representation of the direct product of two groups **N** and **H**.

Indeed, notice that in $\mathbf{N}\times\mathbf{H}$ we have that $N^* := \{(a,1) \mid a \in N\}$ is the kernel of the projection from $\mathbf{N}\times\mathbf{H}$ onto **H** and that $H^* := \{(1,b) \mid b \in H\}$ is the kernel of the other projection function. Observe that we have the following properties:

(a) $\mathbf{N}^* \triangleleft \mathbf{N}\times\mathbf{H}$.

(b) $\mathbf{H}^* \triangleleft \mathbf{N}\times\mathbf{H}$.

(c) $\mathbf{N}^*\mathbf{H}^* = \mathbf{N}\times\mathbf{H}$.

(d) $\mathbf{N}^* \cap \mathbf{H}^*$ is trivial.

(e) $\mathbf{N}^* \cong \mathbf{N}$.

(f) $\mathbf{H}^* \cong \mathbf{H}$.

On the other hand, let us start with a group **G** and subgroups **N** and **H** such that

(a) $\mathbf{N} \triangleleft \mathbf{G}$.

(b) $\mathbf{H} \triangleleft \mathbf{G}$.

(c) $\mathbf{NH} = \mathbf{G}$.

(d) $N \cap H$ is trivial.

Then it is an enjoyable task for hard-working graduate students to verify that $\mathbf{G} \cong \mathbf{N} \times \mathbf{H}$.

So we will say that **G** is the **(internal) direct product** of its subgroups **N** and **H** provided

(a) $\mathbf{N} \triangleleft \mathbf{G}$.

(b) $\mathbf{H} \triangleleft \mathbf{G}$.

(c) $\mathbf{NH} = \mathbf{G}$.

(d) $\mathbf{N} \cap \mathbf{H}$ is trivial.

We write $\mathbf{G} = \mathbf{N} \otimes \mathbf{H}$ to mean that **G** is the internal direct product of **N** and **H**. Of course, $\mathbf{N} \otimes \mathbf{H} \cong \mathbf{N} \times \mathbf{H}$.

Here is a fact that hard-working graduate students should enjoy proving.

Fact. Let **G** be a finite group so that each of its Sylow subgroups is normal. Then **G** is the (internal) direct product of its Sylow subgroups.

Recall that we should say that a group **G** is directly indecomposable provided

- **G** is nontrivial and,
- if $\mathbf{G} = \mathbf{N} \otimes \mathbf{H}$, then either **N** or **H** is trivial.

The Krull-Schmidt Theorem. *Any finite group can be decomposed as a direct product of directly indecomposable groups. Any two such decompositions of the same finite group must have the same number (counting multiplicity) of direct factors and, after some rearranging of the factors, the corresponding direct factors in each decomposition are isomorphic.*

Thus a finite group has **unique direct factorization property**: it can be directly decomposed into directly indecomposable factors and the decomposition is unique (in the sense expressed in the theorem).

Even though I called this the Krull-Schmidt Theorem (as it is commonly called in the literature) it was known to J. H. M. Wedderburn and R. Remak in the early years of the 20th century.

This theorem is somewhat more troublesome to prove than the Structure Theorem for Finitely Generated Modules over a PID. Moreover, a description of the directly indecomposable finite groups seems currently out of reach (even though a century has passed since this theorem was first proved). The lack of such a description limits some of the usefulness of the Krull-Schmidt Theorem.

No proof is included here (but there are a number of accessible proofs in the literature).

The Krull-Schmidt Theorem has been extended in a number of ways. It remains true (and is still called the Krull-Schmidt Theorem) when the finite group is expanded by one-place operations that are endomorphisms of the original group. These kinds of expanded groups are called **groups with operators**. It also remains true, even in the expanded form, when the finiteness restriction is weakened to the restriction that the congruence lattice of the group (with operators) satisfies the finite chain condition. This is what

Krull and Schmidt did in the 1920's. There is also a Krull-Schmidt Theorem for modules satisfying the finite chain condition on their lattices of submodules.

There are more difficult and more far-reaching theorems that extend the Krull-Schmidt Theorem that are due to Garrett Birkhoff and to Bjarni Jónsson. These theorems depend on properties of the congruences of the algebras and of their congruence lattices and will not be formulated here.

In another direction there is a really striking extension of the Krull-Schmidt Theorem due to Bjarni Jónsson and Alfred Tarski. An algebra $\mathbf{A}$ is said to be an **algebra with a zero** provided $\mathbf{A}$ has among its basic operations an element designated by 0 and a two-pace operation + satisfying the following properties:

(a) The set $\{0\}$ is closed under all the basic operations.

(b) The equations $x + 0 = 0 + x = x$ hold in the algebra $\mathbf{A}$.

The Jónsson-Tarski Theorem. *Every finite algebra with a zero is uniquely factorable.*

Algebras with a zero retain just a whiff of groupness: a two-place operation with a two-sided identity element so that the identity element constitutes a one-element subuniverse. No associativity is assumed nor any inverses. The other basic operations can be completely unrestricted, apart from the stipulation that if 0 is plugged into each input position, then the output is also 0. This whiff is enough!

18.2 Decomposing a group using a chain of subgroups

We saw another way to take a group apart. When $\mathbf{G}$ is a finite p-group, where p is a prime number, we saw that there was a sequence

$$\mathbf{G} = \mathbf{G}_0 \rhd \mathbf{G}_1 \rhd \cdots \rhd \mathbf{G}_s$$

of normal subgroups of $\mathbf{G}$ such that $\mathbf{G}_s$ is trivial and each $\mathbf{G}_k/\mathbf{G}_{k+1}$ is a cyclic group of order p. So we conceive $\mathbf{G}$ as a sort of increasing union where the steps $\mathbf{G}_k/\mathbf{G}_{k+1}$ are especially simple.

We weaken this in a couple of ways to reach the notion of a solvable group. We say a group $\mathbf{G}$ is **solvable** provided there is a finite sequence of subgroups of $\mathbf{G}$ such that

$$\mathbf{G} = \mathbf{G}_0 \rhd \mathbf{G}_1 \rhd \cdots \rhd \mathbf{G}_s$$

where $\mathbf{G}_s$ is trivial and the factor groups $\mathbf{G}_k/G_{k+1}$ are Abelian for all $k < s$. Here we did not insist that each $\mathbf{G}_k$ was a normal subgroup of $\mathbf{G}$. We also only required the factor groups to be Abelian rather than the more stringent requirement that they be cyclic.

Sequences like the one appearing in the definition of solvable, but without the stipulation about the factor groups, are called **normal series**. Some authors call them *subnormal* series since the groups involved may not actually be normal subgroups of $\mathbf{G}$. Since this label might bear a demeaning psychological connotation, other authors use *normal series.*

The following fact just records an obvious point and restates a previous Fact.

Fact. Each Abelian group and each finite p-group, where p is a prime number, is solvable.

With this definition in hand, the hard-working graduate student should also be in a position to prove that both $\mathbf{S}_3$ and $\mathbf{S}_4$ are solvable.

Recall that a group $\mathbf{G}$ is said to be **simple** provided it has exactly two normal subgroups. This is equivalent to saying $\mathbf{G}$ is nontrivial and its only normal subgroups are the trivial subgroup and $\mathbf{G}$ itself.

A **composition series** of a group is a normal series in which every factor group is simple. In view of the Correspondence Theorem, another way to say this is that for any link $\mathbf{G}_i \rhd \mathbf{G}_{i+1}$ in the series there is no group $\mathbf{H}$ properly between $\mathbf{G}_i$ and $\mathbf{G}_{i+1}$ so that $\mathbf{G}_i \rhd \mathbf{H} \rhd \mathbf{G}_{i+1}$. In other words, the normal series cannot be made longer by inserting additonal groups in the series. The series we devised for finite p-groups was a composition series since each factor was a cyclic group of prime order and such groups are simple.

It is clear that every normal series for a finite group can be enhanced by the insertion of additional groups until a composition series is obtained. In particular, every finite group has at least one composition series.

Fact. Let **G** be a finite group. **G** is solvable if and only if **G** has a composition series in which each factor group is a finite cyclic group of prime order.

Proof. Since every composition series is a normal series and since every cyclic group is Abelian, we see that the condition about composition series implies that **G** is solvable.

For the converse, suppose **G** is solvable. Let

$$\mathbf{G} = \mathbf{G}_0 \rhd \mathbf{G}_1 \rhd \cdots \rhd \mathbf{G}_s$$

witness the solvability of **G**. Obtain from this series a composition series by inserting additional subgroups along the series. So a part of this composition series would be

$$\mathbf{G}_i \rhd \mathbf{H}_1 \rhd \cdots \rhd \mathbf{H}_k \rhd \mathbf{G}_{i+1}.$$

Now consider the situation where we have three groups so that $\mathbf{K} \rhd \mathbf{L} \rhd \mathbf{M}$ and we know that $\mathbf{K}/\mathbf{M}$ is Abelian. By the Third Isomorphism Theorem we have

$$(\mathbf{K}/\mathbf{M})\ /\ (L/\mathbf{M}) \cong \mathbf{K}/L.$$

Since $\mathbf{K}/\mathbf{M}$ is Abelian and $\mathbf{K}/L$ is a homomorphic image of $\mathbf{K}/\mathbf{M}$, and every homomorphic image of an Abelian group is Abelian, we find that $\mathbf{K}/\mathbf{L}$ is Abelian. We also see that $\mathbf{L}/M$ is a subgroup of $\mathbf{K}/\mathbf{M}$. So $\mathbf{L}/\mathbf{M}$ is also Abelian. This means each time we insert a new subgroup in our normal series we get a longer normal series for which all the factor groups are still Abelian. This means that the composition series we ultimately obtain has the property that each of its factor groups is a finite simple Abelian group. But the hard-working graduate students will have no trouble convincing themselves that the finite simple Abelian groups are exactly the (cyclic) groups of prime order. Of course many different primes might be associated in this way with our composition series. □

There is another characterization of the notion of solvable group which we will find useful. It involves the notion of the commutator of two normal subgroups of a group. We start by devising a new two-place operation on a group. Let **G** be a group and $a, b \in G$. We put $[a, b] := a^{-1}b^{-1}ab$ and call it the **commutator** of a and b. Notice that if a and b commute with each other in **G**, then $[a, b] = 1$. Also it proves convenient to know that $[a, b]^{-1} = [b^{-1}ab, b^{-1}]$, as can be verified by hard-working graduate students with a bit of calculation.

Now let **N** and **K** be normal subgroups of the group **G**. We let $[\mathbf{N},\mathbf{K}]$ denote the subgroup of **G** that is generated by the set $\{[a, b] \mid a \in N \text{ and } b \in K\}$. We call $[\mathbf{N},\mathbf{K}]$ the **commutator** of **N** and **K**.

Fact. Let **G** be a group with normal subgroups **N** and **K**. Then the elements of $[\mathbf{N},\mathbf{K}]$ are exactly the elements of G of the form

$$[a_0, b_0][a_1, b_1]\dots[a_k, b_k]$$

where k is some natural number and $a_i \in N$ and $b_i \in K$ for each $i \le k$.

This Fact just depends on the normality of **N** and **K**, and on the fact that **H** is a subgroup of **G**, in view of the description of $[a, b]^{-1}$ given above. Some crucial properties of the commutator are gathered in the next Fact. Its proof requires a straightforward pleasant effort from the hard-working graduate students.

Fact. Let **G** be a group with normal subgroups **N** and **K** and subgroup **H**. Then

(a) $[\mathbf{N},\mathbf{K}] \lhd \mathbf{G}$.

(b) $[\mathbf{N},\mathbf{K}] \leq \mathbf{N} \cap \mathbf{K}$.

(c) $[\mathbf{H},\mathbf{H}] \leq [\mathbf{G},\mathbf{G}]$.

(d) $[\mathbf{H}/N,\mathbf{H}/N] = [\mathbf{H},\mathbf{H}]/N$.

In conclusion (d) above we mean by $\mathbf{H}/N$ and $[\mathbf{H},\mathbf{H}]/N$ the subgroups of $\mathbf{G}/N$ that are the images of $\mathbf{H}$ and of $[\mathbf{H},\mathbf{H}]$ respectively under the quotient map. Some care is needed in noting this, because it may well be that $\mathbf{N}$ is neither a subgroup of $\mathbf{H}$ nor of $[\mathbf{H},\mathbf{H}]$. We also observe that $[\mathbf{H},\mathbf{H}]$ and $[\mathbf{H}/N,\mathbf{H}/N]$ are to be understood for the commutator in th groups $\mathbf{H}$ and $\mathbf{H}/N$ respectively.

Let $\mathbf{G}$ be a group. The **derived group** of $\mathbf{G}$ is the group $\mathbf{G}' := [\mathbf{G},\mathbf{G}]$. We can iterate this formation of derived groups by the following recursion.

$$\mathbf{G}^{(0)} := \mathbf{G}$$
$$\mathbf{G}^{(k+1)} := [\mathbf{G}^{(k)},\mathbf{G}^{(k)}] \text{ for all natural numbers } k$$

So $\mathbf{G}' = \mathbf{G}^{(1)}, (\mathbf{G}')' = \mathbf{G}^{(2)}$, and so on.

Perhaps the next Fact gives more support to the label “commutator”.

Fact. Let $\mathbf{G}$ be a group. Then $\mathbf{G}/[\mathbf{G},\mathbf{G}]$ is Abelian; moreover, if $\mathbf{N} \lhd \mathbf{G}$ and $\mathbf{G}/\mathbf{N}$ is Abelian, then $[\mathbf{G},\mathbf{G}] \leq \mathbf{N}$.

Proof. Let $a, b \in G$. The cosets $a[\mathbf{G},\mathbf{G}]$ and $b[\mathbf{G},\mathbf{G}]$ are arbitrary elements of $\mathbf{G}/[\mathbf{G},\mathbf{G}]$. To say that they commute is just to assert $ab[\mathbf{G},\mathbf{G}] = ba[\mathbf{G},\mathbf{G}]$. But this is evidently the same as asserting $[a,b] = (ba)^{-1}(ab) \in [\mathbf{G},\mathbf{G}]$. This assertion is certainly true since we took the elements of the form $[a,b]$ as generators of $[\mathbf{G},\mathbf{G}]$.

Now suppose $\mathbf{N} \lhd \mathbf{G}$ and $\mathbf{G}/N$ is Abelian. Let $a, b \in G$. So we see that $abN = baN$. But this means $[a,b] \in N$. So all the generators of $[\mathbf{G},\mathbf{G}]$ belong to N. Since $\mathbf{N}$ is a subgroup, this entails that $[\mathbf{G},\mathbf{G}] \leq \mathbf{N}$. □

So we see that for an arbitrary group $\mathbf{G}$ we have

$$\mathbf{G} = \mathbf{G}^{(0)} \rhd \mathbf{G}^{(1)} \rhd \mathbf{G}^{(2)} \rhd \cdots \rhd \mathbf{G}^{(k)} \rhd \mathbf{G}^{(k+1)} \rhd \ldots$$

As far as it goes it is a normal series (and moreover each of the groups is even a normal subgroup of $\mathbf{G}$) and each factor group is Abelian. Here is our characterization of solvability.

Fact. The group $\mathbf{G}$ is solvable if and only if $\mathbf{G}^{(n)}$ is the trivial group, for some natural number n.

Proof. In the event that $\mathbf{G}^{(n)}$ is trivial, the series

$$\mathbf{G} = \mathbf{G}^{(0)} \rhd \mathbf{G}^{(1)} \rhd \mathbf{G}^{(2)} \rhd \cdots \rhd \mathbf{G}^{(n)}$$

witnesses that $\mathbf{G}$ is solvable.

For the converse, suppose that $\mathbf{G}$ is solvable and let

$$\mathbf{G} = \mathbf{G}_0 \rhd \mathbf{G}_1 \rhd \cdots \rhd \mathbf{G}_n$$

be a normal series that witnesses the solvability of $\mathbf{G}$. So $\mathbf{G}_n$ is trivial and all the factor groups are Abelian. Consider the first link $\mathbf{G} \rhd \mathbf{G}_1$. We certainly get $\mathbf{G}_1 \rhd [\mathbf{G},\mathbf{G}] = \mathbf{G}^{(1)}$. Similarly, at the next link we see $\mathbf{G}_2 \rhd [\mathbf{G}_1,\mathbf{G}_1]$. But we already know $\mathbf{G}_1 \rhd \mathbf{G}^{(1)}$. Since we know the commutator respects the inclusion ordering, we get $[\mathbf{G}_1,\mathbf{G}_1] \rhd [\mathbf{G}^{(1)},\mathbf{G}^{(1)}] = \mathbf{G}^{(2)}$. Putting things together, we get $\mathbf{G}_2 \rhd \mathbf{G}^{(2)}$. Continuing in this way, we get $\mathbf{G}_k \rhd \mathbf{G}^{(k)}$ in general. So at the end we have $\mathbf{G}_n \rhd \mathbf{G}^{(n)}$. Since $\mathbf{G}_n$ is trivial, we find that $\mathbf{G}^{(n)}$ is also trivial, as desired. □

Fact. Every subgroup of a solvable group is solvable. Every homomorphic image of a solvable group is solvable. Let **N** be a normal subgroup of the group **G**. **G** is solvable if and only if both **N** and $\mathbf{G}/N$ are solvable.

Proof. Let **G** be a solvable group.

For any normal subgroup **N** of **G** we know $[\mathbf{G}/N, \mathbf{G}/N] = [\mathbf{G}, \mathbf{G}]/N$. Another way to write this is $(\mathbf{G}/N)^{(1)} = \mathbf{G}^{(1)}/N$. Using this equality, we also see

$$(\mathbf{G/N})^{(2)} = [(\mathbf{G/N})^{(1)}, (\mathbf{G/N})^{(1)}] = [\mathbf{G}^{(1)}/\mathbf{N}, \mathbf{G}^{(1)}/\mathbf{N}] = [\mathbf{G}^{(1)}, \mathbf{G}^{(1)}]/\mathbf{N} = \mathbf{G}^{(2)}/\mathbf{N}$$

Proceeding in this way we find $(\mathbf{G/N})^{(k)} = \mathbf{G}^{(k)}/\mathbf{N}$. So if $\mathbf{G}^{(n)}$ turns out to be the trivial group, then so will $(\mathbf{G/N})^{(n)}$. This means that if **G** is solvable, then so is its homomorphic image **G/N**.

For any subgroup **H** of **G** we know that $\mathbf{H}^{(1)} = [\mathbf{H}, \mathbf{H}] \leq [\mathbf{G}, \mathbf{G}] = \mathbf{G}^{(1)}$. An easy induction argument shows that $\mathbf{H}^{(n)} \leq \mathbf{G}^{(n)}$ for all natural numbers n. So if **G** is solvable so must its subgroup **H** be solvable.

Now suppose that **G** is an arbitrary group and that **N** is a normal subgroup such that both **N** and **G/N** are solvable. Pick a natural number so that $(\mathbf{G}/N)^{(n)}$ is trivial. Since we now know that $(\mathbf{G/N})^{(n)} = \mathbf{G}^{(n)}/\mathbf{N}$ it follows that $\mathbf{G}^{(n)} \leq \mathbf{N}$. But **N** is solvable, so we know all its subgroups are solvable. This means we can pick a natural number m so that $(\mathbf{G}^{(n)})^{(m)}$ is trivial. But it is easy to discover that $(\mathbf{G}^{(n)})^{(m)} = \mathbf{G}^{(n+m)}$, which must be trivial. So **G** is solvable. □

A somewhat different proof could be mounted that involves manipulating the normal series witnessing the various solvability constraint. Those proofs make heavy use of the isomorphism theorems.

So far this approach to decomposing a group using a normal series has concentrated on existence. We have seen that at least every finite group has a composition series (where the factor groups are all simple). For solvable groups we even got the existence of a composition series where the factor groups were cyclic groups of prime order. What about uniqueness?

Even for finite Abelian groups it easy to find examples where there are several different composition series. This is something like the situation with direct decompositions—one could get a different decomposition just by rearranging the direct factors in the direct product and swapping out some factors with isomorphic copies. So we aim to prove a kind of uniqueness theorem for composition series.

Let **G** be a group. We will say that two normal series for **G**

$$\mathbf{G} = \mathbf{G}_0 \rhd \mathbf{G}_1 \rhd \cdots \rhd G_n$$
$$\mathbf{G} = \mathbf{H}_0 \rhd \mathbf{H}_1 \rhd \cdots \rhd H_m$$

are **equivalent** provided $n = m$ and for some permutation σ of $\{0, 1, \ldots, n\}$ we have

$$\mathbf{G}_i/G_{i+1} \cong \mathbf{H}_{\sigma(i)}/H_{\sigma(i)+1} \text{ for all } i < n$$

That is, the series are the same length and the sequence of factor groups along one of the normal series can be rearranged to obtain, within isomorphism, the sequence of factor groups along the other normal series.

Our aim is to prove

The Jordan-Hölder Theorem. *Any composition series of a group is equivalent to any other composition series.*

We will be able to obtain this theorem as an immediate consequence of another theorem.

Let **G** be a group. We say one normal series for **G** is a **refinement** of another if the first can be obtained from the second by inserting some finite number of additional subgroups along the series. The Jordan-Hölder Theorem is an immediate consequence of

Schreier's Refinement Theorem. *Any two normal series for a group have refinements that are equivalent to each other.*

Proof. Let the group **G** have two normal series

$$\mathbf{G} = \mathbf{A}_0 \rhd \mathbf{A}_1 \rhd \cdots \rhd \mathbf{A}_n$$
$$\mathbf{G} = \mathbf{B}_0 \rhd \mathbf{B}_1 \rhd \cdots \rhd \mathbf{B}_m.$$

So we know that both $\mathbf{A}_n$ and $\mathbf{B}_m$ are the trivial subgroup of **G**.

We will invoke the Zassenhaus Butterfly Lemma to construct the two refinements we require. From the coarsest view, that lemma allows us to insert in

$$\mathbf{A} \rhd \mathbf{A}^*$$
$$\mathbf{B} \rhd \mathbf{B}^*$$

two additional groups each so that

$$\mathbf{A} \rhd \geq \mathbf{A}_u \rhd \mathbf{A}_\ell \geq \mathbf{A}^*$$
$$\mathbf{B} \rhd \geq \mathbf{B}_u \rhd \mathbf{B}_\ell \geq \mathbf{B}^*$$
$$\mathbf{A}_u / A_\ell \cong \mathbf{B}_u / B_\ell$$
$$.$$

This coarse view is not adequate for our purposes because some of the subgroup relations are not normal. But the Butterfly Lemma is certainly tempting due to that isomorphism. Fortunately, the actual Butterfly Lemma carries more detail.

Here is what works. Let

$$\mathbf{C}_{i,j} := \mathbf{A}_{i+1}(\mathbf{A}_i \cap \mathbf{B}_j) \text{ and } \mathbf{D}_{i,j} := \mathbf{B}_{j+1}(\mathbf{B}_j \cap \mathbf{A}_i)$$

These are the groups that come up in full detail in the Butterfly Lemma. What the Butterfly Lemma say about them is

$$\mathbf{C}_{i,j} \rhd \mathbf{C}_{i,j+1}$$
$$\mathbf{D}_{i,j} \rhd \mathbf{D}_{i+1,j}$$
$$\mathbf{C}_{i,j} / C_{i,j+1} \cong \mathbf{D}_{i,j} / D_{i+1,j}$$

To understand better what is going on, fix a value of i. Then we see

$$\begin{aligned}
\mathbf{C}_{i,0} &= \mathbf{A}_{i+1}(\mathbf{A}_i \cap \mathbf{B}_0) = \mathbf{A}_{i+1}(\mathbf{A}_i \cap \mathbf{G}) = \mathbf{A}_i \\
\mathbf{C}_{i,1} &= \mathbf{A}_{i+1}(\mathbf{A}_i \cap \mathbf{B}_1) \\
&\vdots \\
\mathbf{C}_{i,m} &= \mathbf{A}_{i+1}(\mathbf{A}_i \cap \mathbf{B}_m) = \mathbf{A}_{i+1}
\end{aligned}$$

where the last line comes about since $\mathbf{B}_m$ is the trivial subgroup. Moreover, the Butterfly Lemma tells us

$$\mathbf{A}_i = \mathbf{C}_{i,0} \rhd \mathbf{C}_{i,1} \rhd \cdots \rhd \mathbf{C}_{i,m} = \mathbf{A}_{i+1}$$

In this way, we see that the $\mathbf{C}_{i,j}$'s, once they are arranged in the proper order, form a normal series for **G** that refines the series of the $\mathbf{A}_i$'s. The proper order is the lexicographic order on $\{(i,j) \mid i < n \text{ and } j < m\}$ given by

$$(i,j) \geq (\ell,k) \Leftrightarrow i \geq \ell \text{ or else } i = \ell \text{ and } j \geq k.$$

Of course an entirely similar analysis leads us to the conclusion that the $\mathbf{D}_{i,j}$'s, arranged properly, form a normal series that refines the series of the $\mathbf{B}_j$'s.

But these two refinements are equivalent since for all $i \le n$ and all $j \le m$ we have

$$\mathbf{C}_{i,j}/C_{i,j+1} \cong \mathbf{D}_{i,j}/D_{i+1,j}.$$

□

You should notice that we did not insist that the normal subgroups along a normal series be proper subgroups. It would be sensible to insist on this since it gives a cleaner connotation to the length of a series. Then in the proof above one must systematically delete one of the groups when

$$\mathbf{C}_{i,j} = \mathbf{C}_{i,j+1} \text{ or } \mathbf{D}_{i,j} = \mathbf{D}_{i+1,j}.$$

Since we know $\mathbf{C}_{i,j}/C_{i,j+1} \cong \mathbf{D}_{i,j}/D_{i+1,j}$ every deletion from the series of $\mathbf{C}_{i,j}$'s must be accompanied by a deletion from the series of $\mathbf{D}_{i,j}$'s, and vice versa. So after all the deletions, the resulting refinements will still have the same length.

The Jordan-Hölder Theorem was proved, in some form, late in the 19th century when Otto Hölder put the finishing touches on a proof of Camille Jordan.

There are a number of different ways to prove the Jordan-Hölder Theorem. For finite groups, it is possible to devise a proof by induction of the size of the group. It is also possible to make a proof by induction of the length of the composition series involved. Otto Schreier's proof of the Jordan-Hölder Theorem using the Refinement Theorem was published in 1928. Hölder was still alive but Jordan had died six years earlier. Hans Zassenhaus gave a new proof of Shreier's Refinement Theorem using his own Butterfly Lemma in the early 1930's while he was still a hard-working graduate student under the direction of Emil Artin. Zassenhaus himself became a prolific mathematician with over 200 papers and 41 PhD students. He died in 1991.

The theorem has been called the Jordan-Hólder-Schreier Theorem or even the Jordan-Hölder-Schreier-Zassenhaus Theorem. It attaches to every finite group a sequence of finite simple groups that provides some structural information about the finite group.

18.3 Addendum: A notion related to solvability

It is an easy observation that every finite Abelian group can be decomposed as a direct product of its Sylow subgroups. In fact one path to the Fundamental Structure Theorem for Finite Abelian Groups starts from this observation. One could consider the class of all finite groups that can the decomposed as a direct product of their Sylow subgroups. This proves to be a class that is wider than the class of finite Abelian groups but narrower than the class of finite solvable groups. Finite groups that are the direct product of their Sylow subgroups are called *nilpotent*. Just as the class of solvable groups can be characterized using the commutator of normal subgroups, so can the class of nilpotent groups.

Let $\mathbf{G}$ be a group. The $\mathbf{G}^{[n]}$ is defined by the following recursion:

$$\begin{aligned}\mathbf{G}^{[0]} &= \mathbf{G}\\ \mathbf{G}^{[k+1]} &= [\mathbf{G}^{[k]}, \mathbf{G}] \text{ for all natural numbers } k.\end{aligned}$$

An easy induction shows the $\mathbf{G}^{(k)}$ is always a subgroup of $\mathbf{G}^{[k]}$.

The group $\mathbf{G}$ is said to be **nilpotent** provided $\mathbf{G}^{[n]}$ is trivial for some natural number n. This definition works for groups that might not be finite. So it is true that every nilpotent group is solvable, although the converse fails. The class of nilpotent groups is a proper subclass of the class of solvable groups. It is also a theorem that a finite group is nilpotent if and only if it is the direct product of its Sylow subgroups. The theories of nilpotent and of solvable groups have elaborate developments, exhibiting many parallels and interconnections.

It was, of course, Galois that first noticed the significance of the class of (finite) solvable groups in his investigations of the solvability of polynomial equations of degree n by means of radicals.

One of the most notable theorems about solvable groups is

The Feit-Thompson Theorem. *Every group of odd order is solvable.*

The proof of this one theorem extends to hundreds of pages—longer than this whole account of algebra for first year graduate students!

18.4 Problem Set 17

Algebra Homework, Edition 17
Eighteenth Week
Mr. Sylow deals with finite Abelian groups and other matters

Problem 81.
Prove that if $\mathbf{G}$, $\mathbf{H}$, and $\mathbf{K}$ are finite Abelian groups and $\mathbf{G} \times \mathbf{H} \cong \mathbf{G} \times \mathbf{K}$, then $\mathbf{H} \cong \mathbf{K}$.

Problem 82.
Prove that every group of order 35 is cyclic.

Problem 83.
Describe, up to isomorphism, all groups of order 1225.

Problem 84.
Let $\mathbf{G}$ be a finite Abelian group. Prove that if $|G|$ is not divisible by k^2 for any $k > 1$, then $\mathbf{G}$ is cyclic.

Problem 85.
Let p be a prime number. For any finite group $\mathbf{G}$, let $\mathbb{B}(\mathbf{G})$ denote the subgroup of $\mathbf{G}$ generated by all the Sylow p-subgroups of $\mathbf{G}$.

(a) Show that $\mathbb{B}(\mathbf{G})$ is the unique normal subgroup of $\mathbf{G}$ minimal with respect to the property that its index is not divisible by p.

(b) Let $\mathbf{L}$ be a normal subgroup of the finite group $\mathbf{G}$. Show that $\mathbb{B}(\mathbf{L})$ is normal in $\mathbf{G}$ and that $\mathbb{B}(\mathbf{L}) = \mathbb{B}(\mathbf{G})$ if $[\mathbf{G} : \mathbf{L}]$ is not divisible by p.

(c) Now let $\mathbf{H}$ be a subgroup of the finite group $\mathbf{G}$ with $[\mathbf{G} : \mathbf{H}] = p$. If $\mathbf{L}$ is the largest normal subgroup of $\mathbf{G}$ contained in $\mathbf{H}$, prove that $[\mathbf{H} : \mathbf{L}]$ is not divisible by p and deduce that $\mathbb{B}(\mathbf{H})$ is normal in $\mathbf{G}$.

Problem 86.
Let $\mathbf{G}$ be finite group with an automorphism σ such that $\sigma^2 = \mathrm{id}$ and the only element fixed by σ is the identity of $\mathbf{G}$. Show $\mathbf{G}$ is Abelian.

LECTURE 19

WHERE TO FIND THE ROOTS OF A POLYNOMIAL

A leading motivation for the rest of the semester is the project: to describe all the roots of a given polynomial in one variable with coefficients from some field.

Let **F** be a field and $f(x) \in F[x]$ be a polynomial with coefficients from F. For instance, **F** might be the field $\mathbb{Q}$ of rational numbers and $f(x)$ might be $x^2 - 2$. This polynomial has no roots in $\mathbb{Q}$, but on the other hand, $f(x)$ is also a polynomial over the field $\mathbb{R}$ of real numbers and in this larger field we find two roots of $f(x)$, namely $\sqrt{2}$ and $-\sqrt{2}$. After a bit of reflection, observing that $\mathbb{Q}$ is countable and $\mathbb{R}$ is uncountable, we see that on the one hand there is quite a gap between $\mathbb{Q}$ and $\mathbb{R}$, while on the other hand $\mathbb{R}$ is not really adequately supplied with roots—the polynomial $x^2 + 1$, has no roots in $\mathbb{R}$.

Wanting to describe the roots of the polynomials from the ring $\mathbf{F}[x]$, we see that we might well consider fields **K** that extend **F**. There is an unlimited supply of these. The principle of parsimony leads us to look for the ones that are some way or another close to **F** but still rich enough to allow us to have a full complement of roots of $f(x)$ or, what is the same, to be able to factor $f(x)$ into a product of polynomials of degree 1.

19.1 ALGEBRAIC EXTENSIONS OF FIELDS

There are two key insights that are the starting point of our efforts. The first is that if **F** is a subfield of **K**, then **K** can be construed as a vector space over **F**. This allows us to use one of the most well-understood and thoroughly developed branches of mathematics, the theory of vector spaces. We might even hope that the most interesting extension fields **K** will turn out to be finite dimensional over **F**. We use $[\mathbf{K}:\mathbf{F}]$ to denote the dimension of **K** as a vector space over **F**. We also refer to this dimension as the **degree** of the extension. It may be an infinite cardinal number.

The second insight is that, when **K** has a full complement of roots of $f(x)$, then every automorphism of **K** that has all the coefficients of $f(x)$ as fixed points must permute the roots of $f(x)$ that are in **K**. The set of roots of $f(x)$ is a finite subset of K. So we see emerging a finite subgroup of the concrete group of all permutations of this set of roots. This allows us to bring in the theory of (finite) groups.

So we see our enterprise as a mixture of ring theory (to understand the rings like $\mathbf{F}[x]$ and $\mathbf{K}[x]$), the theory of fields, the theory of vector spaces, and group theory.

The first step we will take is to lay our hands on the minimal extension **K** of **F** that has a complete set of roots of $f(x)$.

We say that $f(x)$ **splits** over the field **K** provided $f(x) \in K[x]$ and the irreducible factors of $f(x)$ in $K[x]$ all have degree 1. We start looking for at least one root.

Kronecker's Theorem, (1887). *Let* $\mathbf{F}$ *be a field and let* $f(x) \in \mathbf{F}[x]$ *be irreducible. Then there is a field* $\mathbf{K}$ *extending* $\mathbf{F}$ *such that* $f(x)$ *has a root* $s \in K$ *and if* $\mathbf{L}$ *is any field extending* $\mathbf{F}$ *such that* $f(x)$ *has a root* $r \in L$*, then there is an embedding of* $\mathbf{K}$ *into* $\mathbf{L}$ *that fixes each element of* F *and sends* s *to* t*. Moreover, the dimension of* $\mathbf{K}$ *as a vector space over* $\mathbf{F}$ *is the degree of* $f(x)$*.*

Proof. Because $\mathbf{F}[x]$ is a principal ideal domain we know that irreducible and prime elements coincide and that so do the prime ideals and the maximal ideals. So $(f(x))$ is a maximal ideal of $\mathbf{F}[x]$. Consequently, $\mathbf{F}[x]/(f(x))$ is a field. Essentially, this is the field we desire and the element $x/(f(x))$ is the root s. A bothersome point is that it does not actually extend the field $\mathbf{F}$, but rather has a subfield (with underlying set equal to $\{a/(f(x)) \mid a \in F\}$) easily seen to be isomorphic to $\mathbf{F}$. So one must do some set theoretic surgery, snipping out the isomorphic copy and stitching in its place the field $\mathbf{F}$ itself. The result is the field $\mathbf{K}$.

Now let $\mathbf{L}$ be any field extending $\mathbf{F}$ that has a root r of $f(x)$. We know that we can map $\mathbf{F}[x]$ into $\mathbf{K}$ via a homomorphism Ψ that extends the identity map on $\mathbf{F}$ and so that $\Psi(x) = r$. We see that $\Psi(f(x)) = f(r) = 0$ since Ψ is a homomorphism and r is a root of $f(x)$ in $\mathbf{L}$. This means that $f(x) \in \ker\Psi$. On the other hand, if $g(x) \in \mathbf{F}[x]$ and $g(r) = 0$ in $\mathbf{L}$, then $f(x)$ and $g(x)$ have a common factor $x - r$ in $\mathbf{L}[x]$. So they are not relatively prime. This means they cannot be relatively prime in $\mathbf{F}[x]$ either. Since $f(x)$ is prime in $\mathbf{F}[x]$ we find that $f(x) \mid g(x)$. Hence every polynomial belonging to the kernel of Ψ is a multiple of $f(x)$. So $(f(x)) = \ker\Psi$. This means, according to the Homomorphism Theorem, that $\mathbf{F}[x]/(f(x)) \cong \mathbf{L}'$ where $\mathbf{L}'$ is the image of $\mathbf{F}[x]$ under Ψ. But this means that $\mathbf{K} \cong \mathbf{L}'$ (so $\mathbf{L}'$ is actually a field) and we see that $\mathbf{K}$ embeds into $\mathbf{L}$ by a map that fixes each element of F and sends s to r.

Finally, suppose s is a root of $f(x)$ in K and suppose that n is the degree of $f(x)$. Let Φ be a homomorphism from $\mathbf{F}[x]$ onto $\mathbf{K}$ that fixes every element of F and maps x to s. So every element of $\mathbf{K}$ is the image of some $h(x) \in F[x]$ under Φ. But in $\mathbf{F}[x]$ we can pick (uniquely) polynomials $q(x)$ and $r(x)$ so that $h(x) = q(x)f(x) + r(x)$ such that either $r(x)$ is the zero polynomial or else the degree of $r(x)$ is strictly less than n, the degree of $f(x)$. So we find $h(s) = q(s)f(s) + r(s) = r(s)$. But $r(s)$ is a linear combination with scalars from F of $\{1, s, \ldots, s^{n-1}\}$. So the latter set spans $\mathbf{K}$ as a vector space over $\mathbf{F}$. But our contention is that this set is also linearly independent. Were it otherwise, we would have a nontrivial linear combination of these that would be 0. This would give us a nonzero polynomial $g(x)$ in $\mathbf{F}[x]$ that has s as a root. So we would see that $f(x)$ and $g(x)$ are not relatively prime. But $f(x)$ is prime (in $\mathbf{F}[x]$) and so $f(x) \mid g(x)$ in $\mathbf{F}[x]$, which is impossible since $g(x)$ is a nonzero polynomial of degree strictly less than the degree of $f(x)$. So the degree of $f(x)$ is the dimension of $\mathbf{K}$ as a vector space of $\mathbf{F}$. □

There is a bit more mileage to be had from the proof of Kronecker's Theorem. Let $\mathbf{K}$ be a field extending the field $\mathbf{F}$. We say that an element $s \in K$ is **algebraic** over $\mathbf{F}$ provided s is a root of some polynomial of positive degree from $\mathbf{F}[x]$. Of course, since every such polynomial can be factored into irreducible polynomials, and since $\mathbf{K}$ is an integral domain, we must have that every algebraic element of K actually is the root on an irreducible monic polynomial from $\mathbf{F}[x]$. This monic irreducible polynomial is called the **minimal polynomial** of s. An element of K that is not algebraic over $\mathbf{F}$ is said to be **transcendental** over $\mathbf{F}$. So here are some corollaries of our proof of Kronecker's Theorem.

Corollary 19.1.1. *Let the field* $\mathbf{K}$ *extend the field* $\mathbf{F}$ *and let* $s \in K$ *be algebraic over* $\mathbf{F}$*. Then the smallest subring of* $\mathbf{K}$ *that includes* $F \cup \{s\}$ *is, in fact, a subfield of* $\mathbf{K}$*.*

In general, when $\mathbf{K}$ is a field extending the field $\mathbf{F}$ and $s \in K$, we use the notation $\mathbf{F}[s]$ for the subring of $\mathbf{K}$ generated by $F \cup \{s\}$ and the notation $\mathbf{F}(s)$ for the subfield of $\mathbf{K}$ generated by $F \cup \{s\}$. The corollary above says that if s is algebraic over $\mathbf{F}$, then $\mathbf{F}(s) = \mathbf{F}[s]$.

Corollary 19.1.2. *Let the field* $\mathbf{K}$ *extend the field* $\mathbf{F}$ *and let* $s \in K$ *be algebraic over* $\mathbf{F}$*. Then every element of* $\mathbf{F}[s]$ *is algebraic of* $\mathbf{F}$*.*

In general, we say that **K** is an **algebraic extension** of **F** provided every element of K is a root of some polynomial in $\mathbf{F}[x]$ that has positive degree. So this corollary asserts that if $s \in K$ is algebraic over **F**, then $\mathbf{F}[s]$ is an algebraic extension of **F**. If **K** is not an algebraic extension of **F** we call it a **transcendental extension** of **F**.

Now the field **K** given to us in Kronecker's Theorem provides us with an essentially unique extension of **F** that contains at least one root r of our irreducible polynomial $f(x) \in \mathbf{F}[x]$. So in $\mathbf{K}[x]$ we can factor $f(x)$ at least a little bit: there is $q(x) \in \mathbf{K}[x]$ so that $f(x) = (x-r)q(x)$. But we are not assured that $q(x)$, which still might have large degree, can be factored any further. So while **K** has at least one root of $f(x)$ it may not have a full complements of roots. Of course, the remedy is obvious. The degree of $q(x)$ is smaller than the degree of $f(x)$, so first we factor $q(x)$ into irreducibles in $\mathbf{K}[x]$ and then we invoke Kronecker on each of these, doing this again and again until some field **L** is reached in which $f(x)$ splits. While each step in this envisioned construction yields an essentially unique way to get to the next field extension, there are lots of arbitrary choices that have to be made along the way. The question of which irreducible polynomial to address next can be resolved at any stage in a number of ways. So maybe there are lots of different fields like **L** in which $f(x)$ splits, with any one of them as minimal as it can be. Fortunately, the whole business works out better than that.

Let **F** be a field and let $\mathcal{S}$ be a collection of of polynomials of positive degree, all drawn from $\mathbf{F}[x]$. A field **K** that extends **F** is said to be a **splitting field** of $\mathcal{S}$ over **F** provided

- in $\mathbf{K}[x]$ every polynomial in $\mathcal{S}$ factors as a product of polynomials of degree 1, and
- **K** is generated by $F \cup \{r \mid r \in K \text{ and } r \text{ is a root of some polynomial in } \mathcal{S}\}$.

We say that **K** is a splitting field of $f(x)$ over **F** instead of that **K** is a splitting field of $\{f(x)\}$ over **F**. Then the step-by-step, recursive extension of Kronecker's Theorem outlined above gives us

Corollary 19.1.3. *Let* **F** *be a field and* $f(x)$ *be a polynomial of positive degree that belongs to* $F[x]$. *Then* $f(x)$ *has a splitting field over* **F**.

We would like to see that the splitting field is essentially unique, that it is an algebraic extension of **F**, and that it is finite dimensional as a vector space over **F** (and even more, we would like to lay hands on this dimension).

The Dimension Formula. *Let the field* **L** *be an extension of the field* **K** *that is in turn an extension of the field* **F**. *Then* $[\mathbf{L}:\mathbf{F}] = [\mathbf{L}:\mathbf{K}][\mathbf{K}:\mathbf{F}]$.

Proof. Let B be a basis for the vector space **K** over the field **F** and let C be a basis for the vector space **L** over the field **K**. Put

$$BC := \{bc \mid b \in B \text{ and } c \in C\}.$$

Our contention is that the set BC is a basis for the vector space **L** over the field **F** and that $|BC| = |B||C|$. So we must prove that BC spans **L** (with scalars chosen from F), that BC is linearly independent, and that there is a one-to-one correspondence between $B \times C$ and BC.

Contention. BC spans **L** as a vector space over **F**.

Let $w \in L$. Since C spans **L** over **K**, pick $c_0, c_1, \dots, c_{n-1} \in C$ and $d_0, d_1, \dots, d_{n-1} \in K$ so that

$$w = \sum_{i<n} d_i c_i.$$

Now consider any $i < n$. We have $d_i \in K$. Since B spans **K** over **F**, pick $b_{i,0}, b_{i,1}, \dots . b_{i,m_i-1} \in B$ and $a_{i,0}, a_{i,}, \dots, a_{i,m_i-1} \in F$ so that

$$d_i = \sum_{j<m_i} a_{i,j} b_{i,j}.$$

Putting these two pieces together, we get

$$w = \sum_{i<n} d_i c_i = \sum_{i<n} \left(\sum_{j<m_i} a_{i,j} b_{i,j} \right) c_i.$$

In this way, we see

$$w = \sum_{i<n, j<m_i} a_{i,j}(b_{i,j} c_i),$$

which is a linear combination of elements of BC using scalars from F.

Contention. The set BC is a linearly independent subset of the vector space $\mathbf{L}$ over the field $\mathbf{F}$.

Let us suppose that $a_0, a_1, \ldots, a_{n-1} \in F$, $b_0, b_1, \ldots, b_{n-1} \in B$, and $c_0, c_1, \ldots, c_{n-1} \in C$ have been chosen so that

$$\sum_{i<n} a_i(b_i c_i) = 0$$

and that $b_0c_0, b_1c_1, \ldots, b_{n-1}c_{n-1}$ are distinct. Now perhaps not all the c_i's are distinct. However, by rearranging the indices we may suppose that $c_0, \ldots, c_{\ell-1}$ are all distinct but that any c with a later index is equal to one of these ℓ distinct c's. For each $k < \ell$ we put $I_k = \{i \mid c_i = c_k\}$. Then we can reorganize the sum above as

$$\begin{aligned} 0 &= \sum_{i\in I_0}(a_i b_i)c_0 + \sum_{i\in I_1}(a_i b_i)c_1 + \cdots + \sum_{i\in I_{\ell-1}}(a_i b_i)c_{\ell-1} \\ &= \left(\sum_{i\in I_0} a_i b_i\right)c_0 + \left(\sum_{i\in I_1} a_i b_i\right)c_1 + \cdots + \left(\sum_{i\in I_{\ell-1}} a_i b_i\right)c_{\ell-1}. \end{aligned}$$

Because C is linearly independent (over $\mathbf{K}$) and because $c_0, \ldots, c_{\ell-1}$ are distinct, we find, for each $k < \ell$, that

$$0 = \sum_{i\in I_k} a_i b_i.$$

Now suppose $i, j \in I_k$ and $i \neq j$. So we know that $b_i c_i \neq b_j c_j$ but also that $c_i = c_j = c_k \neq 0$, with the last $\neq$ following because 0 cannot be in any linearly independent set like C. So we see that $b_i c_i \neq b_j c_i$ and $c_i \neq 0$. Dividing away the c_i, we conclude that $b_i \neq b_j$. This means that the b_i's occurring in $0 = \sum_{i\in I_k} a_i b_i$ are all distinct. Since B is linearly independent (over $\mathbf{F}$) we find that $a_i = 0$ for all $i \in I_k$ and for all $k < \ell$. This means that $a_i = 0$ for all $i < n$, and the set BC is linearly independent, as desired.

Contention. The map from $B \times C$ to BC that sends (b, c) to bc for all $(b, c) \in B \times C$ is a one-to-one correspondence.

According to the definition of BC, this map is onto BC. So it remains to show that it is one-to-one. So pick $b_0, b_1 \in B$ and $c_0, c_1 \in C$ so that $b_0c_0 = b_1c_1$. We need to show that $b_0 = b_1$ and $c_0 = c_1$. We note that none of b_0, b_1, c_0, and c_1 can be 0 since 0 belongs to no linearly independent set. There are two cases: either $c_0 = c_1$ or else $c_0 \neq c_1$. In the first case we can cancel the c's from $b_0c_0 = b_1c_1$ to obtain as well that $b_0 = b_1$, our desire. In the second case we see that $b_0c_0 - b_1c_1 = 0$. Since in this case c_0 and c_1 are distinct and linearly independent, we find that $b_0 = -b_1 = 0$, which we have already observed is impossible. So we must reject the second case.

This establishes the Dimension Formula. □

Notice that since the product on any two infinite cardinals is always an infinite cardinal (in fact, the larger of the two), we see that in the Dimension Formula, $[\mathbf{L} : \mathbf{F}]$ is infinite if and only if at least one of $[\mathbf{L} : \mathbf{K}]$ and $[\mathbf{K} : \mathbf{F}]$ is infinite.

The following fact will be useful in our ensuing work.

Fact. Let the field $\mathbf{K}$ extend the field $\mathbf{F}$ and suppose that $[\mathbf{K}:\mathbf{F}]$ is finite. Then $\mathbf{K}$ is an algebraic extension of $\mathbf{F}$.

Proof. Let $s \in K$. Since $\mathbf{K}$ is a finite dimensional vector space over $\mathbf{F}$ it cannot happen that all the elements on the list below are distinct and linearly independent:

$$1, s, s^2, s^3, s^4, \ldots .$$

This means that there must be elements $a_0, a_1, \ldots, a_n \in F$ that are not all 0 such that

$$a_0 + a_1 s^1 + a_2 s^2 + \cdots + a_n s^n = 0$$

It does no harm to suppose that $a_n \neq 0$. Notice that $n \neq 0$. So we see that s is a root of the polynomial $a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n \in F[x]$ of positive degree. Therefore s is algebraic over $\mathbf{F}$. □

Extensions like the one in this Fact are called **finite extensions**. Another way to frame the Fact is “Finite extensions are algebraic extensions”.

The Algebraic Extension Theorem. *Let the field* $\mathbf{L}$ *be an algebraic extension of the field* $\mathbf{K}$ *and let* $\mathbf{K}$ *be an algebraic extension of the field* $\mathbf{F}$*. Then* $\mathbf{L}$ *is an algebraic extension of* $\mathbf{F}$.

Proof. Let $s \in L$. Since $\mathbf{L}$ is an algebraic extension of $\mathbf{K}$, we pick $c_0, c_1, \ldots, c_n \in K$ with $n > 0$ and $c_n \neq 0$ such that s is a root of $c_0 + c_1 x + \cdots + c_n x^n$. Now let

$$\begin{aligned} \mathbf{K}_0 &= \mathbf{F}[c_0] \\ \mathbf{K}_1 &= \mathbf{K}_0[c_1] \\ &\vdots \\ \mathbf{K}_n &= \mathbf{K}_{n-1}[c_n]. \end{aligned}$$

Using the Dimension Formula repeatedly we find

$$[\mathbf{K}_n : \mathbf{F}] = [\mathbf{K}_n : \mathbf{K}_{n-1}][\mathbf{K}_{n-1} : \mathbf{K}_{n-2}] \cdots [\mathbf{K}_1 : \mathbf{K}_0].$$

But we know each of the dimensions on the right is finite. So $[\mathbf{K}_n : \mathbf{F}]$ is finite. But s is a root of a polynomial in $\mathbf{K}_n[x]$. So $[\mathbf{K}_n[s] : \mathbf{K}_n]$ is finite. Invoking the Dimension Formula one more time yields that $[\mathbf{K}_n[s] : \mathbf{F}]$ is finite. Since $s \in K_n[s]$, we see that s is algebraic over $\mathbf{F}$, which is just what we want. □

There is one more thing to do. Tackle the uniqueness of splitting fields.

The Basic Fact about Extending Isomorphisms. *Let* $\mathbf{F}$ *and* $\mathbf{F}^*$ *be fields. Let* Φ *be an isomorphism from* $\mathbf{F}$ *onto* $\mathbf{F}^*$*. Let* $\mathbf{L}$ *be a field extending* $\mathbf{F}$ *and let* $\mathbf{L}^*$ *be a field extending* $\mathbf{F}^*$*. Let* $s \in L$ *be algebraic over* $\mathbf{F}$ *with minimal polynomial* $f(x)$*. Then* Φ *can be extended to an embedding of* $\mathbf{F}[s]$ *into* $\mathbf{L}^*$ *if and only if* $f^*(x)$ *has a root in* $\mathbf{L}^*$*, in which case the number of distinct extensions of* Φ *is the same as the number of distinct roots of* $f^*(x)$ *in* $\mathbf{L}^*$*. (Here* $f^*(x) \in F^*[x]$ *is obtained from* $f(x)$ *by applying* Φ *to each of its coefficients.)*

Proof. Suppose first that Φ has an extension Ψ. Applying Ψ to the equation $f(r) = 0$, which is true in $\mathbf{L}$, gives $f^*(\Psi(s)) = 0$, which is true in $\mathbf{L}^*$. Hence $\Psi(s) \in L^*$ is a root of $f^*(x)$.

On the other hand, let r be any root of $f^*(x)$ that belongs to L^*. We apply Kronecker’s Theorem twice. So we have isomorphisms

$$\begin{aligned} \Lambda &: \mathbf{F}[x]/(f(x)) \rightarrowtail\!\!\!\!\twoheadrightarrow \mathbf{F}[s] \text{ with } \Lambda(x/(f(x))) = s \\ \Theta &: \mathbf{F}^*[x]/(f^*(x)) \rightarrowtail\!\!\!\!\twoheadrightarrow \mathbf{F}^*[r] \text{ with } \Theta(x/(f^*(x))) = r. \end{aligned}$$

But the isomorphism Φ from $\mathbf{F}$ to $\mathbf{F}^*$ induces an isomorphism $\Pi : \mathbf{F}[x]/(f(x)) \rightarrowtail\!\!\!\!\!\twoheadrightarrow \mathbf{F}^*[x]/(f^*(x))$. Putting things together we get an isomorphism $\Theta \circ \Pi \circ \Lambda^{-1}$ from $\mathbf{F}[s]$ onto $\mathbf{F}[r]$ that sends s to r and extends Φ. Since every embedding from $\mathbf{F}[s]$ into $\mathbf{L}^*$ that extends Φ is determined by what image it gives to s, we see that there are precisely as many extensions of Φ as there are distinct roots of $f^*(x)$ in L^*. □

Existence and Uniqueness of Splitting Fields. *Let* $\mathbf{F}$ *be a field and* $f(x) \in F[x]$ *be a polynomial of degree* $n > 0$. *Then there is a field* $\mathbf{E}$ *extending* $\mathbf{F}$ *such that*

(a) $\mathbf{E}$ *is a splitting field of* $f(x)$ *over* $\mathbf{F}$,

(b) $[\mathbf{E} : \mathbf{F}] \leq n!$, *and*

Moreover, suppose that $\mathbf{E}$ *and* $\mathbf{E}^*$ *are splitting fields of* $f(x)$ *over* $\mathbf{F}$. *Then*

(c) $\mathbf{E}$ *and* $\mathbf{E}^*$ *are isomorphic via an isomorphism that fixes each element of* F, *and*

(d) *The number of isomorphisms from* $\mathbf{E}$ *onto* $\mathbf{E}^*$ *that fix each element of* F *is no greater than* $[\mathbf{E} : \mathbf{F}]$ *and it is equal to* $[\mathbf{E} : \mathbf{F}]$ *if* $f(x)$ *has* n *distinct roots in* E.

Proof. Let us prove the existence part by induction of n. The base step is that $f(x) = ax + b$ where $a, b \in F$ and $a \neq 0$. So $\frac{b}{a} \in F$ is a root of $f(x)$. So we take $\mathbf{E} = \mathbf{F}$.

For the induction step we take $f(x)$ to be a polynomial of degree $k+1$ and we assume the (existence parts of) the theorem *over arbitrary fields* for $n < k+1$. Let $p(x) \in F[x]$ be an irreducible factor of $f(x)$. According to Kronecker's Theorem there is an extension $\mathbf{E}$ of $\mathbf{F}$ and an $s \in K$ so that s is a root of $f(x)$. So in $\mathbf{F}[s]$ we can factor $f(x) = (x - s)g(x)$ where $g(x) \in F[s][x]$ has degree k. Using the induction hypothesis we obtain a splitting field $\mathbf{E}$ of $g(x)$ over $\mathbf{F}[s]$ such that $[\mathbf{E} : \mathbf{F}[s]] \leq k!$.

In $\mathbf{E}$ we see that $f(x)$ factors into a product of polynomials of degree one and that the roots of $f(x)$ in E consist of the element s and the roots of $g(x)$. Because $\mathbf{E}$ is a splitting field of $g(x)$ over $\mathbf{F}[s]$ we know that it is generated by $F[s] \cup R$, where R is the set of all roots of $g(x)$ in E. But $\mathbf{F}[s]$ is generated by $F \cup \{s\}$. So $F \cup \{s\} \cup R$ generates $\mathbf{E}$. In this way we see that $\mathbf{E}$ is a splitting field of $f(x)$ over $\mathbf{F}$. So condition (a) is met in the inductive step.

We know, by Kronecker, that $[\mathbf{F}[s] : \mathbf{F}]$ is the degree of the irreducible factor $p(x)$ of $f(x)$. So the degree of $p(x) \leq k+1$, which is the degree of $f(x)$. By the Dimension Formula, we see

$$\begin{aligned}[\mathbf{E} : \mathbf{F}] &= [\mathbf{E} : \mathbf{F}[s]][\mathbf{F}[s] : \mathbf{F}] \\ &\leq k!(k+1) = (k+1)!\end{aligned}$$

So condition (b) is met in the inductive step.

For the rest, it proves more convenient to prove something a bit stronger.

Let $\mathbf{F}^*$ be a field and $\Phi : \mathbf{F} \rightarrowtail\!\!\!\!\!\twoheadrightarrow \mathbf{F}^*$. Let $f^*(x)$ be the polynomial over $\mathbf{F}^*$ obtained from $f(x)$ by applying Φ to each coefficient.

Now suppose $\mathbf{E}$ and $\mathbf{E}^*$ are splitting fields of $f(x)$ over $\mathbf{F}$ and of $f^*(x)$ over $\mathbf{F}^*$ respectively. Instead of considering maps that fix each element of F (i.e. those extending the identity map on F) we consider maps extending Φ.

We proceed by induction on $[\mathbf{E} : \mathbf{F}]$.

For the base step, we will have $F = E$ and $f(x)$ factors into polynomials of degree 1 over $\mathbf{F}$. So all the roots of $f(x)$ lie in F. Likewise for $f^*(x) \in \mathbf{F}^*[x]$. Since the roots of $f(x)$ together with F itself generate $\mathbf{E}$ and likewise for $f^*(x)$ and $\mathbf{E}^*$, we see that $\mathbf{E} = \mathbf{F} \overset{\Phi}{\rightarrowtail\!\!\!\!\!\twoheadrightarrow} \mathbf{F}^* = \mathbf{E}^*$ and there is only one isomorphism between $\mathbf{E}$ and $\mathbf{E}^*$ that extends Φ, namely the map Φ itself. So the appropriately modifications of conditions (c) and (d) both hold in the base step.

Now we turn to the inductive step, where we have $[\mathbf{E}:\mathbf{F}] > 1$. So there must be a root $r \in E$ that does not belong to F. Let $p(x)$ be the minimal polynomial of r_0. The degree of $p(x)$ must be at least 2. Since we see that $p(x)$ is a factor of $f(x)$ so the corresponding $p^*(x)$ is a factor of $f^*(x)$ and $p^*(x)$ must split in $\mathbf{E}^*$. Let us say that it has $m > 1$ roots in $\mathbf{E}^*$. Then by our basic fact about extending isomorphisms there are exactly m distinct extensions of Φ to embeddings of $\mathbf{F}[r]$ into $\mathbf{E}^*$. Consider one of them Φ' and let $s = \Phi'(r)$. Now $[\mathbf{E}:\mathbf{F}] = [\mathbf{E}:\mathbf{F}[r]][\mathbf{F}[r]:\mathbf{F}]$ since $[\mathbf{F}[r]:\mathbf{F}]$ is the degree of $p(x)$, which is at least 2, we see that $[\mathbf{F}:\mathbf{F}[r]] < [\mathbf{E}:\mathbf{F}]$. But $\mathbf{E}$ is a splitting field of $f(x)$ over $\mathbf{F}[r]$ and $\mathbf{E}^*$ is a splitting field of $f^*(x)$ over $\mathbf{F}^*[s]]$. So we can appeal to the induction hypothesis to get at least one extension of Φ' to an isomorphism between $\mathbf{E}$ and $\mathbf{E}^*$. Evidently, such an isomorphism also extends Φ and we obtain, in the inductive step, the appropriate modification of condition (c).

The induction hypothesis also tells us that the number of such extensions of Φ' is not greater than $[\mathbf{E}:\mathbf{F}[r]]$ and is equal to $[\mathbf{E}:\mathbf{F}[r]]$ if the number of distinct roots of $f(x)$ in E coincides with the degree of $f(x)$. Recall that Φ' was one of the m extensions of Φ that embed $\mathbf{F}[r]$ into $\mathbf{E}^*$. So the number of extensions of Φ to an isomorphism between $\mathbf{E}$ and $\mathbf{E}^*$ is no greater than the product of $[\mathbf{E}:\mathbf{F}[r]]$ and m. But m, the number of distinct roots of $p^*(x)$ in $\mathbf{E}^*$, can be no greater than the degree of $p(x)$, which we know is $[\mathbf{F}[r]:\mathbf{F}]$. So $[\mathbf{E}:\mathbf{F}] = [\mathbf{E}:\mathbf{F}[r]][\mathbf{F}[r]:\mathbf{F}]$ is an upper bound on the number of ways Φ can be extended to an isomorphism between $\mathbf{E}$ and $\mathbf{E}^*$. Last suppose that $f(x)$ has distinct roots. Then so must $p(x)$. This means that $m = [\mathbf{F}[r]:\mathbf{F}]$. In this case we know there are precisely $[\mathbf{F}[r]:\mathbf{F}]$ ways to extend Φ to an embedding of $\mathbf{F}[r]$ into $\mathbf{E}^*$ and, for each such extension, there are precisely $[\mathbf{E}:\mathbf{F}[r]]$ ways to extend it to an isomorphism between $\mathbf{E}$ and $\mathbf{E}^*$. So the number of extensions of Φ to an isomorphism from $\mathbf{E}$ onto $\mathbf{E}^*$ is precisely $[\mathbf{E}:\mathbf{F}]$ by the Dimension Formula. So condition (d) holds in the inductive step.

The proof is complete. □

19.2 Transcendental Extensions of Fields

While our main effort is centered on describing the roots of polynomials (and so describing the extensions of fields by adjoining algebraic elements), it is useful to give some consideration to forming field extensions by adjoining transcendental elements. Lecture 29 is devoted to the Lindemann-Weierstrass Theorem. One consequence of that theorem is that the real number π is transcendental over the field $\mathbb{Q}$ of rational numbers. What does the subfield $\mathbb{Q}(\pi)$ of the field of real numbers look like? Well, it is the subfield generated by $\mathbb{Q} \cup \{\pi\}$. So for each polynomial $f(x) \in \mathbb{Q}[x]$, the number $f(\pi)$ belongs to $\mathbb{Q}(\pi)$. Moreover, so long as $f(x)$ is not the zero polynomial, $\frac{1}{f(\pi)}$ will also belong to $\mathbb{Q}(\pi)$, since then $f(\pi) \neq 0$ because π is transcendental and cannot be a root of $f(x)$. Consider the set

$$\left\{ \frac{f(\pi)}{g(\pi)} \;\middle|\; f(x), g(x) \in \mathbb{Q}[x] \text{ with } g(x) \text{ not the zero polynomial} \right\}.$$

I leave it in the capable hands of the graduate students to check that this set of real numbers is actually a subfield of $\mathbb{R}$—so it must be $\mathbb{Q}(\pi)$. But notice that the set displayed above is very similar to

$$\left\{ \frac{f(x)}{g(x)} \;\middle|\; f(x), g(x) \in \mathbb{Q}[x] \text{ with } g(x) \text{ not the zero polynomial} \right\}.$$

You can see that this is just the field of fractions of the integral domain $\mathbb{Q}[x]$. It should come as no surprise (and as no real bother to check) that the evalutation map

$$\frac{f(x)}{g(x)} \mapsto \frac{f(\pi)}{g(\pi)}$$

is a homomorphism from the field of fractions onto $\mathbb{Q}(\pi)$. Since fields are simple, this map must actually be an isomorphism. Now the only feature of π that was important in the discussion was that π is transcendental over $\mathbb{Q}$. It follows that if $a, b \in \mathbb{R}$ so that a and b are transcendental over $\mathbb{Q}$, then $\mathbb{Q}(a)$ and $\mathbb{Q}(b)$ are isomorphic via an isomorphism that fixes each element of $\mathbb{Q}$. But no special properties of $\mathbb{Q}$ or of $\mathbb{R}$ came into the discussion above either.

This means we can have a version of the Basic Fact about Extending Isomorphisms that applies to transcendental elements. Here it is

The Basic Fact about Extending Isomorphisms, Transcendental Version. *Let* $\mathbf{F}$ *and* $\mathbf{F}^*$ *be fields. Let* Φ *be an isomorphism from* $\mathbf{F}$ *onto* $\mathbf{F}^*$. *Let* $\mathbf{L}$ *be a field extending* $\mathbf{F}$ *and let* $\mathbf{L}^*$ *be a field extending* $\mathbf{F}^*$. *Let* $s \in L$ *be transcendental over* $\mathbf{F}$ *and let* $s^* \in L^*$ *be transcendental over* $\mathbf{F}^*$. *Then* Φ *can be extended to an isomorphism of* $\mathbf{F}(s)$ *onto* $\mathbf{F}^*(s^*)$ *that sends* s *to* s^*.

LECTURE 20

Algebraically Closed Fields

A field $\mathbf{F}$ is said to be **algebraically closed** provided irreducible polynomials in $\mathbf{F}[x]$ coincide with the polynomials of degree 1. This is the same as saying that every polynomial in $\mathbf{F}[x]$ of positive degree has a root in $\mathbf{F}$. It is also evidently equivalent to the requirement that $\mathbf{F}$ has no proper algebraic extensions.

Neither the field $\mathbb{Q}$ of rational numbers nor the field $\mathbb{R}$ of real numbers is algebraically closed. It is a nontrivial fact (which we will prove later in the semester) that the field $\mathbb{C}$ of complex numbers is algebraically closed.

20.1 Algebraic Closures

An extension $\mathbf{K}$ of the field $\mathbf{F}$ is an **algebraic closure** of $\mathbf{F}$ provided $\mathbf{K}$ is an algebraically closed algebraic extension of $\mathbf{F}$.

Consider how we might arrive at an algebraic closure of the field $\mathbb{Q}$ of rational numbers. We might begin by making a list of all the polynomials of positive degree in $\mathbb{Q}[x]$. This list is countably infinite and it takes a bit of work to arrange these polynomials like the natural numbers are arranged. But image we have made this list: $f_0(x), f_1(x), f_2(x), \ldots$. Now we could proceed by letting $\mathbf{F}_0$ be the splitting field of $f_0(x)$ over $\mathbb{Q}$. Next, we let $\mathbf{F}_1$ be the splitting field of $f_1(x)$ over $\mathbf{F}_0$. We continue in this way to split, one after another, all the polynomials on our list. We get in this way a chain of fields, each extending the one before. Fearlessly, we form the union of this chain of fields to arrive at $\mathbf{K}_0$. A little thought shows us that $\mathbf{K}_0$ is an algebraic extension of $\mathbb{Q}$, that it is generated over $\mathbb{Q}$ by the roots of all those polynomials, and that all those polynomials split in $\mathbf{K}_0$. Unfortunately, along the way we have added a lot of new elements and these new elements can be coefficients of polynomials in $\mathbf{K}_0[x]$ that haven't yet been addressed. So now we must list all of these polynomials, build another infinite chain of splitting fields, and finally arrive at the union $\mathbf{K}_1$. Now many more polynomials have been split but many more new elements have also been introduced. But we continue anyhow to construct $\mathbf{K}_2$, then $\mathbf{K}_3, \ldots$. Finally, we take one last union of this chain to obtain the field $\mathbb{A}$. We would be able to show that $\mathbb{A}$ is an algebraically closed algebraic extension of $\mathbb{Q}$ and even that $\mathbb{A}$ is countably infinite. The idea behind this proof sketch could be made to work starting with any field (although the countability of the algebraic closure has to be modified if the field we start with is uncountable). In general, this construction requires making a lot of choices along the way (particularly, choices about how to order the polynomials at each step).

We can avoid the complexity of this construction by invoking Zorn's Lemma.

The Existence Theorem for Algebraic Closures. *Every field has an algebraic closure.*

Proof. The basic idea is that we will take $\mathcal{F}$ to be the collection of all algebraic extensions of the given field $\mathbf{F}$. Using Zorn's Lemma we will extract a maximal member of $\mathcal{F}$ that will turn out to be an algebraic extension of $\mathbf{F}$ that is algebraically closed. There is one stumbling block to this scheme: the collection $\mathcal{F}$ turns out to be too large and wild to be a set.

Let $\mathbf{F}$ be a field. Let U be an uncountably infinite set of cardinality properly larger than the cardinality of F so that $F \subseteq U$. Take $\mathcal{F}$ to be the collection of all algebraic extensions $\mathbf{K}$ of $\mathbf{F}$ so that $K \subseteq U$.

To invoke Zorn's Lemma, consider any nonempty chain $\mathcal{C} \subseteq \mathcal{F}$. We contend $\mathcal{C}$ has an upper bound in $\mathcal{F}$. Indeed, let $L = \bigcup\{K \mid \mathbf{K} \in \mathcal{C}\}$. Of course $0, 1 \in L$. We impose $+$ and $\cdot$ on L in the natural way: for $a, b \in L$, using the fact that $\mathcal{C}$ is a chain, pick $\mathbf{K} \in \mathcal{C}$ so that $a, b \in K$. Take $a + b$ and $a \cdot b$ in $\mathbf{L}$ just as they are understood in $\mathbf{K}$. (The hard-working graduate students should confirm that the particular $\mathbf{K}$ chosen works no bad idiosyncratic influence here.) This, of course, is just another case of the union of a chain of algebraic entities resulting in another entity of the same kind. It is straightforward to provide the details showing that $\mathbf{L}$ is a field that extends $\mathbf{F}$ and of course $L \subseteq U$. It is also clear the $\mathbf{L}$ is an upper bound of $\mathcal{C}$. To conclude that $\mathbf{L} \in \mathcal{F}$, we need to show that $\mathbf{L}$ is also an algebraic extension of $\mathbf{F}$. But this is clear: let $a \in L$ and pick $\mathbf{K} \in \mathcal{C} \subseteq \mathcal{F}$ so that $a \in K$. Since $\mathbf{K} \in \mathcal{F}$ it is an algebraic extension of $\mathbf{F}$. So a is a root of a polynomial in $\mathbf{F}[x]$. That is, a is algebraic over $\mathbf{F}$, as desired.

So Zorn's Lemma provides us with a field $\mathbf{M}$ that is a maximal element of $\mathcal{F}$. In particular, $\mathbf{M}$ is an algebraic extension of $\mathbf{F}$. Now the idea is to take any irreducible polynomial $p(x) \in \mathbf{M}[x]$. Applying Kronecker's Theorem, we obtain an algebraic extension $\mathbf{M}[r]$ of $\mathbf{M}$ so that r is a root of $p(x)$. We know that an algebraic extension of $\mathbf{M}$ must be an algebraic extension of $\mathbf{F}$, since $\mathbf{M}$ is an algebraic extension of $\mathbf{F}$. Were we able to appeal to the maximality of $\mathbf{M}$, we would conclude that $M = M[r]$, so that $r \in M$ and the arbitrary irreducible polynomial $p(x)$ has a root in $\mathbf{M}$. Thus $\mathbf{M}$ would be algebraically closed.

The point of difficulty is that $M[r]$ might not be contained in U.

This would present no trouble if there were enough room in $U \setminus M$ to fit in a copy of $M[r] \setminus M$. So what is the size of $M \setminus M[r]$? Well, we know that $\mathbf{M}[r]$ is a vector space over $\mathbf{M}$ with dimension equal to the degree d of the minimal polynomial of r. So, as with any finite dimensional vector space, we find $|M[r]| = |M|^d$. So we see that $|M[r] \setminus M| \leq |M[r]| = |M|^d$. We could argue (maybe the curious graduate student will do it) that M must be infinite. We take $\kappa = |M|$ if M is infinite (as it is) and otherwise take κ to be the smallest infinite cardinal (namely $|\mathbb{N}|$). One useful fact from the arithmetic of infinite cardinals is that $\kappa \cdot \kappa = \kappa$. So a touch of induction shows that $|M[r] \setminus M| \leq \kappa$.

How big is $|U \setminus M|$? Every element of M is a root of some irreducible polynomial in $\mathbf{F}[x]$ and each such polynomial has only finitely many roots. How many polynomials are there? The zero polynomial together with the polynomials of degree 0 make up F. So $|F| = |F|^1$ is an upper bound on this collection. The polynomials of degree 1 each have two coefficients. So $|F| \cdot |F| = |F|^2$ is an upper bound on the number of these polynomials. In general, $|F|^d$ bounds the number of polynomials of degree d. Now each polynomial of degree d can have at most d distinct roots in M. Altogether, we set that

$$\sum_{1 \leq d < \omega} d|F|^d$$

is an upper bound on the number of elements of M. Let $\mu = |F|$ if F is infinite and let μ be the least infinite cardinal ω otherwise. Then

$$\sum_{1 \leq d < \omega} d|F|^d \leq \sum_{1 \leq d < \omega} d\mu^d = \sum_{1 \leq d < \omega} \mu \leq \omega \cdot \mu = \mu.$$

Our choice of the size of U at the beginning of the proof ensures that $|U| > \mu$ and hence that

$$|U \setminus M| = |U| > \mu \geq \kappa \geq |M[r] \setminus M|.$$

This means that there is enough room left over in U, after M is in hand, to construct a copy of $\mathbf{M}[r]$. Now we can really appeal to the maximality of $\mathbf{M}$ to complete the proof. □

The Uniqueness Theorem for Algebraic Closures.

The Basic Fact about Extending Isomorphisms. *Let* $\mathbf{F}$ *and* $\mathbf{F}^*$ *be fields. Let* Φ *be an isomorphism from* $\mathbf{F}$ *onto* $\mathbf{F}^*$. *Let* $\mathbf{L}$ *be a field extending* $\mathbf{F}$ *and let* $\mathbf{L}^*$ *be a field extending* $\mathbf{F}^*$. *Let* $s \in L$ *be algebraic over* $\mathbf{F}$ *with minimal polynomial* $f(x)$. *Then* Φ *can be extended to an embedding of* $\mathbf{F}[s]$ *into* $\mathbf{L}^*$ *if and only if* $f^*(x)$ *has a root in* $\mathbf{L}^*$, *in which case the number of distinct extensions of* Φ *is the same as the number of distinct roots of* $f^*(x)$ *in* $\mathbf{L}^*$. *(Here* $f^*(x) \in F * [x]$ *is obtained from* $f(x)$ *by applying* Φ *to each of its coefficients.)*

Let $\mathbf{F}$ *be a field and let* $\mathbf{A}$ *and* $\mathbf{K}$ *be algebraic extensions of* $\mathbf{F}$ *which are algebraically closed. Then there is an isomorphism from* $\mathbf{A}$ *onto* $\mathbf{K}$ *which fixes each element of* $\mathbf{F}$.

Proof. Let $\mathcal{I}$ be the set of all isomorphisms with domains which are subfields of $\mathbf{A}$ that extend $\mathbf{F}$, whose images are subfields of $\mathbf{K}$ that extend $\mathbf{F}$, and which fix every element of $\mathbf{F}$.

Recalling that each function is a set of ordered pairs, we see that $\mathcal{I}$ is partially ordered by $\subseteq$. It is easy to see that this ordering is the same as the ordering by extension of functions.

To invoke Zorn's Lemma, we need to see that any chain $\mathcal{C}$ in $\mathcal{I}$ has an upper bound. If $\mathcal{C}$ is empty, then the identity function of $\mathbf{F}$ is an upper bound of $\mathcal{C}$ and it belongs to $\mathcal{I}$. Consider the case when $\mathcal{C}$ is not empty. Let $\Phi = \bigcup \mathcal{C}$.

Claim. Φ is a function.

Proof. Suppose $(a,b),(a,c) \in \Phi$. Pick $\varphi,\psi \in \mathcal{C}$ so that $(a,b) \in \varphi$ and $(a,c) \in \psi$. Since $\mathcal{C}$ is a chain, either $\varphi \subseteq \psi$ or $\psi \subseteq \varphi$. Without loss of generality, let us suppose that $\psi \subseteq \varphi$. Then $(a,b),(a,c) \in \varphi$. But φ is a function, so $b = c$. Consequently, Φ is a function. □

Claim. The domain of Φ is a subfield of $\mathbf{A}$ which extends $\mathbf{F}$.

Proof. A routine argument shows that $\operatorname{dom}\Phi = \bigcup\{\operatorname{dom}\varphi \mid \varphi \in \mathcal{C}\}$. Since $\mathcal{C}$ is not empty and $\mathbf{F} \subseteq \operatorname{dom}\varphi$ for each $\varphi \in \mathcal{C}$, we see that $\mathbf{F} \subseteq \operatorname{dom}\Phi$. Let $a,b \in \operatorname{dom}\Phi$. As above, pick $\varphi \in \mathcal{C}$ so that $a,b \in \operatorname{dom}\varphi$. Since $\operatorname{dom}\varphi$ is a subfield of $\mathbf{A}$, we see that $a+b, ab \in \operatorname{dom}\varphi \subseteq \operatorname{dom}\Phi$ and also that $a^{-1} \in \operatorname{dom}\Phi$ if $a \neq 0$. This means that $\operatorname{dom}\Phi$ is a subfield of $\mathbf{A}$. □

Claim. Φ is a homomorphism.

Proof. Let $a,b \in \operatorname{dom}\Phi$. Pick $\varphi,\psi \in \mathcal{C}$ so that $a \in \operatorname{dom}\varphi$ and $b \in \operatorname{dom}\psi$. As above, without loss of generality we suppose that $a,b \in \operatorname{dom}\varphi$. Now φ is a homomorphism, so $(a+b,\varphi(a)+\varphi(b))$ and $(ab,\varphi(a)\varphi(b))$ both belong to φ, as do $(a,\varphi(a))$ and $(b,\varphi(b))$. But $\varphi \subseteq \Phi$. So those four ordered pairs belong to the function Φ. Translated into usual usage we have

$$\Phi(a+b) = \varphi(a+b) = \varphi(a)+\varphi(b) = \Phi(a)+\Phi(b)$$
$$\Phi(ab) = \varphi(ab) = \varphi(a)\varphi(b) = \Phi(a)\Phi(b).$$

We see, even more easily, that $\Phi(0) = 0$ and $\Phi(1) = 1$. So Φ is a homomorphism. □

Claim. Φ fixes each element of $\mathbf{F}$.

Proof. This is too easy to prove. □

Since Φ is a homomorphism from one field into another and it fixes each element of the subfield $\mathbf{F}$, we see it cannot collapse everything to one value. So it must be one-to-one (after all, fields are simple). So Φ is an isomorphism. All this, taken together, means that $\Phi \in \mathcal{I}$. So Φ is an upper bound of $\mathcal{C}$ as desired.

Invoking Zorn's Lemma, we see that $\mathcal{I}$ must have a maximal member. Let Ψ be such a maximal member. It remains to show that $\mathbf{A}$ is the domain of Ψ and that $\mathbf{K}$ is the image of Ψ.

The field $\mathbf{A}$ is an algebraic extension of $\operatorname{dom}\Psi$, because any element of $\mathbf{A}$ is a root of some polynomial of positive degree with coefficients in $\mathbf{F} \subseteq \operatorname{dom}\Psi$. Let $u \in \mathbf{A}$. Let $p(x) \in \operatorname{dom}\Psi[x]$ be an irreducible polynomial with root u. Say $p(x) = a_0 + a_1x + \cdots + a_nx^n$ with $a_n \neq 0$. Then the polynomial $\Psi(a_0) + \Psi(a_1)x + \cdots + \Psi(a_n)x^n$ is irreducible over the field that is the image of Ψ. Denote the image of Ψ by $\mathbf{B}$. But $\mathbf{K}$ is algebraically closed, so this polynomial must have a root v in $\mathbf{K}$. Consider the possibility that the degree of $p(x)$ is bigger than 1. In that event, $u \notin \operatorname{dom}\Psi$ and $v \notin \mathbf{B}$. Then we can extend Ψ to an isomorphism from $\operatorname{dom}\Psi[u]$ onto $\mathbf{B}[v]$. This extension also belongs to $\mathcal{I}$. In this way the maximality of Ψ is violated. So $p(x)$ must have degree 1. But this entails that $u \in \operatorname{dom}\Psi$. Since u was an arbitrary element of $\mathbf{A}$, we see that $\mathbf{A} = \operatorname{dom}\Psi$.

It remains to see that the image $\mathbf{B}$ of Ψ is $\mathbf{K}$. Since we have seen, by this point, that $\mathbf{B}$ is isomorphic to $\mathbf{A}$, and we know that $\mathbf{A}$ is algebraically closed, we conclude that $\mathbf{B}$ is also algebraically closed. Now $\mathbf{K}$ is an algebraic extension of $\mathbf{B}$. But algebraically closed fields cannot have proper algebraic extensions. So $\mathbf{K}$ is not a proper extension of $\mathbf{B}$. This means $\mathbf{B} = \mathbf{K}$, concluding our proof. □

In one of the problem sets you will be asked to prove that no algebraically closed field can be finite. In the proof of existence of algebraic closure we saw that the algebraic closure of a finite field in countably infinite. For any infinite field, our argument shows that the algebraic closure is the same cardinality as the original field. So the algebraic closure of the field of rational numbers is countably infinite.

The field of complex numbers turns out to be algebraically closed, a fact customarily referred to as the Fundamental Theorem of Algebra. Proofs of this fact were offered in the 18th century, notably by Euler, Lagrange, and Laplace. These proofs all had gaps. Roughly speaking, these gaps are filled by Kronecker's Theorem. In his 1799 doctoral dissertation, Gauss devoted a lot of space to picking out the flaws in these proofs, and then supplied a flawed proof of his own. (This proof by Gauss had an ingenious geometric turn—the gap was finally filled by Ostrowski in 1920). The first complete proof was given in 1806 by Argand. Gauss later gave two further proofs. In Lecture 27 you will find a proof due to Emil Artin and Otto Schreier from 1927.

The Fundamental Theorem of Algebra no longer plays a fundamental role in algebra. At some level, it is basically an analytical rather than an algebraic theorem, although we will give toward the end of these lectures a largely algebraic account. The quickest modern proofs appeal to theorems in complex analysis, for example to Liouville's Theorem.

As a consequence, the algebraic closure of the rationals can be construed as a subfield of the complex numbers.

20.2 On the Uniqueness of Uncountable Algebraically Closed Fields

Two algebraically closed fields might differ in some obvious way. They might have different characteristics. They might have different cardinalities: the algebraic closure of the rationals is countable, whereas the field of complex numbers is an algebraically closed field that is uncountable. Getting a bit more subtle, let $\mathbb{A}$ be the algebraic closure of the rationals, construed as a subfield of $\mathbb{C}$. The transcendental extension $\mathbb{A}(\pi)$ is again countable. Let $\mathbb{B}$ be the algebraic closure of $\mathbb{A}(\pi)$. So $\mathbb{A}$ and $\mathbb{B}$ will be countable algebraically closed fields of characteristic 0. But the are not isomorphic. We could go on to extend $\mathbb{B}$ in a similar way, to eventually obtain an increasing tower of field extensions, each countable and each algebraically closed and no two of them would be isomorphic. It is a striking discovery made by Ernst Steinitz in 1910 that this is basically the whole story. Here is one part of his discovery.

The Steinitz's Theorem on Isomorphisms between Algebraically Closed Fields. *Let* $\mathbf{F}, \mathbf{A}, \mathbf{E}$ *and* $\mathbf{B}$ *be algebraically closed fields so that* $\mathbf{F}$ *is a subfield of* $\mathbf{A}$ *and* $\mathbf{E}$ *is a subfield of* $\mathbf{B}$*. Further suppose that* Φ *is an isomorphism from* $\mathbf{F}$ *onto* $\mathbf{E}$*. If* $\mathbf{A}$ *and* $\mathbf{B}$ *have the same cardinality* κ *and* κ *is larger than the cardinality of* $\mathbf{F}$*, then there is an isomorphism* Φ^* *from* $\mathbf{A}$ *onto* $\mathbf{B}$ *that extends* Φ*. Thus, any two uncountable algebraically closed fields of the same characteristic and the same cardinality are isomorphic.*

Proof. I would like to emulate the proof of the uniqueness of the algebraic closure and employ Zorn's Lemma. However, the situation at hand is a bit more delicate and Zorn's Lemma seems too blunt an instrument. Instead, the desired isomorphism is built in stages. The proof itself requires a very modest understanding of cardinal and ordinals.

To start, let κ be the common cardinality of $\mathbf{A}$ and $\mathbf{B}$. The first little bit of knowledge about cardinals and ordinals is that κ itself is an ordinal and consists of the set of all ordinals of cardinality less than κ. Thus κ itself is well-ordered and each of its proper initial segments has cardinality smaller than κ. So we have

$$\kappa = \{\alpha \mid \alpha < \kappa\}.$$

We use κ to enumerate both A and B. So $A = \{a_\alpha \mid \alpha \in \kappa\}$ and $B = \{b_\alpha \mid \alpha \in \kappa\}$. It does no harm to suppose that a_0 and b_0 are the zero's in their respective fields.

For any set X, we use $|X|$ to denote the cardinality of X. The addition of cardinals is just the cardinality of there disjoint union. So $|X| + |Y|$ is just the maximum of $|X|$ and $|Y|$, provided at least one of X and Y is infinite. The product of two cardinals is just the cardinality of their direct product. That is $|X||Y| := |X \times Y|$. In case at least one of these sets is infinite, it is true that $|X||Y|$ is the maximum of $|X|$ and $|Y|$. This fact is somewhat subtle—its proof lies outside the scope of this book.

The construction here has a stage for each ordinal $\alpha \in \kappa$. At Stage α we will have in hand for each earlier Stage β:

(a) two algebraically closed subfields $\mathbf{A}_\beta$ of $\mathbf{A}$ and $\mathbf{B}_\beta$ of $\mathbf{B}$, so that for each ordinal $\gamma < \beta$, $\mathbf{A}_\beta$ extends $\mathbf{A}_\gamma$ and $\mathbf{B}_\beta$ extends $\mathbf{B}_\gamma$;

(b) an isomorphism Φ_β between $\mathbf{A}_\beta$ and $\mathbf{B}_\beta$ that extends all earlier isomorphisms in the construction;

(c) the common cardinality of A_β and B_β will be no more than $|\beta| + |F|$;

(d) for each $\beta < \alpha$ we will have that $a_\beta \in A_\beta$ and $b_\beta \in B_\beta$.

The constraint (c) concerning cardinalities is perhaps obscure, but it ensures that the construction at Stage α can be carried out. The task at Stage α is to produce $\mathbf{A}_\alpha$ and $\mathbf{B}_\alpha$ and an isomorphism Φ_α between them to keep the construction going by fulfilling for α all the attributes of the construction attributed to β above.

We take $\mathbf{A}_0 = \mathbf{F}$ and $\mathbf{B}_0 = \mathbf{E}$ in $\mathbf{A}$ and $\Phi_0 = \Phi$. In this way, Stage 0 of our construction is completed.

To finish describing the construction by stages we need another little bit of knowledge about ordinals and cardinals. Ordinals (which are representatives up to isomorphism of well-orderings) come in two kinds: successor ordinals and limit ordinals. Successor ordinals are those that have an immediate predecessor (and they represent those well-orderings with a largest element). Any other kind of ordinal is a limit ordinal. The natural numbers coincide with the finite ordinals. The natural number 0 is a limit ordinal and the other natural numbers are all successor ordinals. The whole set of natural numbers is the next ordinal after all the finite ordinals and it is a limit ordinal.

There are two kinds of stages, depending on whether α is a limit ordinal or an successor ordinal.

The Construction at Stage α, in case the ordinal α is a limit ordinal
We already have the case $\alpha = 0$ in hand. So consider that α is an infinite limit ordinal. We let

$$\begin{aligned} A'_\alpha &= \bigcup_{\beta<\alpha} A_\beta \\ B'_\alpha &= \bigcup_{\beta<\alpha} B_\beta \\ \Phi'_\alpha &= \bigcup_{\beta<\alpha} \Phi_\beta \end{aligned}$$

Recalling that $\langle \mathbf{A}_\beta \mid \beta < \alpha\rangle$ and $\langle \mathbf{B}_\beta \mid \beta < \alpha\rangle$ are towers of field extensions, the diligent graduate student should find it a straightforward matter to check that the sets A'_α and B'_α are closed under all the operations (including forming inverses of nonzero elements). This gives us fields $\mathbf{A}'_\alpha$ and $\mathbf{B}'_\alpha$. A bit more effort should reveal that Φ'_α is an isomorphism between these two fields. Finally, a little more work shows that each of the fields $\mathbf{A}'_\alpha$ and $\mathbf{B}'_\alpha$ is algebraically closed. Now here is the last bit a knowledge about cardinals and ordinals that we need. Since for each $\beta < \alpha$ we know that the cardinality $|A'_\beta|$ of A_β isnolarger than $|\beta| + |F|$, then we know that

$$|A'_\alpha| = |\bigcup_{\beta<\alpha} A_\beta| \leq \sum_{\beta<\alpha} |A_\beta| \leq \sum_{\beta<\alpha} (|\beta| + |F|) \leq \sum_{\beta<\alpha} (|\alpha| + |F|) \leq |\alpha|(|\alpha| + |F|) = |\alpha| + |F|.$$

That is $|A'_\alpha| \leq |\alpha| + |F|$. Likewise that $|B'_\alpha|$ is bounded in the same way. The bit of knowledge about cardinals then ensures us that $|A'_\alpha|$ and $|B'_\alpha|$ must both be smaller than κ. According to how we have defined things here, we know that for each $\beta < \alpha$ that $a_\beta \in A'_\alpha$ and that $b_\beta \in B'_\alpha$. But we need a_α and b_α to be in the fields we construct at this stage. Now $|A'_\alpha| < \kappa$. This means that there are elements of A that are not in A'_α. Let c be the least such element in our enumeration of A. So $c = a_\alpha$, unless a_α is already in A'_α. Since $\mathbf{A}'_\alpha$ is algebraically closed, c is transcendental over $\mathbf{A}'_\alpha$. Likewise, let d be the least member of B not in B'_α. According to the transcendental version of the Basic Fact about Extending Isomorphisms, the fields $\mathbf{A}'_\alpha(c)$ and $\mathbf{B}'_\alpha(d)$ are isomorphic via an isomorphism Φ''_α that extends Φ'_α. Finally, let $\mathbf{A}_\alpha$ be the algebraic closure of $\mathbf{A}'_\alpha(c)$ and let $\mathbf{B}_\alpha$ be the algebraic closure of $\mathbf{B}'_\alpha(d)$ and let Φ_α be an isomorphism between these two algebraic closures that extends Φ''_α. The careful graduate student should check that these two fields and the isomorphism between them fulfill all the desires (a)–(d) we had listed above.

The Construction at Stage α, in case the ordinal α is a successor ordinal
In this case, take α to be the successor of β. Then we have in hand the algebraically closed fields $\mathbf{A}_\beta$ and $\mathbf{B}_\beta$ and an isomorphism Φ_β between them. Furthermore, we know that, for each ordinal $\gamma \leq \beta$ that $a_\gamma \in A_\beta$ and $b_\gamma \in B_\beta$. We also know that either both $|A_\beta|$ and $|B_\beta|$ are no larger than $|\beta| + |F| < \kappa$. So, as above, we let c be the least element of A not in A_β and d be the least element of B not in B_β. The elements c and d are transcendental over $\mathbf{A}_\beta$ and $\mathbf{B}_\beta$ respectively. Also as above, a_α will be a member of the extension $\mathbf{A}_\beta(c)$ and likewise b_α will be in the field $\mathbf{B}_\beta(d)$. So take $\mathbf{A}_\alpha$ to be the algebraic closure of $\mathbf{A}_\beta(c)$ in $\mathbf{A}$ and $\mathbf{B}_\alpha$ to be the algebraic closure of $\mathbf{B}_\beta(d)$ in $\mathbf{B}$. Then Φ_β can be extended to an isomorphism Φ_α from $\mathbf{A}_\alpha$ onto $\mathbf{B}_\alpha$. The remaining properties that we desire can be established through a small effort of the graduate student.

Now having the construction in hand, it only remains to show that

$$\Phi^* : = \bigcup_{\alpha<\kappa} \Phi_\alpha$$

is an isomorphism from $\mathbf{A}$ onto $\mathbf{B}$. Notice that the sequence $\langle \Phi_\alpha \mid \alpha < \kappa\rangle$ is a sequence of isomorphisms, each extending all the preceding ones. A small argument will establish that the Φ^* with the definition displayed above will be itself an isomorphism from a subfield of $\mathbf{A}$ onto a subfield of $\mathbf{B}$. But the construction was carefully designed to include all the a_β's in the domain of Φ^* and also to include all the b_β's among the outputs of Φ^*. So the domain of Φ is A and Φ^* maps A onto B, as desired. This establishes that **A** and **B** are isomorphic by an isomorphism that extends Φ.

Now consider the last contention in the theorem. Let **A** and **B** are uncountable algebraically closed fields of the same cardinality and with the same characteristic. Then each of these fields has a smallest subfield. In the case of characteristic p, these smallest subfields will be isomorpic the $\mathbb{Z}_p$. In the case of characteristic 0, the will both be isomorphic to $\mathbb{Q}$. Take **F** and **E** to be the algebraic closures of these smallest subfields (these algebraic clusores will be countably infinite) and let Φ be any isomorphism between them. This sets up the main part of the theorem and the conclusion follows. □

20.3 Problem Set 18

Algebra Homework, Edition 18

Nineteenth Week

Roots of polynomials over fields

Problem 87.
Show that any finite field has an extension of degree 2.

Problem 88.
Let $\mathbf{F}$ be a field and let n be a positive integer such that $\mathbf{F}$ has no nontrivial field extensions of degree less than n. Let $\mathbf{L} = \mathbf{F}[u]$ be an extension field with $u^n \in F$. Prove that every element in L is a product of elements of the form $au + b$ where $a, b \in F$.

Problem 89.
Show that $\big[\mathbb{Q}[\sqrt[5]{2} + \sqrt{5}] : \mathbb{Q}\big] = 10$.

Problem 90.
Let $f(x) \in \mathbb{Q}[x]$ be an irreducible polynomial of odd degree and suppose that u is a root of $f(x)$. Prove $\mathbb{Q}[u] = \mathbb{Q}[u^{2^k}]$ for every natural number k.

Problem 91.
Let the field $\mathbf{L}$ extend the field $\mathbf{F}$ so that $[\mathbf{L} : \mathbf{F}] = 4$.

(a) Prove that there are no more than 5 fields $\mathbf{K}$ with $\mathbf{F} \subseteq K \subseteq \mathbf{L}$. Given an example where there are 5 such intermediate fields.

(b) Give an example where the only intermediate fields are $\mathbf{F}$ and $\mathbf{L}$.

Problem 92.
Prove that every algebraically closed field is infinite.

LECTURE 21

Constructions by Straightedge and Compass

Euclid's book, sadly now fallen from the mathematician's bookshelf, should properly still be the property of every mathematician. It is filled with theorems proved by means of constructions using straightedge and compass. Loosely speaking, these constructions started with a given finite configuration of points on the plane and then proceeded in a step-by-step fashion to construct further points. The new points could only arise in one of three ways:

(a) As the point of intersection of two line segments, each drawn with the help of the straightedge through two distinct points already at hand. Here the endpoints of the segments, and indeed almost all the points on the segments, need not be among the constructed points.

(b) As points of intersection between a line segment and a circle, where the line segment arises as above and the circle is drawn with the help of the compass by placing the foot and the drawing points of the compass on points constructed at some prior step. Again the only points on the line segment and the circle that qualify as constructed are the points of intersection.

(c) As points of intersection between two circles, each circle drawn as described above.

Among the problems left as unsolved by the geometers of this classical period were the following:

The Trisection of Angles. *Given an arbitrary angle, to trisect it by means of straightedge and compass.*

The Duplication of the Cube or the Delian Problem. *Given an arbitrary cube, to construct a cube of twice the volume by means of straightedge and compass.*

A legend behind this problem concerns a serious plague. The advice of the great oracle of Apollo at Delphi was sought. The Altar of Apollo was an impressive cube. The oracle advised that Apollo would intercede once the altar had been exactly doubled in volume. Apollo never interceded.

Squaring the Circle. *Given a circle, to construct, by means of straightedge and compass, a square with the same area as the circle.*

The Construction of Regular Polygons. *Given a line segment to construct, by means of straightedge and compass, a regular polygon of n sides each of length the same as the given line segment.*

We are now in a position to present the solutions to most of these problems.

If we identify Euclid's plane with $\mathbb{R}\times\mathbb{R}$, we can convert these geometric problems into algebraic problems. To set a unit length, we start our analysis with two points $(0,0)$ and $(1,0)$. Let $\mathcal{C}$ be the totality of all points that can be constructed by straightedge and compass from these first two points. Since at any stage only finitely many new points are constructed and since any constructible point is reached after some finite sequence of steps, we see that $\mathcal{C}$ is countable. We say a real number r is **constructible** provided r is one of the coordinates of a point that belongs to $\mathcal{C}$. It is an informative exercise in straightedge and compass construction to show that r is constructible if and only if $|r|$ is the length of a line segment joining to constructible points (including degenerate segments of length 0). Let E be the set of constructible real numbers.

[The leading sentence of the last paragraph might well have given you pause. Is it really permissible to identify Euclid's plane with $\mathbb{R}\times\mathbb{R}$? What would you have to do to prove this statement?]

It is clear that $0,1\in E$. We will show next E is closed under addition and multiplication, and that every nonzero constructible real has a constructible multiplicative inverse. In this way, we will arrive at $\mathbf{E}$, the field of constructible reals. Actually, $\mathbf{E}$ has very special properties. Let us say that a subfield $\mathbf{K}$ of $\mathbb{R}$ is **closed under the extraction of square roots** provided a real number r belongs to K whenever $r^2\in K$. This property holds for $\mathbb{R}$, essentially by default, but not for $\mathbb{Q}$.

Let $\mathbf{F}$ be a field. By a **square root tower** over $\mathbf{F}$ we mean a finite sequence

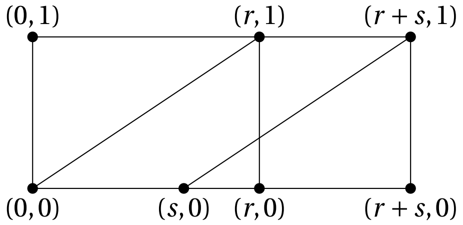
$$\mathbf{F}=\mathbf{F}_0\leq\mathbf{F}_1\leq\cdots\leq\mathbf{F}_n$$

of field extensions such that for each $j<n$ there is some u_j so that $u_j^2\in F_j$ and $\mathbf{F}_j[u_j]=\mathbf{F}_{j+1}$. That is, we obtain the next field up the tower $\mathbf{F}_{j+1}$ by adjoining to $\mathbf{F}_j$ a square root of an element belonging to F_j. Let $\mathbf{K}$ be a field extending $\mathbf{F}$ and let $r\in K$. We say r is **captured in a square root tower** over $\mathbf{F}$ provided $r\in F_n$ for some square root tower $\mathbf{F}=\mathbf{F}_0\leq\cdots\leq\mathbf{F}_n$.

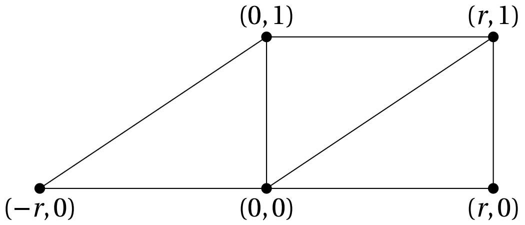
The Basic Theorem for the Field of Constructible Reals. *The constructible real numbers constitute the smallest subfield* $\mathbf{E}$ *of* $\mathbb{R}$ *that is closed under the extraction of square roots. Moreover,* $r\in E$ *if and only if* r *captured in a square root tower over* $\mathbb{Q}$*. In particular,* $\mathbf{E}$ *is algebraic over* $\mathbb{Q}$ *and* $[\mathbb{Q}[r]:\mathbb{Q}]$ *is a power of* 2*, for all* $r\in E$*.*

Proof. That E is closed under addition, multiplication, inversion of nonzero elements, and square roots of positive elements will follow from a series of diagrams. The intention of the diagrams is to display how the construction by straightedge and compass should proceed. Some familiarity with the use of straightedge and compass is needed to interpret the diagrams. For instance,

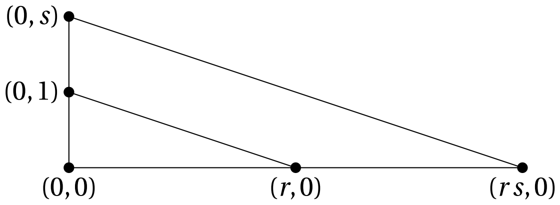
- given a line segment (i.e. its endpoints) and a point on the line segment, there is a straightedge and compass construction of a second line segment perpendicular to the first at the given point;
- given a line segment ℓ and a point $\mathbf{p}$ not collinear with ℓ, there is a straightedge and compass construction of a point $\mathbf{q}$ collinear with ℓ so that the segment joining $\mathbf{p}$ and $\mathbf{q}$ is perpendicular to ℓ (extended as required to include the point $\mathbf{q}$);
- given a line segment ℓ and a point $\mathbf{p}$ not collinear with ℓ, there is a straightedge and compass construction of a line segment through $\mathbf{p}$ that is parallel to ℓ.



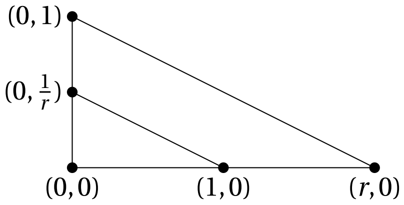
Construction of $r+s$ when $0 \leq s \leq r$



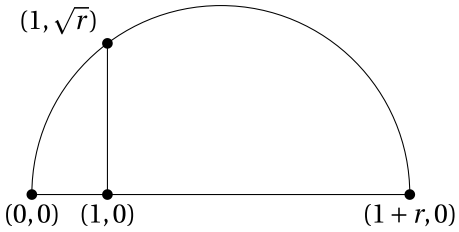
Construction of $-r$ when $0 < r$



Construction of rs when $1 \leq s$ and $0 \leq r$



Construction of $\frac{1}{r}$ when $0 < r$



Construction of $\sqrt{r}$ when $0 < r$

At this point we see that **E** is indeed a subfield of the field $\mathbb{R}$ of real numbers and it is closed under the extraction of square roots. Of course, it is also an extension of the field $\mathbb{Q}$ of rationals.

Now suppose that $r \in E$. Pick $s \in E$ so that $(r,s) \in \mathcal{C}$. There is a finite sequence

$$\mathbf{p}_0, \ldots, \mathbf{p}_n = (r,s)$$

of points so that $\mathbf{p}_0$ is constructible from the points $(0,0)$ and $(1,0)$ in one step by straightedge and compass, and for each $k < n$ the point $\mathbf{p}_{k+1}$ is constructible in one step from points in $\{(0,0),(1,0),\mathbf{p}_0,\ldots,\mathbf{p}_k\}$. For each $j < n$ put $\mathbf{p}_j = (r_j, s_j)$. Let $\mathbf{K} = \mathbb{Q}[r_0, s_0, r_1, s_1, \ldots, r_{n-1}, s_{n-1}]$.

Contention. $r \in K[u]$ for some $u \in \mathbb{R}$ so that u is positive and $u^2 \in K$.

There are three kinds of straightedge and compass steps.

The first produces a point of intersection of two lines. Each of the lines is determined by two distinct points. Construing this algebraically, we have a system of two linear equations in two unknowns, which are the coordinates of the point of intersection. We can solve this system just using the field operations. This yields that $r \in K$ and so $r \in K[1]$.

The second produces points of intersection of a line and a circle. The line is determined by two distinct points (a,b) and (c,d), and the circle is determined by its center (a',b') and one point (c',d') on the circle. Construing this algebraically we arrive at a system of two equations:

$$(c-a)(y-b) = (d-b)(x-a)$$
$$(x-a')^2+(y-b')^2 = (c'-a')^2+(d'-b')^2$$

The point (r,s) is a root of this system. Using just the field operations solve the first equation for one of the unknowns in terms of the other (taking care not to divide by 0). Substituting the result into the second equation yields a quadratic equation with coefficients in K. Invoking the quadratic formula will produce values for the unknown. The formula involves the extraction of a square root of a nonnegative number $u \in K$. The value of the other unknown can be determined just using the field operations. So $r \in K[u]$ where $u^2 \in K$ in this case.

The last kind of straightedge and compass step produces the points of intersection of two circles, each determined by its center and a point of the circle. Algebraically, this yields the system

$$(x-a)^2+(y-b)^2 = (c-a)^2+(d-b)^2$$
$$(x-a')^2+(y-b')^2 = (c'-a')^2+(d'-b')^2$$

Subtracting these equations eliminates the x^2's and y^2's. The resulting equation is of the form $Ax+By=C$. This equation can be solved for one of the unknowns in terms of the other and the result substituted into the first displayed equation. From this point the argument proceeds as before.

In this way the contention is established.

Evidently, square root towers can be extended by square root towers to obtain longer square root towers. So we see, inductively, that $r_0, s_0, r_1, s_1, \ldots, r_n, s_n$ are all contained in a single square root tower over $\mathbb{Q}$.

So every element of E is captured in some square root tower over $\mathbb{Q}$.

Conversely, since square roots of nonnegative constructible numbers are themselves constructible, we see, via induction, that any real number captured in a square root tower over $\mathbb{Q}$ must be constructible.

Now observe that if **K** is any subfield of $\mathbb{R}$ that is closed under the extraction of square roots, then every square root tower over $\mathbb{Q}$ is included in **K**. Therefore, **E** is a subfield of **K**. Since **E** is itself closed under the extraction of square roots, we see that indeed it must be the smallest subfield of $\mathbb{R}$ that is closed under the extraction of square roots.

Finally, let $r \in E$ and let $\mathbb{Q} = \mathbf{F}_0 \le \mathbf{F}_1 \le \cdots \le \mathbf{F}_n$ be a square root tower that captures r. We know that $1 \le [\mathbf{F}_{k+1} : \mathbf{F}_k] \le 2$. It follows from the Dimension Formula that $[\mathbf{F}_n : \mathbb{Q}]$ must be a power of 2. But $\left[\mathbf{F}_n : \mathbb{Q}\right] = \left[\mathbf{F}_n : \mathbb{Q}[r]\right]\left[\mathbb{Q}[r] : \mathbb{Q}\right]$. So we conclude that $[\mathbb{Q}[r] : \mathbb{Q}]$ must also be a power of 2, and also that r is algebraic of $\mathbb{Q}$. □

While the Basic Theorem for the Field of Constructible Reals is probably close to the spirit of Euclid, it is convenient to recast this result over the field of complex numbers, which can itself be construed as the Euclidean plane. It helps to recall that a complex number can be written as

$$z = re^{i\theta} = r(\cos\theta + i\sin\theta)$$

where r is a nonnegative real number and $0 \le \theta < 2\pi$. A bit of trigonometry yields

$$\sqrt{z} = \sqrt{r}(\cos\frac{\theta}{2} + i\sin\frac{\theta}{2})$$

So every element of the field of complex numbers has a square root back in the field (for the field of real numbers, the corresponding statement was only true for nonnegative reals). Now we already saw how to construct $\sqrt{r}$ with straightedge and compass. As Euclid taught us how to bisect angles with straightedge and compass. What more would the willing graduate student need to convert the proof above into a proof of the theorem below?

The Basic Theorem for the Field of Constructible Complex Numbers. *The constructible complex numbers constitute the smallest subfield* $\mathbf{E}_c$ *of* $\mathbb{C}$ *that is closed under the extraction of square roots. Moreover,* $r \in E_c$ *if and only if* r *captured in a square root tower over* $\mathbb{Q}$*. In particular,* $\mathbf{E}_c$ *is algebraic over* $\mathbb{Q}$ *and* $[\mathbb{Q}[r]:\mathbb{Q}]$ *is a power of* 2*, for all* $r \in E$*.*

Using these basic theorems about constructible numbers, we are in position to tackle problems that were inaccessible to the ancient geometers.

The Impossibility of General Angle Trisection. *The angle* $\frac{\pi}{3}$ *radians (whose construction was given in Euclid's First Proposition), cannot be trisected with straightedge and compass.*

Proof. This is the angle in an equilateral triangle. Its trisection results in an angle of $\frac{\pi}{9}$ radians. The construction of such an angle entails the constructibility of a right triangle with hypotenuse 1 and legs of length $\cos(\frac{\pi}{9})$ and $\sin(\frac{\pi}{9})$. In turn, this entails that $\cos(\frac{\pi}{9})$ would be a constructible real number. We will see that this is not the case.

The hard-working graduate student can verfiy the following trigonometric identity

$$\cos 3\alpha = 4\cos^3\alpha - 3\cos\alpha.$$

Since $\cos 3\frac{\pi}{9} = \frac{1}{2}$, we see that $\cos\frac{\pi}{9}$ is a root of $4x^3 - 3x - \frac{1}{2}$. This polynomial is irreducible over $\mathbb{Q}$ (the verification of this is left to the enjoyment of the graduate students). This means, according to Kronecker, that $[\mathbb{Q}[\frac{\pi}{9}]:\mathbb{Q}] = 3$. Since 3 is not a power of 2, we see that $\cos\frac{\pi}{9}$ is not constructible. □

The Impossibility of Duplicating the Unit Cube. *A line segment the cube of whose length is* 2 *cannot be constructed by straightedge and compass.*

Proof. Evidently, $\sqrt[3]{2}$ is a root of $x^3 - 2$. According to Eisenstein, this polynomial is irreducible over $\mathbb{Q}$ and according to Kronecker $[\mathbb{Q}[\sqrt[3]{2}]:\mathbb{Q}] = 3$. Since 3 is not a power of 2 we are done. □

The Impossibility of Squaring the Unit Circle. *A line segment the square of whose length is the area of the unit circle cannot be constructed by straightedge and compass.*

We are not yet in a position to prove this theorem. The area of the unit circle is π. A square of this area would have sides of length $\sqrt{\pi}$. The constructibility of this number would entail the constructibility of π. But it turns out that π is not even algebraic over $\mathbb{Q}$. We will prove this later, finally putting this old problem to rest.

This leaves the problem of constructing regular polygons. Here is what is known.

A prime number p is said to be a **Fermat prime** provided $p = 2^a + 1$ for some positive natural number number a. Here are the first five Fermat primes: $3, 5, 17, 257$, and 65537. These five were already known to Fermat—no further Fermat primes have been found in the ensuing years, even with the very considerable computational power now at our disposal. It was conjectured by Eisenstein that there are infinitely many Fermat primes. It is even conceivable that Eisenstein was wrong and that those we now know are all that there are.

Gauss's Theorem on Constructible Regular Polygons. *Let* $n \geq 3$*. It is possible to construct by straightedge and compass a regular* n*-gon if and only if* n *has the form*

$$n = 2^e p_1 p_2 \dots p_m$$

where e *is a natural number and* $p_1, \dots, p_m$ *are distinct Fermat primes.*

We also have to defer the proof of this theorem. Essentially, we can identify the vertices of a regular n-gon as the complex numbers that are the roots of $x^n - 1$. We will take up a detailed study of these roots of unity later. At that time we can provide a proof of this theorem of Gauss. Even then, our grasp of the situation is incomplete since our knowledge of Fermat primes is so sketchy.

LECTURE 22

GALOIS CONNECTIONS

In his investigation of the solvability of polynomial equations by radicals, Galois came across a way to connect (the splitting field of) a polynomial with a finite combinatorial object (in fact a finite group) which proved more amenable to analysis than the splitting field, which was an infinite object. It turns out that the connection Galois discovered is a particular instance of what has turned out to be a common phenomena. Because more general situation is not encumbered with all the details of Galois's particular connection, and because the idea requires hardly any mathematical background, I will present the general situation first.

22.1 ABSTRACT GALOIS CONNECTIONS

Consider any two classes A and B and any two-place relation $R \subseteq A \times B$. Two-place relations are ubiquitous in mathematics. Here are some examples:

- Take $A = \mathbb{Z} = B$ and let R be the divisibility relation.
- Take A and B each to be the set of rational numbers in the unit interval and let R be their usual ordering $\leq$.
- Let A and B both be the set of vertices of some graph and let R be the relation of adjacency.
- Let A and B both be the class of all groups and let R be the relation of one group being a homomorphic image of another.
- Let A be the ring of polynomials in 5 variables over the complex numbers, let B be the vector space of 5-tuples of complex numbers. Take R to be the relation between a polynomial and its solutions.
- Let A and B be the class of all sets and take R to be the membership relation.
- Let A be the points in the Euclidean plane and let B be the collection of 2-element subsets of A. Let R relate the point p to $\{a, b\}$ provided p is on the line segment joining a and b.
- Imagine two or three more examples.

We call a system (A, B, R), where $R \subseteq A \times B$, a **Galois connection**. Each Galois connection induces two maps ${}^{\rightarrow}$ and ${}^{\leftarrow}$, called the **polarities** of the Galois connection.

$$^{\rightarrow} : \mathcal{P}(A) \to \mathcal{P}(B)$$

where for each $X \subseteq A$

$$X^{\rightarrow} := \{y \mid y \in B \text{ and } (x, y) \in R \text{ for all } x \in X\}$$

and

$$^{\leftarrow} : \mathcal{P}(B) \to \mathcal{P}(A)$$

where for each $Y \subseteq B$

$$Y^{\leftarrow} := \{x \mid x \in A \text{ and } (x, y) \in R \text{ for all } y \in Y\}$$

I read $X^{\rightarrow}$ as "X going over" and $Y^{\leftarrow}$ as "Y coming back."

An example is in order. Let $A = \mathbb{Z} = B$ and let R be the divisibility relation. So

$$\{6, 9\}^{\rightarrow} = \{r \mid r \in \mathbb{Z} \text{ and } 6 \mid r \text{ and } 9 \mid r\}$$

That is, $\{6, 9\}^{\rightarrow}$ is just all the multiples of 18, since 18 is the least common multiple of 6 and 9. So even though we started with a finite set, by going over we got an infinite set—but it is a nice one, an ideal of the ring of integers. Now let's see what we get by coming back.

$$\{6, 9\}^{\rightarrow\leftarrow} = \{s \mid s \in \mathbb{Z} \text{ and } s \mid r \text{ for all } r \in \{6, 9\}^{\rightarrow}\}$$

With a little thought the industrious graduate students will find that

$$\{6, 9\}^{\rightarrow\leftarrow} = \{1, -1, 2, -2, 3, -3, 6, -6, 9, -9, 18, -18\}.$$

This is just the set of divisors of 18. So we started with a set with two elements and then by going over and coming back we arrive at a set with 12 elements that includes our original set. Now observe

$$\varnothing^{\rightarrow} = \mathbb{Z} \text{ and } \varnothing^{\leftarrow} = \mathbb{Z}.$$

Also observe

$$\mathbb{Z}^{\rightarrow} = \{0\} \text{ and } \mathbb{Z}^{\leftarrow} = \{1, -1\}.$$

Now $\mathcal{P}(A)$ and $\mathcal{P}(B)$ are partially ordered by the inclusion relation $\subseteq$. The basic properties of Galois connections concern how the polarities and this order relation interact.

The Polarity Theorem for Galois Connections. *Let (A, B, R) be any Galois connection. All of the following hold.*

(a) *If $X_0 \subseteq X_1 \subseteq A$, then $X_1^{\rightarrow} \subseteq X_0^{\rightarrow}$. If $Y_0 \subseteq Y_1 \subseteq B$, then $Y_1^{\leftarrow} \subseteq Y_0^{\leftarrow}$.*

(b) *If $X \subseteq A$, then $X \subseteq X^{\rightarrow\leftarrow}$. If $Y \subseteq B$, then $Y \subseteq Y^{\leftarrow\rightarrow}$.*

(c) *If $X \subseteq A$, then $X^{\rightarrow} = X^{\rightarrow\leftarrow\rightarrow}$. If $Y \subseteq B$, then $Y^{\leftarrow} = Y^{\leftarrow\rightarrow\leftarrow}$.*

(d) *If $X_0, X_1 \subseteq A$ and $X_0^{\rightarrow\leftarrow} = X_1^{\rightarrow\leftarrow}$, then $X_0^{\rightarrow} = X_1^{\rightarrow}$.*
If $Y_0, Y_1 \subseteq B$ and $Y_0^{\leftarrow\rightarrow} = Y_1^{\leftarrow\rightarrow}$, then $Y_0^{\leftarrow} = Y_1^{\leftarrow}$.

(e) *For all $X \subseteq A$ and all $Y \subseteq B$,*

$$X \subseteq Y^{\leftarrow} \textit{ if and only if } Y \subseteq X^{\rightarrow}.$$

The proof of this theorem is left in the trustworthy hands of the graduate students. It is even part of an official problem set. Can you deduce the other parts of this theorem from (a) and (e)?

We say that subsets of A of the form $Y^{\leftarrow}$ are **closed**, as are the subsets of B of the form $X^{\rightarrow}$. Part of the content of this theorem about polarities is that restricted to the closed sets on each side of the Galois connection, the polarities are inverses of each other and they are order reversing. So viewed as ordered sets, the systems of closed sets on each side are *anti-isomorphic*: one looks like the upside down version of the other. The polarities are the anti-isomorphisms.

The intersection of any nonempty collection of closed sets from one side of a Galois connection will be again a closed set. There is a unique smallest closed set on each side. The least closed subset of A is, of course, $B^{\leftarrow}$. This means that any collection of closed subsets from one side of a Galois connection always has a greatest lower bound. It follows via the polarities (which are anit-isomorphisms, sometimes called dual isomorphisms) that every collection of closed sets from one side of a Galois connection always has a least upper bound. A partially ordered set with these properties is called a **complete lattice**. So the closed sets from any one side of a Galois connection always constitute a complete lattice. In the example we worked with, the integers with divisibility, the closed sets on the right side of the Galois connection turn out to be the ideals of the ring of integers.

22.2 The Connection of Galois

The Galois connection discovered by Evariste Galois was not listed among our examples in the section above. We describe it in this section.

Let $\mathbf{E}$ be a field that extends a field $\mathbf{F}$. The set E will be the left side of Galois' connection. Let

$$\operatorname{Gal}\mathbf{E}/\mathbf{F} = \{\sigma \mid \sigma \text{ is an automorphism of } \mathbf{E} \text{ and } \sigma(a) = a \text{ for all } a \in F\}.$$

$\operatorname{Gal}\mathbf{E}/\mathbf{F}$ is the group of automorphisms of $\mathbf{E}$ that fix each element of $\mathbf{F}$. It is called the **Galois group** of $\mathbf{E}$ over $\mathbf{F}$. It is the right side of Galois' connection. The relation that connects these two sides is

$$\{(a,\sigma) \mid a \in E \text{ and } \sigma \in \operatorname{Gal}\mathbf{E}/\mathbf{F} \text{ and } \sigma(a) = a\}.$$

The polarities of Galois' connection are given as follows, for any $X \subseteq E$ and any $Y \subseteq \operatorname{Gal}\mathbf{E}/\mathbf{F}$:

$$\begin{aligned} X^{\rightarrow} &= \operatorname{Gal} X = \{\sigma \mid \sigma \in \operatorname{Gal}\mathbf{E}/\mathbf{F} \text{ and } \sigma(x) = x \text{ for all } x] \in X\} \\ Y^{\leftarrow} &= \operatorname{Inv} Y = \{a \mid a \in E \text{ and } \sigma(a) = a \text{ for all } \sigma \in Y\} \end{aligned}$$

We abandon the arrow notation in favor of Gal and Inv. We leave it in the trustworthy hands of the graduate students to work out that $\operatorname{Gal} X$ is always a subgroup of $\operatorname{Gal}\mathbf{E}/\mathbf{F}$ and that $\operatorname{Inv} Y$ is always a subfield of $\mathbf{E}$ that extends $\mathbf{F}$. $\operatorname{Inv} Y$ is called the **fixed field** of Y and $\operatorname{Gal} X$ is called the Galois group of X. In Galois' investigations, the field $\mathbf{E}$ was the splitting field of some polynomial $f(x)$ with coefficients from $\mathbf{F}$. Galois realized that an automorphism of $\mathbf{E}$ that leaves the coefficients of the polynomial fixed must send roots of $f(x)$ to roots of $f(x)$. And as the roots of $f(x)$ determine the elements of the splitting field $\mathbf{E}$, this meant that $\operatorname{Gal}\mathbf{E}/\mathbf{F}$ was, in essence, just some group consisting of certain permutations of the roots of $f(x)$. Such a group is finite since $f(x)$ can have only finitely many roots. In this way, Galois saw that it might be possible to understand the roots of $f(x)$ by understanding this finite group instead of trying to understand how the roots were situated in the (usually) infinite field $\mathbf{E}$. Galois succeeded.

The next two lectures are devoted to understanding the closed sets on each side of Galois' connection.

22.3 Problem Set 19

Algebra Homework, Edition 19
Twentieth Week
Galois Connections

In Problem 93 to Problem 97 below, let A and B be two classes and let R be a binary relation with $R \subseteq A \times B$. For $X \subseteq A$ and $Y \subseteq B$ put

$$X^{\rightarrow} = \{b \mid x \mathrel{R} b \text{ for all } x \in X\}$$
$$Y^{\leftarrow} = \{a \mid a \mathrel{R} y \text{ for all } y \in Y\}$$

Problem 93.
Prove that if $W \subseteq X \subseteq A$, then $X^{\rightarrow} \subseteq W^{\rightarrow}$. (Likewise if $V \subseteq Y \subseteq B$, then $Y^{\leftarrow} \subseteq V^{\leftarrow}$.)

Problem 94.
Prove that if $X \subseteq A$, then $X \subseteq X^{\rightarrow\leftarrow}$. (Likewise if $Y \subseteq B$, then $Y \subseteq Y^{\leftarrow\rightarrow}$.)

Problem 95.
Prove that $X^{\rightarrow\leftarrow\rightarrow} = X^{\rightarrow}$ for all $X \subseteq A$ (and likewise $Y^{\leftarrow\rightarrow\leftarrow} = Y^{\leftarrow}$ for all $Y \subseteq B$).

Problem 96.
Prove that the collection of subclasses of A of the form $Y^{\leftarrow}$ is closed under the formation of arbitrary intersections. (As is the collection of subclasses of B of the form $X^{\rightarrow}$.) We call classes of the form $Y^{\leftarrow}$ and the form $X^{\rightarrow}$ closed.

Problem 97.
Let $A = B = \{q \mid 0 < q < 1 \text{ and } q \text{ is rational}\}$. Let R be the usual ordering on this set. Identify the system of closed sets. How are they ordered with respect to inclusion? Linearly? Densely?

LECTURE 23

THE FIELD SIDE OF GALOIS' CONNECTION

Suppose that $\mathbf{E}$ is the splitting field of a polynomial $f(x)$ over the field $\mathbf{F}$ and that $\mathbf{K}$ is a field intermediate between $\mathbf{F}$ and $\mathbf{E}$. From general facts about Galois connections, we know that $\operatorname{Inv}\operatorname{Gal}(\mathbf{E}/\mathbf{K})$ is a subfield of $\mathbf{E}$ that extends $\mathbf{K}$. Our hope is that $\mathbf{K} = \operatorname{Inv}\operatorname{Gal}(\mathbf{E}/\mathbf{K})$. While this hope cannot be realized in general, there is an important case in which it does hold.

Let us say that an irreducible polynomial $p(x) \in \mathbf{F}[x]$ is *separable* provided the number of distinct roots it has in its splitting field over $\mathbf{F}$ is the same as its degree. This means that when $p(x)$ is completely factored over its splitting field, then all the factors are distinct, that is all the roots are distinct or separated. We say a polynomial $f(x) \in \mathbf{F}[x]$ is *separable* provided each of its irreducible factors is separable.

Notice that a separable polynomial of degree n may have fewer than n distinct roots in its splitting field. For example $x^2+2x+1 = (x+1)^2$ has degree 2 but it has only one root (namely -1). In general, a polynomial $f(x) \in \mathbf{F}[x]$ factors over $\mathbf{F}$ as

$$f(x) = g_0(x)^{e_0} g_1(x)^{e_1} \dots g_{m-1}(x)^{e_{m-1}}$$

where each $g_k(x)$ is irreducible and disinct from the other g's and each e_k is a positive integer. This polynomial must have repeated roots in its splitting field as long as some $e_k > 1$. But consider the polynomial

$$h(x) = g_0(x) g_1(x) \dots g_{m-1}(x).$$

The polynomials $f(x)$ and $h(x)$ have the same splitting field over $\mathbf{F}$, and $h(x)$ will have distinct roots, provided $f(x)$ (and hence $h(x)$) is separable.

The Galois Field Closure Theorem. *Let* $\mathbf{E}$ *be the splitting field of a separable polynomial over the field* $\mathbf{F}$*. Then* $\operatorname{Inv}\operatorname{Gal}(\mathbf{E}/\mathbf{K}) = \mathbf{K}$*, for every field* $\mathbf{K}$ *intermediate between* $\mathbf{F}$ *and* $\mathbf{E}$*.*

Proof. Let $f(x) \in \mathbf{F}[x]$ be a separable polynomial from $\mathbf{F}[x]$ and let $\mathbf{E}$ be the splitting field of $f(x)$ over $\mathbf{F}$. Let $\mathbf{K}$ be a field intermediate between $\mathbf{F}$ and $\mathbf{E}$ and put $\mathbf{L} = \operatorname{Inv}\operatorname{Gal}(\mathbf{E}/\mathbf{K})$. For the general properties of polarities for Galois connections, we see that $K \subseteq L$ and that $\operatorname{Gal}(\mathbf{E}/\mathbf{K}) = \operatorname{Gal}(\mathbf{E}/\mathbf{L})$. But $\mathbf{E}$ is the splitting field of $f(x)$ over both $\mathbf{K}$ and $\mathbf{L}$. By the Existence and Uniqueness Theorem for Splitting Fields, we see that $[\mathbf{E}:\mathbf{K}] = |\operatorname{Gal}(\mathbf{E}/\mathbf{K})| = |\operatorname{Gal}(\mathbf{E}/\mathbf{L})| = [\mathbf{E}:\mathbf{L}]$. But we know that $[\mathbf{E}:\mathbf{K}] = [\mathbf{E}:\mathbf{L}][\mathbf{L}:\mathbf{K}]$. It follows that $[\mathbf{L}:\mathbf{K}] = 1$. Hence, $\mathbf{K} = \mathbf{L}$, as desired. □

23.1 Perfect Fields

This leaves us with the question of when an irreducible polynomial is separable. It turns out that just a bit of formal calculus does the trick. Consider a polynomial

$$f(x) = a_0 + a_1x + a_2x^2 + \cdots + a_nx^n.$$

We can define the derivative of $f'(x)$ as follows

$$f'(x) = a_1 + 2a_2x + 3a_3x^2 + \cdots + na_nx^{n-1}.$$

But we have to be careful. The exponents are natural numbers, but our field $\mathbf{F}$, might not have in it any natural numbers. So, to be necessarily more fussy, we define

$$f'(x) = a_1 + ((1+1)a_2)x + ((1+1+1)a_3)x^2 + \cdots + ((\underbrace{1+\cdots+1}_{n\text{-times}})a_n)x^{n-1}.$$

But after this we will write it as we did at first. Notice that this definition always produces another polynomial, regardless of the field over which we are working. No limits or other notion of convergence enters here.

It is left to the eager graduate students to verify that the derivatives of sums and products (and even compositions) of polynomials work out just like they do in calculus.

Fact. Let $f(x)$ be an irreducible polynomial with coefficients in the field $\mathbf{F}$. Then $f(x)$ is separable if and only if $f(x)$ and $f'(x)$ are relatively prime.

Proof. It is harmless to suppose that $f(x)$ is monic. Let $\mathbf{E}$ be the splitting field of $f(x)$ over $\mathbf{F}$. Let $r_0, \ldots, r_{m-1}$ be the distinct roots of $f(x)$ in E. Then we see that

$$f(x) = (x - r_0)^{e_0}(x - r_1)^{e_1} \ldots (x - r_{m-1})^{e_{m-1}},$$

for certain positive integers $e_0, \ldots, e_{m-1}$.

Suppose first that $f(x)$ is separable. This only means that $e_0 = \cdots = e_{m-1} = 1$. Under this supposition, $f'(x)$ is, according to the product rule, just the sum of all terms made by deleting single factors from the factorization above. This entails (with the graduate students fiddling down the details) that none of the irreducible factors (over $\mathbf{E}$) of $f(x)$ can divide $f'(x)$. This means that $f(x)$ and $f'(x)$ are relatively prime over $\mathbf{E}$. Therefore, they must be relatively prime over $\mathbf{F}$.

Now suppose that $f(x)$ is not separable. This means that some $e_k > 0$. Hence, $f(x) = (x - r)^2 g(x)$ is a factorization over $\mathbf{E}$ for some $r \in E$ and some $g(x) \in \mathbf{E}[x]$. The product rule now tells us that $f'(x) = 2(x - r)g(x) + (x - r)^2 g'(x)$. This means that $x - r$ is a common divisor of $f(x)$ and $f'(x)$. So $f(x)$ and $f'(x)$ are not relatively prime over $\mathbf{E}$. Hence (why?) they are not relatively prime over $\mathbf{F}$. □

Actually, the proof above does not make significant use of the irreducibility of $f(x)$. The graduate students should be able to reformulate the statement of this fact so as to remove the irreducibility condition.

Of course, from our perspective the best thing that can happen is for all polynomials of positive degree to turn out to be separable. A field $\mathbf{F}$ with this property is called *perfect*.

Here is an important corollary of the Fact above.

Corollary 23.1.1. *Every field of characteristic* 0 *is perfect.*

Proof. We only need to pay attention to irreducible polynomials. Observe that in a field of characteristic 0, if $f(x)$ has positive degree, then $f'(x)$ cannot be the zero polynomial and must have degree properly smaller than the degree of $f(x)$. (Over fields of prime characteristic it is possible for $f'(x)$ to be the zero polynomial.) Since we are taking $f(x)$ to be irreducible, we see that $f(x)$ and $f'(x)$ must be relatively prime, since $f(x)$ cannot divide $f'(x)$, the degree of $f'(x)$ being too small. □

So what happens for fields of prime characteristic?

Suppose that **F** is a field of characteristic p, where p is a prime number. There is an interesting thing that happens. According to the Binomial Theorem (that holds in every commutative ring) in $\mathbf{F}[x]$ we have

$$(a+b)^p = \sum_{k\leq p} \binom{p}{k} a^k b^{p-k}$$

for all $a, b \in F$. Recall that $\binom{p}{0} = 1 = \binom{p}{p}$ but that $p \mid \binom{p}{k}$ when $0 < k < p$. Recalling the fussy point made above about positive integer multiples, we see that $\binom{p}{k}$ reduces to 0 in the characteristic p case, whenever $0 < k < p$. This means that

$$(a+b)^p = a^p + b^p$$

for all $a, b \in F$. But we also know that

$$(ab)^p = a^p b^p$$

for all $a, b \in F$. This means that the map $a \mapsto a^p$ for all $a \in F$, must be a homomorphism from **F** into **F**. Now fields are simple, that is they have just two ideals. So the kernel of this special map must either be $\{0\}$ (in which case the map is one-to-one) or F itself (in which case the map sends every element of F to 0). Since $1^p = 1 \neq 0$, we see that our map is actually one-to-one, that is it is an embedding of **F** into **F**. This map is known as the *Frobenius embedding.*

Theorem on Perfect Fields of Characteristic p. *Let* **F** *be a field of prime characteristic p. The field* **F** *is perfect if and only if every element of F has a p^{th} root in F.*

Proof. Suppose first that there is some $a \in F$ so that a has no p^{th} root in F. We contend that the polynomial $x^p - a$ is irreducible. Suppose otherwise. So $x^p - a = g(x)h(x)$ where $g(x)$ is a monic polynomial of positive degree $k < p$. Let **E** be the splitting field of $x^p - a$ over **F** and let $b \in E$ be a root of $g(x)$. Now notice $b^p = a$ so $b \notin F$ and

$$x^p - a = x^p - b^p = (x-b)^p$$

in $\mathbf{E}[x]$. By unique factorization, $g(x) = (x-b)^k$. As b^k is the constant term of $g(x)$, we find that $b^k \in F$. But k and p are relatively prime integers (since p is prime and $0 < k < p$). Pick integers u and v so that $1 = uk + vp$. But then

$$b = b^{uk+vp} = (b^k)^u (b^p)^v = (b^k)^u a^v \in F.$$

This provides a contradiction to our supposition. So $x^p - a$ is irreducible. Its derivative is the zero polynomial. So we see that it is not separable. This means that **F** is not perfect.

For the converse, suppose that every element of F has a p^{th} root in F. Consider any irreducible polynomial $f(x)$. The only barrier to $f(x)$ being separable is that $f'(x)$ might be the zero polynomial. This can only happen when

$$f(x) = a_0 + a_p x^p + a_{2p} x^{2p} + \cdots + a_{np} x^{np}.$$

Since every element of F is a p^{th} power we can pick $b_0, \ldots, b_{np} \in F$ so that $a_{kp} = (b_{kp})^p$ for all $k \leq n$. This gives us

$$f(x) = b_0^p + b_p^p x^p + \cdots + b_{np}^p x^{np} = (b_0 + b_p x + b_{2p} x^2 + \cdots + b_{np} x^n)^p.$$

But $f(x)$ is irreducible. So it cannot happen that $f'(x)$ is the zero polynomial. Consequently, $f(x)$ is separable and **F** must be a perfect field. □

A nice corollary of this theorem is

Corollary 23.1.2. *Every finite field is perfect.*

The reason is that the Frobenius map from a finite field to itself must be onto the field since it is one-to-one (and any one-to-one map of a finite set to itself must be onto). So we see that every element of a finite field is a p^{th} power of some other element, where p is the characteristic.

23.2 Galois Extensions

The key hypothesis of the Galois Field Closure Theorem is that **E** should be the splitting field of a separable polynomial over **F**. In this case, we say that **E** is a **Galois extension** of **F**. There are several useful ways to characterize this kind of extension.

Theorem Characterizing Galois Extensions. *Let* **E** *be a finite extension of the field* **F**. *The following conditions are equivalent.*

(a) **E** *is a Galois extension of* **F**.

(b) *Every element of E is a root of a separable polynomial in* $\mathbf{F}[x]$ *and every irreducible polynomial in* $\mathbf{F}[x]$ *that has a root in E splits over* **E**.

(c) $\mathbf{F} = \operatorname{Inv}\operatorname{Gal}\mathbf{E}/\mathbf{F}$.

We say that **E** is a **separable extension** of **F** provided every element of E is a root of a separable polynomial in $\mathbf{F}[x]$. We say that **E** is a **normal extension** of **F** provided every polynomial of $\mathbf{F}[x]$ that has a root is E splits over **E**. So condition (b) in this theorem says that **E** is a normal separable extension of **F**.

Proof.
(a)⇒(c)
According to the Galois Field Closure Theorem, every intermediate field between **F** and **E** is closed. In particular, **F** is closed. This entails that $\mathbf{F} = \operatorname{Gal}(\mathbf{E}/\mathbf{F})$.

(c)⇒(b)
Let $r \in E$. Since $[\mathbf{E}:\mathbf{F}]$ is finite we know that r is algebraic over **F**. Let $m(x)$ be the minimal polynomial of r over **F**. We need to show that $m(x)$ is separable and that it splits over **E**. Now for each $\sigma \in \operatorname{Gal}\mathbf{E}/\mathbf{F}$ we know that $\sigma(r)$ is also a root of $m(x)$. Let $r_0, r_1, \dots, r_{\ell-1}$ be a list of all the distinct images of r by automorphisms belonging to $\operatorname{Gal}(\mathbf{E}/\mathbf{F})$. (This is just the orbit of r under the action of $\operatorname{Gal}(\mathbf{E}/\mathbf{F})$.) Let $f(x) = (x-r_0)(x-r_1)\dots(x-r_{\ell-1})$. The coefficients of $f(x)$ are fixed by each automorphism belonging to $\operatorname{Gal}(\mathbf{E}/\mathbf{F})$. That is these coefficients belong to $\operatorname{Inv}\operatorname{Gal}\mathbf{E}/\mathbf{F}$. So by (c) we find that $f(x) \in F[x]$. So $m(x) \mid f(x)$. On the other hand, $(x - r_i) \mid m(x)$ for each $i < \ell$. This means $f(x) \mid m(x)$. Since both $m(x)$ and $f(x)$ are monic, we see $m(x) = f(x)$. So $m(x)$ is separable and splits over **E**.

(b)⇒(a)
Since $[\mathbf{E}:\mathbf{F}]$ is finite, there are finitely many elements $s_0, \dots, s_{n-1} \in E$ so that $\mathbf{E} = \mathbf{F}[s_0, \dots, s_{n-1}]$. Let $m_i(x)$ be the minimal polynomial of s_i over **F**, for each $i < n$. According to (b), each of these polynomial is separable and splits over **E**. Let $f(x) = m_0(x)m_1(x)\dots m_{n-1}(x)$. So $f(x)$ is a separable polynomial that splits over **E**. Evidently, **E** is the splitting field of $f(x)$ over **F**. So **E** is a Galois extension of **F**. □

23.3 Problem Set 20

Algebra Homework, Edition 20
Twenty First Week
Field extensions

Problem 98.
Let $\mathbf{E}$ and $\mathbf{F}$ be fields. Prove that $\mathbf{E}$ is an algebraic closure of $\mathbf{F}$ if and only if $\mathbf{E}$ is an algebraic extension of $\mathbf{F}$ and for every algebraic extension $\mathbf{K}$ of $\mathbf{F}$ there is an embedding of $\mathbf{K}$ into $\mathbf{E}$ which fixes each element of $\mathbf{F}$.

Problem 99.
Prove that if $\mathbf{E}$ extends the field $\mathbf{F}$ and $[\mathbf{E}:\mathbf{F}] = 2$, then $\mathbf{E}$ is a normal extension of $\mathbf{F}$.

Problem 100.
Let $\mathbf{E}$ be a field extending the field $\mathbf{F}$. Let $\mathbf{L}$ and $\mathbf{M}$ be intermediate fields such that $\mathbf{L}$ is the splitting field of a separable polynomial in $\mathbf{F}[x]$. Let $\mathbf{L} \vee \mathbf{M}$ denote the smallest subfield of $\mathbf{E}$ that extends both $\mathbf{L}$ and $\mathbf{M}$. Prove that $\mathbf{L} \vee \mathbf{M}$ is a finite normal separable extension of $\mathbf{M}$ and that $\operatorname{Aut}_{\mathbf{M}}(\mathbf{L} \vee \mathbf{M}) \cong \operatorname{Aut}_{\mathbf{M} \cap \mathbf{L}} \mathbf{L}$.

Problem 101.
Let $\mathbf{L}$ and $\mathbf{M}$ be fields. Then the collection of functions from $\mathbf{L}$ into $\mathbf{M}$ can be regarded as a vector space over $\mathbf{M}$. (Add functions like we do in calculus...). Prove that the collection of field embeddings from $\mathbf{L}$ into $\mathbf{M}$ is a linearly independent set in this vector space.

Problem 102.
Let $\mathbf{F}$ be a field. We use $\mathbf{F}^\times$ to denote the group of nonzero elements of $\mathbf{F}$ under multiplication and the formation of multiplicative inverses. Show that every finite subgroup of $\mathbf{F}^\times$ is a cyclic group.

Problem 103.
Let $\mathbf{K} \subseteq \mathbf{F} \subseteq \mathbf{E}$ be fields with $[\mathbf{E}:\mathbf{F}]$ finite and let $\mathbf{A}$ be the subfield of $\mathbf{E}$ consisting of all elements of $\mathbf{E}$ that are algebraic over $\mathbf{K}$. Assume that $F \cap A = K$.

(a) Suppose $\alpha \in A$ and $f(x)$ is the monic minimal polynomial of α over $\mathbf{F}$. Show that all the coefficients of $f(x)$ lie in K.

(b) Now assume that $\mathbf{K}$ has characteristic 0. If $\mathbf{B}$ is a field with $\mathbf{K} \subseteq \mathbf{B} \subseteq \mathbf{A}$ and $[\mathbf{B}:\mathbf{K}]$ finite. Prove that $[\mathbf{B}:\mathbf{K}] \leq [\mathbf{E}:\mathbf{F}]$.

(c) Conclude that $[\mathbf{A}:\mathbf{K}] \leq [\mathbf{E}:\mathbf{F}]$.

LECTURE 24

THE GROUP SIDE OF GALOIS' CONNECTION AND THE FUNDAMENTAL THEOREM

24.1 CLOSED SUBGROUPS OF A GALOIS GROUP

Now we want to determine what the closed subgroups of a Galois group are. Since on the field side we found it convenient to look at Galois extensions, here we will focus on the case when **E** is a Galois extension of **F**.

Our first step is to develop more information on the field side. We begin with a theorem of Ernst Steinitz

Theorem on Primitive Elements—Steinitz, 1910. *Let* **E** *be a finite extension of* **F**. *The following are equivalent.*

(a) *There is an element* $r \in E$ *so that* $\mathbf{E} = \mathbf{F}[r]$.

(b) *There are only finitely many fields intermediate between* **F** *and* **E**.

The element r mentioned in (a) is a **primitive** element of **E** with respect to **F**.

Proof. Since **E** is a finite extension of **F**, we observe that E is finite if and only if F is finite. Let us first dispose of the case when either (and hence both) of these fields is finite. Of course, we have that (b) holds in this case. So to see that (a) and (b) are equivalent, we must only prove that (a) is also true. Let $\mathbf{E}^\times$ denote the group of nonzero elements of E under multiplication. This is a finite subgroup of $\mathbf{E}^\times$ (of course). But we saw last semester that such finite subgroups must be cyclic. Let r be any generator of the group $\mathbf{E}^\times$. Evidently, $\mathbf{E} = \mathbf{F}[r]$ and we have found our primitive element.

So now we turn to the case when F is infinite.

Let $\mathcal{F} = \{\mathbf{K} \mid \mathbf{F} \leq \mathbf{K} \leq \mathbf{E}\}$. That is, $\mathcal{F}$ is the collection of intermediate fields.

(a) $\Longrightarrow$ (b)
Let $r \in E$ such that $\mathbf{E} = \mathbf{F}[r]$. Let $f(x)$ be the minimal polynomial of r over **F**. Let

$$\mathcal{P} := \{g(x) \mid g(x) \text{ is a monic polynomial in } \mathbf{E}[x] \text{ that divides } f(x)\}.$$

By unique factorization for $\mathbf{E}[x]$ we see that $\mathcal{P}$ is finite. So our proof of (a) $\Longrightarrow$ (b) will be complete when we show that $\mathcal{F}$ can be mapped into $\mathcal{P}$ by a one-to-one map.

So suppose $\mathbf{K} \in \mathcal{F}$. Let $g_{\mathbf{K}}(x)$ be the minimal polynomial of r over $\mathbf{K}$. Then $g_{\mathbf{K}}(x)$ is certainly monic and irreducible. Also $g_{\mathbf{K}}(x)$ must divide $f(x)$ since the set of polynomials over $\mathbf{K}$ that have r as a root is just the ideal generated by $g_{\mathbf{K}}(x)$ and $f(x)$ belongs to this ideal since $\mathbf{F} \leq \mathbf{K}$. So let $\Phi : \mathcal{F} \to \mathcal{P}$ be defined so that

$$\Phi(\mathbf{K}) := g_{\mathbf{K}}(x) \text{ for all } \mathbf{K} \in \mathcal{F}.$$

We need to prove that Φ is one-to-one, or, what is the same, that $\mathbf{K}$ can be recovered from $g_{\mathbf{K}}(x)$.

So let $g_{\mathbf{K}}(x) = a_0 + a_1x + \cdots + a_{m-1}x^{m-1} + x^m$. Let $\mathbf{L} = \mathbf{F}[a_0, \ldots, a_{m-1}]$. Evidently $\mathbf{F} \leq \mathbf{L} \leq \mathbf{K}$ and $g_{\mathbf{K}}(x)$ is irreducible over $\mathbf{L}$. This means that $g_{\mathbf{K}}(x)$ is the minimal polynomial of r over $\mathbf{L}$. From this and Kronecker we see

$$[\mathbf{E} : \mathbf{L}] = \deg g_{\mathbf{K}}(x) = [\mathbf{E} : \mathbf{K}].$$

But we also have $[\mathbf{E} : \mathbf{K}] = [\mathbf{E} : \mathbf{L}][\mathbf{L} : \mathbf{K}]$. So it follows that $[\mathbf{L} : \mathbf{K}] = 1$. This means that $\mathbf{K} = \mathbf{L} = \mathbf{F}[a_0, \ldots, a_{m-1}]$. Therefore, $\mathbf{K}$ can indeed be recovered from $g_{\mathbf{K}}(x)$. We conclude that Φ is one-to-one and that $\mathcal{F}$ is finite. This establishes (a) $\Longrightarrow$ (b).

(b) $\Longrightarrow$ (a)
So we assume that $\mathcal{F}$ is finite and we want to prove there is a primitive element. We proceed by induction of $[\mathbf{E} : \mathbf{F}]$.

The base step of the induction is the case when $\mathbf{E} = \mathbf{F} = \mathbf{F}[1]$, which almost proves itself.

For the induction step, let $s \in E \setminus F$. Then $[\mathbf{E} : \mathbf{F}[s]] < [\mathbf{E} : \mathbf{F}]$. As there are only finitely many fields between $\mathbf{F}[s]$ and $\mathbf{E}$ condition (b) holds. By the inductive hypothesis pick $t \in E$ so that $\mathbf{E} = \mathbf{F}[s][t] = \mathbf{F}[s, t]$. For each $a \in F$ let $\mathbf{K}_a = \mathbf{F}[s + at]$. Each of the fields $\mathbf{K}_a$ is between $\mathbf{F}$ and $\mathbf{E}$. There are only finitely many intermediate fields but there are infinitely many choices for a since F is infinite. As pigeons know, this means that there are $a, b \in F$ with $a \neq b$ but $\mathbf{K}_a = \mathbf{K}_b$. Now $s + at, s + bt \in K_a$. Subtracting, we find that $(a - b)t \in K_a$. But $a - b \neq 0$ and $a - b \in F \subseteq K_a$. So we can conclude that $t \in K_a$. But $a \in F \subseteq K_a$, so $at \in K_a$. But $s + at \in K_a$ so we arrive at $s \in K_a$. But this means

$$\mathbf{E} = \mathbf{F}[s, t] \leq \mathbf{K}_a = \mathbf{F}[s + at] \leq \mathbf{E}.$$

So we can take our primitive element to be $r = s + at$. □

Corollary: Artin's Primitive Element Theorem. *Let* $\mathbf{E}$ *be a finite separable extension of* $\mathbf{F}$. *Then* $\mathbf{E}$ *has a primitive element with respect to* $\mathbf{F}$.

Proof. Since $\mathbf{E}$ is a finite extension of $\mathbf{F}$, we pick $s_0, \ldots s_{m-1} \in E$ so that $\mathbf{E} = \mathbf{F}[s_0, \ldots, s_{m-1}]$. For each $j < m$ let $f_j(x)$ be the minimal polynomial of s_j over $\mathbf{F}$. Since $\mathbf{E}$ is a separable extension, each of these minimal polynomials is separable. Let $f(x)$ be the product of all the $f_j(x)$'s. Then $f(x)$ is also separable. Let $\mathbf{L}$ be a splitting field of $f(x)$ over $\mathbf{F}$. Since $\mathbf{E} = \mathbf{F}[s_0, \ldots, s_{m-1}]$ and each s_j is a root of $f(x)$, we can insist that $\mathbf{E} \leq \mathbf{L}$. Since $\mathbf{L}$ is the splitting field over $\mathbf{F}$ of a separable polynomial, we know that $\mathbf{L}$ is a Galois extension of $\mathbf{F}$. Now $\mathrm{Gal}(\mathbf{L}/\mathbf{F})$ is finite, it is even embeddable into the symmetric group on the set of all roots of $f(x)$ in $\mathbf{L}$, which is a finite set. In particular, $\mathrm{Gal}(\mathbf{L}/\mathbf{F})$ has only finitely many subgroups. We know our Galois connection sets up a one-to-one correspondence between the fields intermediate between $\mathbf{F}$ and $\mathbf{L}$ and certain subgroups of $\mathrm{Gal}(\mathbf{L}/\mathbf{F})$. So there can only be finitely many fields intermediate between $\mathbf{F}$ and $\mathbf{L}$, and hence, between $\mathbf{F}$ and $\mathbf{E}$. By Steinitz' Theorem on Primitive Elements $\mathbf{E}$ must have a primitive element with respect to $\mathbf{F}$. □

Fact. Let $\mathbf{E}$ be a finite extension of $\mathbf{F}$. Each of the following statements holds.

(a) $|\mathrm{Gal}(\mathbf{E}/\mathbf{F})|$ divides $[\mathbf{E} : \mathbf{F}]$.

(b) $|\mathrm{Gal}(\mathbf{E}/\mathbf{F})| = [\mathbf{E} : \mathbf{F}]$ if and only if $\mathbf{E}$ is a Galois extension of $\mathbf{F}$.

Proof. Let $\bar{\mathbf{F}} = \operatorname{Inv}\operatorname{Gal}(\mathbf{F}) = \operatorname{Inv}\operatorname{Gal}(\mathbf{E}/\mathbf{F})$. We see that $\mathbf{E}$ is a Galois extension of $\bar{\mathbf{F}}$. In particular, $\mathbf{E}$ is a finite separable extension of $\bar{\mathbf{F}}$ and so it has a primitive element. Pick $r \in E$ so that $\mathbf{E} = \bar{\mathbf{F}}[r]$. Let $f(x)$ be the minimal polynomial of r over $\bar{\mathbf{F}}$.

Now $\operatorname{Gal}(\mathbf{E}/\bar{\mathbf{F}})$ acts on E. Let $\mathcal{O}$ be the orbit of r under this action. Of course, every automorphism in $\operatorname{Gal}(\mathbf{E}/\bar{\mathbf{F}})$ maps r to some other root of $f(x)$. By Kronecker, there are enough automorphisms in this Galois group to map r to each other root of $f(x)$. So the orbit $\mathcal{O}$ is exactly the set of all roots of $f(x)$. But since $f(x)$ is an irreducible separable polynomial the number of its roots is just its degree. This tells us

$$|\mathcal{O}| = \deg f(x) = [\bar{\mathbf{F}}[r] : \bar{\mathbf{F}}] = [\mathbf{E} : \bar{\mathbf{F}}].$$

But we can count the number of elements in $\mathcal{O}$ using the Key Fact about group actions.

$$|\mathcal{O}| = [\operatorname{Gal}(\mathbf{E}/\bar{\mathbf{F}}) : \mathbf{Stab}\, r].$$

Observe that $\mathbf{Stab}\, r = \{\sigma \mid \sigma \in \operatorname{Gal}(\mathbf{E}/\bar{\mathbf{F}}) \text{ and } \sigma(r) = r\}$. But $\mathbf{E} = \bar{\mathbf{F}}[r]$. So $\mathbf{Stab}\, r$ is just a one element group. This means that $|\mathcal{O}| = |\operatorname{Gal}(\mathbf{E}/\bar{\mathbf{F}})|$. Consequently,

$$[\mathbf{E} : \bar{\mathbf{F}}] = |\operatorname{Gal}(\mathbf{E}/\bar{\mathbf{F}})|.$$

Recalling that $[\mathbf{E} : \mathbf{F}] = [\mathbf{E} : \bar{\mathbf{F}}][\bar{\mathbf{F}} : \mathbf{F}]$ and $\operatorname{Gal}(\mathbf{E}/\bar{\mathbf{F}}) = \operatorname{Gal}(\mathbf{E}/\mathbf{F})$ we obtain (a) and the right to left direction of (b).

To obtain the left to right direction of (b), we need to see that if $|\operatorname{Gal}(\mathbf{E}/\mathbf{F})| = [\mathbf{E} : \mathbf{F}]$ then $\bar{\mathbf{F}} = \mathbf{F}$. From general considerations about Galois connections we know that $\operatorname{Gal}(\mathbf{E}/\mathbf{F}) = \operatorname{Gal}(\mathbf{E}/\bar{\mathbf{F}})$. But by what we saw above

$$[\mathbf{E} : \bar{\mathbf{F}}] = |\operatorname{Gal}(\mathbf{E}/\bar{\mathbf{F}})| = |\operatorname{Gal}(\mathbf{E}/\mathbf{F}) = [\mathbf{E} : \mathbf{F}].$$

Since $\mathbf{F} \leq \bar{\mathbf{F}}$ we draw the desired conclusion that $\mathbf{F} = \bar{\mathbf{F}}$. □

With this groundwork, we are prepared to examine the closed subgroups on the group side of the Galois connection.

The Galois Group Closure Theorem. *Let* $\mathbf{E}$ *be a Galois extension of* $\mathbf{F}$ *and let* $\mathbf{H}$ *be a subgroup of* $\operatorname{Gal}(\mathbf{E}/\mathbf{F})$. *Then* $\operatorname{Gal}\operatorname{Inv}\mathbf{H} = \mathbf{H}$. *In other words, every subgroup of the Galois group is closed with respect to Galois' connection.*

Proof. On general principles we know $\mathbf{H} \leq \operatorname{Gal}\operatorname{Inv}\mathbf{H}$. We need to reverse the order.

Let $\mathbf{K} = \operatorname{Inv}\mathbf{H}$. So $\mathbf{K}$ is the subfield of all elements of E fixed by every automorphism in H. Of course $\operatorname{Gal}\mathbf{K} = \operatorname{Gal}(\mathbf{E}/\mathbf{K})$.

Now $\mathbf{H}$ acts on E. For each $s \in E$ let $f_s(x)$ be the minimal polynomial of s over $\mathbf{K}$. Since $\mathbf{E}$ is a Galois extension of $\mathbf{F}$ it must also be a Galois extension of $\mathbf{K}$, so we know these polynomials are separable. Let $\mathcal{O}_s$ be the orbit of s under the action by $\mathbf{H}$. Then we see

$$[\mathbf{K}[s] : \mathbf{K}] = \deg f_s(x) = |\mathcal{O}_s| = [\mathbf{H} : \mathbf{Stab}\, s].$$

But by Lagrange, we know that $[\mathbf{H} : \mathbf{Stab}\, s]$ divides the order of $\mathbf{H}$, which is finite. This means that $[\mathbf{K}[s] : \mathbf{K}]$ is bounded above by a finite number as s ranges through E. Pick $t \in E$ so that $[\mathbf{K}[t] : \mathbf{K}]$ is as large as possible. Now we also know that $\mathbf{E}$ has a primitive element r with respect to $\mathbf{K}$. So $\mathbf{E} = \mathbf{K}[r]$. As a consequence of the Dimension Formula, we find that $\mathbf{E} = \mathbf{K}[t]$ as well. Putting this together with the Fact proved just above, we get

$$|\operatorname{Gal}\operatorname{Inv}\mathbf{H}| = |\operatorname{Gal}(\mathbf{E}/\mathbf{K})| = [\mathbf{E} : \mathbf{K}] = [\mathbf{H} : \mathbf{Stab}\, t] \leq |H|.$$

But we know $\mathbf{H} \leq \operatorname{Gal}\operatorname{Inv}\mathbf{H}$ and that these groups are finite. Therefore $\mathbf{H} = \operatorname{Gal}\operatorname{Inv}\mathbf{H}$, as desired. □

24.2 The Fundamental Theorem of Galois Theory

It is traditional to gather together the bits and pieces up to this point and package them into one theorem. Here it is.

The Fundamental Theorem of Galois Theory. *Let* $\mathbf{E}$ *be a Galois extension of* $\mathbf{F}$. *Then the following hold.*

(a) *The closed sets on the field side of Galois' connection are exactly the fields intermediate between* $\mathbf{F}$ *and* $\mathbf{E}$.

(b) *The closed sets of the group side of Galois' connection are exactly the subgroups of the Galois group* $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$.

(c) *Polarities of Galois' connection, namely* Inv *and* Gal, *are inverses of each other and establish an anti-isomorphism between the two lattices of closed sets.*

(d) $[\mathbf{E}:\mathbf{K}] = |\mathrm{Gal}\,\mathbf{K}|$ *and* $[\mathbf{K}:\mathbf{F}] = [\mathrm{Gal}\,\mathbf{F}:\mathrm{Gal}\,\mathbf{K}]$, *for each intermediate field* $\mathbf{K}$. *In particular,* $|\mathrm{Gal}(\mathbf{E}/\mathbf{F})| = [\mathbf{E}:\mathbf{F}]$.

(e) *Let* $\mathbf{H}$ *be any subgroup of* $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$. *Then* $\mathbf{H} \lhd \mathrm{Gal}(\mathbf{E}/\mathbf{F})$ *if and only if* $\mathrm{Inv}\,\mathbf{H}$ *is a normal extension of* $\mathbf{F}$. *In this case,* $\mathrm{Gal}(\mathrm{Inv}\,\mathbf{H}/\mathbf{F}) \cong \mathrm{Gal}(\mathbf{E}/\mathbf{F})/\mathbf{H}$.

Proof. The only parts that need attention, perhaps, are (d) and (e).

For (d), notice that $\mathrm{Gal}\,\mathbf{K} = \mathrm{Gal}(\mathbf{E}/\mathbf{K})$. We know that $\mathbf{E}$ is a Galois extension of $\mathbf{K}$, so by the Fact immediately preceding the Galois Group Closure Theorem, we see $|\mathrm{Gal}\,\mathbf{K}| = [\mathbf{E}:\mathbf{K}]$ as well as $|\mathrm{Gal}\,\mathbf{F}| = [\mathbf{E}:\mathbf{F}]$. The Dimension Formula tells us

$$[\mathbf{E}:\mathbf{F}] = [\mathbf{E}:\mathbf{K}][\mathbf{K}:\mathbf{F}]$$

and Lagrange tells us

$$|\mathrm{Gal}\,\mathbf{F}| = [\mathrm{Gal}\,\mathbf{F}:\mathrm{Gal}\,\mathbf{K}]|\mathrm{Gal}\,\mathbf{K}|.$$

A bit of twiddling extracts $[\mathbf{K}:\mathbf{F}] = [\mathrm{Gal}\,\mathbf{F}:\mathrm{Gal}\,\mathbf{K}]$ from these equations. This secures (d).

For (e), suppose first the $\mathbf{H}$ is a normal subgroup of $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$. Let $s \in \mathrm{Inv}\,\mathbf{H}$. We need to see that the minimal polynomial $f(x)$ of s splits in $\mathrm{Inv}\,\mathbf{H}$. Now $f(x)$ certainly splits in $\mathbf{E}$ since $\mathbf{E}$ is a normal extension of $\mathbf{F}$. Let $r \in E$ be a root of $f(x)$. What we need is to show that $r \in \mathrm{Inv}\,\mathbf{H}$. Relying on Kronecker, we pick $\sigma \in \mathrm{Gal}(\mathbf{E}/\mathbf{F})$ so that $\sigma(s) = r$. So we must show that $\tau(r) = r$ for every $\tau \in H$. But $\sigma H \sigma^{-1} = H$ by normality of the subgroup. This means what we have to show is $\sigma \circ \tau \circ \sigma^{-1}(r) = r$. But this is immediate:

$$\sigma \circ \tau \circ \sigma^{-1}(r) = \sigma(\tau(\sigma^{-1}(r))) = \sigma(\tau(s)) = \sigma(s) = r.$$

So we see that $\mathrm{Inv}\,\mathbf{H}$ is a normal extension of $\mathbf{F}$.

Now suppose that $\mathrm{Inv}\,\mathbf{H}$ is a normal extension of $\mathbf{F}$. Let $\sigma \in \mathrm{Gal}(\mathbf{E}/\mathbf{F})$ and let $r \in \mathrm{Inv}\,\mathbf{H}$. Let $f(x)$ be the minimal polynomial of r over $\mathbf{F}$. Then $\sigma(r)$ must also be a root of $f(x)$. But $f(x)$ splits in $\mathrm{Inv}\,\mathbf{H}$. So $\sigma(r) \in \mathrm{Inv}\,\mathbf{H}$. This means that the restriction $\sigma \restriction_{\mathrm{Inv}\,\mathbf{H}}$ belongs to $\mathrm{Gal}(\mathrm{Inv}\,\mathbf{H}/\mathbf{F})$. Now define $\Phi : \mathrm{Gal}(\mathbf{E}/\mathbf{F}) \to \mathrm{Gal}(\mathrm{Inv}\,\mathbf{H}/\mathbf{F})$ via

$$\Phi(\sigma) = \sigma \restriction_{\mathrm{Inv}\,\mathbf{H}}, \text{ for all } \sigma \in \mathrm{Gal}(\mathbf{E}/\mathbf{F}).$$

The eager graduate students will find it easy to show that Φ is a homomorphism onto the group $\mathrm{Gal}(\mathrm{Inv}\,\mathbf{H}/\mathbf{F})$ and that its kernel is $\mathbf{H}$ (because $\mathbf{H}$ is closed). So we see that $\mathbf{H}$ is a normal subgroup of $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$ and, by the Homomorphism Theorem, that

$$\mathrm{Gal}(\mathbf{E}/\mathbf{F})/\mathbf{H} \cong \mathrm{Gal}(\mathrm{Inv}\,\mathbf{H}/\mathbf{F}),$$

as desired. □

24.3 Problem Set 21

Algebra Homework, Edition 21
Twenty Second Week
Galois extensions

Problem 104.
Assume that **L** is a Galois extension of the field $\mathbb{Q}$ of rational numbers and that $\mathbf{K} \subseteq \mathbf{L}$ is the subfield generated by all the roots of unity in **L**. Suppose that $\mathbf{L} = \mathbb{Q}[a]$, where $a^n \in \mathbb{Q}$ for some positive integer n. Show that the Galois group $\mathrm{Gal}(\mathbf{L}/\mathbb{Q})$ is cyclic.

Problem 105.
Let the field **L** be an algebraic extension of the field **K**. An element a of L is called **Abelian** if $\mathbf{K}[a]$ is a Galois extension of **K** with an Abelian Galois group $\mathrm{Gal}(\mathbf{K}[a]/\mathbf{K})$. Show that the set of Abelian elements of L is a subfield of **L** containing **K**.

Problem 106.
Let **L** be a finite extension of a field **K** and **M** be a finite extension of **L**. For each of the extensions $\mathbf{L}/\mathbf{K}, \mathbf{M}/\mathbf{L}, \mathbf{M}/\mathbf{K}$ is it possible to choose the fields so that the extension in question is not Galois while the other two extensions are each Galois? Explain thoroughly.

Problem 107.
Prove that every finite extension of a finite field has a primitive element.

Problem 108.
Let $\mathbb{Q}$ denote the field of rational numbers and $\mathbb{C}$ the field of complex numbers.

(a) Suppose **K** and **L** are subfields of $\mathbb{C}$, each of which is Galois over $\mathbb{Q}$. Show that the field **E** generated by K and L is Galois over $\mathbb{Q}$.

(b) Suppose in part (a), the degrees $[\mathbf{K} : \mathbb{Q}]$ and $[\mathbf{L} : \mathbb{Q}]$ are relatively prime. Show that $\mathrm{Gal}(\mathbf{E}/\mathbb{Q})$ is isomorphic to the direct product of the groups $\mathrm{Gal}(\mathbf{K}/\mathbb{Q})$ and $\mathrm{Gal}(\mathbf{L}/\mathbb{Q})$. Deduce that $[\mathbf{E} : \mathbb{Q}] = [\mathbf{K} : \mathbb{Q}][\mathbf{L} : \mathbb{Q}]$.

(c) Prove that there exists a subfield **F** of $\mathbb{C}$ such that **F** is Galois over $\mathbb{Q}$ with $[\mathbf{F} : \mathbb{Q}] = 55$.

Lecture 25

Galois' Criterion for Solvability by Radicals

Given a field $\mathbf{F}$ and a polynomial $f(x) \in \mathbf{F}[x]$ the task of explicitly describing, in some manner, all the roots of $f(x)$ is a project that is most appropriate carried forward in the splitting field of the polynomial. So let $\mathbf{E}$ be the splitting field of $f(x)$ over $\mathbf{F}$. The set of all roots of $f(x)$ is a finite set, so it would be possible to simply make a list of all these elements. But it is not apparent, just given $f(x)$ how such a list might be made. Just trying to use the field operations and the coefficients of $f(x)$ is not even adequate for describing all the roots of $x^2 - 2$. By permitting the extracting of square roots, we can resolve this case and that of all polynomials of degree no more than 2. By allowing the extraction of cube roots, fourth roots, and so on, one might hope to succeed, at least with some frequency. The problem of determining when success is possible is what Galois undertook.

Recall how we approached the notion of a constructible number. We envisioned a tower of field extensions so that later fields in the tower were obtained by adjoining a square root to an earlier field. We simply expand our horizons by allowing the adjunction of k^{th} roots for any positive integer k. More precisely, we say that

$$\mathbf{F} = \mathbf{F}_0 \leq \mathbf{F}_1 \leq \cdots \leq \mathbf{F}_{m-1}$$

is a **radical tower** over $\mathbf{F}$ provided for all $i < m$

$$\mathbf{F}_{i+1} = \mathbf{F}_i[r] \text{ for some } r \text{ such that } r^k \in F_i \text{ for some positive integer } k.$$

We will say that $\mathbf{K}$ is a **radical extension** of $\mathbf{F}$ provided $\mathbf{K}$ extends $\mathbf{F}$ and $\mathbf{K} \leq \mathbf{F}_{m-1}$ for some radical tower $\mathbf{F}_0 \leq \mathbf{F}_1 \leq \cdots \leq \mathbf{F}_{m-1}$ over $\mathbf{F}$.

We say a polynomial $f(x) \in \mathbf{F}[x]$ is **solvable by radicals** over $\mathbf{F}$ exactly when the splitting field of $f(x)$ over $\mathbf{F}$ is a radical extension of $\mathbf{F}$.

By the **Galois group** of $f(x)$ we mean the group $\operatorname{Gal}(\mathbf{E}/\mathbf{F})$, where $\mathbf{E}$ is the splitting field of $f(x)$ over $\mathbf{F}$.

Galois' Criterion for Solvability by Radicals. *Let $\mathbf{F}$ be a field of characteristic 0 and let $f(x) \in \mathbf{F}[x]$. The polynomial $f(x)$ is solvable by radicals over $\mathbf{F}$ if and only if the Galois group of $f(x)$ over $\mathbf{F}$ is a solvable group.*

We need a few preliminaries to prepare the way.

Lemma 25.0.1. *Over any field $\mathbf{F}$, the polynomial $x^p - a$, with p a prime number and $a \in F$, either has a root in F or it is irreducible over $\mathbf{F}$.*

Proof. This is clear if $a = 0$, so we consider that $a \neq 0$. There are two cases.

Case: p is not the characteristic of F.
In this case we know that $x^p - a$ must have distinct roots. Let **E** be the splitting field of $x^p - a$ over **F**. So

$$x^p - a = (x - r_0)(x - r_1)\ldots(x - r_{p-1})$$

where $r_0, \ldots, r_{p-1}$ are the p distinct roots of $x^p - a$. Notice that $r_0 \neq 0$. So

$$1 = \frac{r_0}{r_0}, \quad \frac{r_1}{r_0}, \ldots, \frac{r_{p-1}}{r_0}$$

must be the p distinct p^{th} roots of unity. Let ζ be a primitive p^{th} root of unity in E. This means that

$$r_0, \zeta r_0, \zeta^2 r_0, \ldots, \zeta^{p-1} r_0$$

are the roots of $x^p - a$.

Now consider the possibility that $x^p - a$ is reducible in $\mathbf{F}[x]$. We desire to show that $x^p - a$ has a root in F. For some k with $1 \le k < p$ we can render a factor of $x^p - a$ as a product of k factors, each of the form $x - \zeta^j r_0$. computing the constant term of this product, we find $\zeta^\ell r_0^k \in F$ for some ℓ. Put $b = \zeta^\ell r_0^k$. Now $b^p = \zeta^{p\ell} r_0^{pk} = a^k$. Since $1 \le k < p$ and p is a prime number, we see that k and p are relatively prime. Pick integers s and t so that $1 = sk + tp$. We get

$$a = a^1 = a^{sk+tp} = (a^k)^s (a^t)^p = (b^p)^s (a^t)^p = (b^s a^t)^p.$$

This means that a has a p^{th} root in F, namely $b^s a^t$. Hence $x^p - a$ has a root in F. This finishes the first case.

Case: F has characteristic p.
In the splitting field **E** pick a root r of $x^p - a$. It follows that $x^p - a = x^p - r^p = (x - r)^p$. Consider the case that $x^p - a$ is reducible in $\mathbf{F}[x]$. This means that for some k with $1 \le k < p$ we will have $(x - r)^k \in \mathbf{F}[x]$. Computing the constant term, we find $r^k \in F$. Put $b = r^k$. Hence $b^p = (r^k)^p = (r^p)^k = a^k$. As above, we have integers s and t so that $1 = sk + tp$. Just as above, we have $a = (b^s a^t)^p$. This makes $b^s a^t$ a root of $x^p - a$, as desired. □

Theorem 25.0.2. *Let p be a prime that is not the characteristic of* **F** *and let $a \in F$. The Galois group of $x^p - a$ over* **F** *is solvable.*

Proof. Let **E** be the splitting field of $x^p - a$ over **F**. As in the proof of the lemma above, **E** has p distinct p^{th} roots of unity. Let ζ be a primitive one. Let r be any root of $x^p - a$. Then we have seen that $\mathbf{E} = \mathbf{F}[\zeta, r]$. We also know that $x^p - a$ is separable. So the Fundamental Theorem of Galois Theory applies here.

The field $\mathbf{F}[\zeta]$ is the splitting field of the separable polynomial $x^p - 1$ over **F**. So $\mathbf{F}[\zeta]$ is a normal extension of **F**. By the Fundamental Theorem, $\text{Gal}(\mathbf{E}/\mathbf{F}[\zeta]$ is a normal subgroup of $\text{Gal}(\mathbf{E}/\mathbf{F})$ and

$$\text{Gal}(\mathbf{F}[\zeta]/\mathbf{F}) \cong \text{Gal}(\mathbf{E}/\mathbf{F}) / \text{Gal}(\mathbf{E}/\mathbf{F}[\zeta]).$$

Observe that every automorphism belonging to $\text{Gal}(\mathbf{F}[\zeta]/\mathbf{F})$ is determined by what it does to ζ (which it must map to another root of unity). Thus restriction is actually an embedding of $\text{Gal}(\mathbf{F}[\zeta]/\mathbf{F})$ into the group of automorphisms of the group of p^{th} roots of unity. But the group of p^{th} roots of unity is just a copy of the cyclic group $\mathbb{Z}_p$. It is an exercise (to be carried out by the diligent graduate students) to prove that $\text{Aut}\,\mathbb{Z}_p \cong \mathbb{Z}_{p-1}$. In this way we see that $\text{Gal}(\mathbf{F}[\zeta]/\mathbf{F})$ is embeddable into the cyclic group $\mathbb{Z}_{p-1}$. But every subgroup of a cyclic group is cyclic, so $\text{Gal}(\mathbf{F}[\zeta])/\mathbf{F})$ is cyclic.

Now consider the group $\text{Gal}(\mathbf{E}.\mathbf{F}[\zeta])$. In case $x^p - a$ has a root in $\mathbf{F}[\zeta]$, then all its roots are in $\mathbf{F}[\zeta]$. This means $\mathbf{E} = \mathbf{F}[\zeta, r] = \mathbf{F}[\zeta]$. So $\text{Gal}(\mathbf{E}/\mathbf{F}[\zeta])$ is a trivial group. In case $x^p - a$ has no root in $\mathbf{F}[\zeta]$ we know by the lemma that $x^p - a$ is irreducible over $\mathbf{F}[\zeta]$. So by Kronecker, $p = [\mathbf{F}[\zeta, r] : \mathbf{F}[\zeta]] = |\mathrm{Gal}(\mathbf{E}/\mathbf{F}[\zeta])|$. This means that $\mathrm{Gal}(\mathbf{E}/\mathbf{F}[\zeta])$ is a cyclic group of order p.

So the normal series $\mathrm{Gal}(\mathbf{E}/\mathbf{F}) \rhd \mathrm{Gal}(\mathbf{E}/\mathbf{F}[\zeta]) \rhd 1$ has cyclic factor groups. Therefore $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$ is solvable. □

Lemma 25.0.3. *Let p be a prime number and suppose that the field $\mathbf{F}$ contains p distinct p^{th} roots of unity. Let $\mathbf{K}$ extend $\mathbf{F}$ so that $[\mathbf{K} : \mathbf{F}] = p$ and so that $\mathrm{Gal}(\mathbf{K}/\mathbf{F})$ is cyclic of order p. Then $\mathbf{K} = \mathbf{F}[d]$ for some d such that $d^p \in F$.*

Proof. Let η generate the cyclic group $\mathrm{Gal}(\mathbf{K}/\mathbf{F})$ and let ζ be a primitive p^{th} root of unity in F.

Begin by picking $c \in K \setminus F$. Then $\mathbf{K} = \mathbf{F}[c]$ because p is prime, by the Dimension Formula there can be no fields properly intermediate between $\mathbf{F}$ and $\mathbf{K}$.

For each $i < p$, put $c_i = \eta^i(c)$. So we get

$$\begin{aligned} c_0 &= c \\ c_{i+1} &= \eta(c_i) \text{ for all } i < p-1 \\ c_0 &= \eta(c_{p-1}) \end{aligned}$$

Put

$$d_i = c_0 + c_1\zeta^i + c_2\zeta^{2i} + \cdots + c_{p-1}\zeta^{(p-1)i} \text{ for } i < p. \tag{$\star$}$$

A straightforward computation shows $\eta(d_i) = \zeta^{-i} d_i$ for all $i < p$. Hence

$$\eta(d_i^p) = (\eta(d_i))^p = (\zeta^{-i} d_i)^p = 1 \cdot d_i^p$$

for all $i < p$. Since the generator of $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$ fixes each d_i^p, we find that each d_i^p belongs to the fixed field, namely to F. (The fixed field must be $\mathbf{F}$, for lack of other intermediate fields.)

It remains to show that $d_i \notin F$ for some i, for then we can take that d_i to be our desired d. Let us render the system ($\star$) of equations in matrix form.

$$\begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & \zeta & \zeta^2 & \dots & \zeta^{p-1} \\ 1 & \zeta^2 & \zeta^4 & \dots & \zeta^{2(p-1)} \\ 1 & \zeta^3 & \zeta^6 & \dots & \zeta^{3(p-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \zeta^{p-1} & \zeta^{2(p-1)} & \dots & \zeta^{(p-1)(p-1)} \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \\ \vdots \\ c_{p-1} \end{pmatrix} = \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ d_3 \\ \vdots \\ d_{p-1} \end{pmatrix}$$

The $p \times p$ matrix displayed above is invertible, since it is a Vandermonde matrix. This means that the column vector of c_i's can be obtained by multiplying the column vector of d_i's by the inverse of the Vandermonde matrix—which is a matrix over $\mathbf{F}$. In particular, this means that $c = c_0$ is an $\mathbf{F}$-linear combination of the d_i's. Since $c \notin F$, we see that at least one of the d_i's must also fail to be in F. This completes the proof of the lemma. □

We need one more lemma.

Lemma 25.0.4. *Let $\mathbf{F}$ be a field of characteristic 0. Any radical extension of $\mathbf{F}$ can be embedded into a separable normal radical extension of $\mathbf{F}$ that has a solvable Galois group over $\mathbf{F}$.*

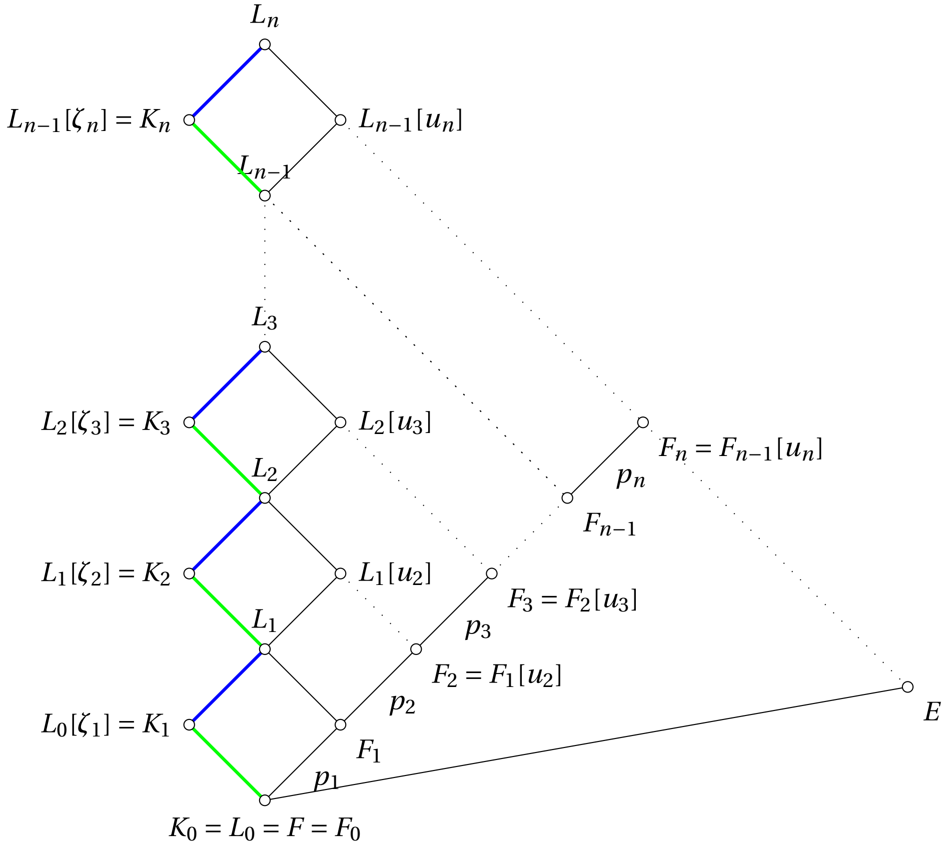
Proof. We do this by induction on the number of intermediate fields in the root tower leading to the radical extension. The base step (when the number of intermediate fields in 0) is evident. So consider the inductive step. Suppose the last step in the given radical tower in $\mathbf{F}_k < \mathbf{F}_k[u_k]$ where $u_k^p \in F_k$ for some prime p. By the inductive hypothesis, we suppose that we have in hand $\mathbf{L}_k$, which is a normal separable radical extension of $\mathbf{F}$ such that $\mathbf{L}_k$ extends $\mathbf{F}_k$, and that $\mathrm{Gal}(\mathbf{L}_k/\mathbf{F})$ is solvable. To conserve notation, let $\mathbf{G} = \mathrm{Gal}(\mathbf{L}_k/\mathbf{F})$ and let $G = \{\sigma_0, \dots, \sigma_{m-1}\}$. Put $a = u_k^p$. Let

$$f(x) = \prod_{\sigma \in G} (x^p - \sigma(a)).$$

Observe that the coefficients of $f(x)$ are fixed by every $\tau \in G$. Since $\mathbf{F}$ is the fixed field of $\mathbf{G}$, we see that $f(x) \in F[x]$. Let $\mathbf{L}_{k+1}$ be the splitting field of $f(x)$ over $\mathbf{L}_k$. Now, as we saw several times above, each $x^p - \sigma(a)$ is a separable polynomial, so in L_{k+1} we must have a primitive p^{th} root ζ of unity. For each $i < m$ pick $r_i \in L_{k+1}$ that is a root of $x^p - \sigma_i(a)$. This entails that $\mathbf{L}_{k+1} = \mathbf{L}_k[\zeta, r_0, \dots, r_{m-1}]$. Now consider the following root tower of fields:

$$\mathbf{F} \le \mathbf{L}_k \le \mathbf{L}_k[\zeta] \le \mathbf{L}_k[\zeta, r_0] \le \cdots \le \mathbf{L}_k[\zeta, r_0, \dots, r_{m-1}] = \mathbf{L}_{k+1}.$$

The field obtained at each step above $\mathbf{L}_k$ is a normal extension of the previous field. Moreover, the Galois group associated to each step are cyclic. This entails that the associated series of subgroups of $\mathbf{G}$ above $\mathrm{Gal}(\mathbf{E}/\mathbf{L}_k)$ have cyclic factors. Since we also have that $\mathrm{Gal}(\mathbf{L}_k, \mathbf{F})$ is solvable, it follows that $\mathrm{Gal}(\mathbf{L}_{k+1}/\mathbf{F})$ is solvable. □



Normalizing a Root Tower over E

Here ζ_i is a primitive p_i^{th} root of unity and $u_i^{p_i} \in F_{i-1}$. The blue lines represent root towers in their own right along which only p_i^{th} roots are extracted, where p_i is the prime labelling the parallel edge in the original root tower. Each L_i is a separable normal extension of F. The zigzag path along the left edge of the diagram is a root tower.

Now we are ready.

Proof of Galois' Criterion. First, let us suppose that $f(x)$ is solvable by radicals. Let $\mathbf{E}$ be the splitting field of $f(x)$ over $\mathbf{F}$. We aim to show that $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$ is a solvable group. Use the lemma just above to obtain a field $\mathbf{L}$ that is a separable normal radical extension of $\mathbf{F}$ that also extends $\mathbf{E}$. Observe $\mathrm{Gal}(\mathbf{L}/\mathbf{E}) \triangleleft \mathrm{Gal}(\mathbf{L}/\mathbf{F})$, since $\mathbf{E}$ is a normal extension of $\mathbf{F}$. Moreover,

$$\mathrm{Gal}(\mathbf{E}/\mathbf{F}) \cong \mathrm{Gal}(\mathbf{L}/\mathbf{F}) / \mathrm{Gal}(\mathbf{L}/\mathbf{E}).$$

This means that $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$ is a homomorphic image of the solvable group $\mathrm{Gal}(\mathbf{L}/\mathbf{F})$. Hence, $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$ is solvable.

For the converse, suppose that $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$ is solvable. Let $n = |\mathrm{Gal}(\mathbf{E}/\mathbf{F})| = [\mathbf{E} : \mathbf{F}]$. Let $r_0, \dots, r_{m-1}$ be the roots of $f(x)$ in E. Let ζ be a primitive n^{th} root of unity. (It might not be in E.) Observe that $\mathbf{E}[\zeta]$ is a splitting field of $f(x)$ over $\mathbf{F}[\zeta]$. Let $\eta \in \mathrm{Gal}(\mathbf{E}[\zeta]/\mathbf{F}[\zeta])$. Then η will permute the elements of $\{r_0, \dots, r_{m-1}\}$. The elements generate $\mathbf{E}$ over $\mathbf{F}$. This means that restricting η to E produces a member of $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$. Indeed, an argument routine by now shows that restriction to E is an embedding of $\mathrm{Gal}(\mathbf{E}[\zeta]/\mathbf{F}[\zeta])$ into $\mathrm{Gal}(\mathbf{E}/\mathbf{F})$. This latter group is solvable. Since subgroups of solvable groups are themselves solvable, we find that $\mathrm{Gal}(\mathbf{E}[\zeta]/\mathbf{F}[\zeta])$ is solvable. Let

$$G(\mathbf{E}[\zeta]/\mathbf{F}[\zeta]) = \mathbf{H}_0 \triangleright \mathbf{H}_1 \triangleright \cdots \triangleright \mathbf{H}_{\ell-1}$$

be a composition series. So its factor groups must be cyclic of prime order. Let

$$\mathbf{F} \leq \mathbf{F}[\zeta] = \mathbf{K}_0 \leq \mathbf{K}_1 \leq \cdots \leq \mathbf{K}_{\ell-1} = \mathbf{E}[\zeta]$$

be the corresponding tower of fixed fields. By our lemmas, each step of this tower is made by adding a k^{th} root of an element of the previous field, for some k. This means that $\mathbf{E}[\zeta]$ is a radical extension of $\mathbf{F}$. Since $\mathbf{E} \leq \mathbf{E}[\zeta]$, we see that $f(x)$ is solvable by radicals. □

LECTURE 26

Polynomials and Their Galois Groups

In order to take advantage of Galois' Criterion for Solvability by Radicals we need at least some way to start with a polynomial $f(x)$ and find its Galois group. With the group in hand, we may be able to determine whether it is a solvable group. Such a group is, after all, finite and even if no more elegant approach is at hand it would be possible to undertake, with computational assistance, the brute force examination of its subgroup lattice. At any rate we see a two step process:

- Given $f(x)$ construct its Galois group $\mathbf{G}$.
- Given a finite group $\mathbf{G}$ determine whether it is solvable.

Even over the field of rational numbers this situation has been the focus of very considerable mathematical effort and is still not well understood.

Here we will see just three results: The determination of which symmetric groups $\mathbf{S}_n$ are solvable and two conditions, each sufficient to ensure that the Galois group of $f(x)$ is not solvable.

The following theorem is due to Galois.

The Solvability of Symmetric Groups. *The groups $\mathbf{S}_1, \mathbf{S}_2, \mathbf{S}_3$, and $\mathbf{S}_4$ are solvable. The groups $\mathbf{S}_n$, where $4 < n$, are not solvable.*

Proof. The groups $\mathbf{S}_1$ and $\mathbf{S}_2$ have one and two elements respectively. They are evidently solvable. Observe the $\mathbf{S}_3 \triangleright \mathbf{A}_3$ (since $[\mathbf{S}_3 : \mathbf{A}_3] = 2$) and that $\mathbf{A}_3$ is the three element group, which is simple and Abelian . This gives us a normal series with Abelian factors, witnessing the solvability of $\mathbf{S}_3$. We can try the same thing with $\mathbf{S}_4$:

$$\mathbf{S}_4 \triangleright \mathbf{A}_4 \triangleright \dots$$

Now $\mathbf{A}_4$ has twelve elements. Let $\mathbf{V}$ be the subgroup of $\mathbf{A}_4$ consisting of the identity permutation and the following three permutations:

$$(0,1)(2,3),\ (0,2)(1,3),\ \text{and}\ (0,3)(1,2).$$

Direct calculation reveals that these elements constitute an Abelian subgroup of $\mathbf{A}_4$ and that this subgroup is normal. (It is the Sylow 2-subgroup of $\mathbf{A}_4$). So we see

$$\mathbf{S}_4 \triangleright \mathbf{A}_4 \triangleright \mathbf{V} \triangleright \mathbb{1}$$

is a normal series with Abelian factors, witnessing that $\mathbf{S}_4$ is solvable.

To see, on the other hand, that $\mathbf{S}_n$ is not solvable when $n > 4$ we will show that $\mathbf{A}_n$ is simple: it has no proper nontrivial normal subgroups. This will mean that $\mathbf{S}_n \rhd \mathbf{A}_n \rhd \mathbb{1}$ is a composition series. Since $\mathbf{A}_n$ is not Abelian this will demonstrate that $\mathbf{S}_n$ is not solvable. Or we might simply note that $\mathbf{A}_n$ is not solvable, itself—so neither is $\mathbf{S}_n$, since every subgroup of a solvable group is also solvable.

So let $\mathbf{N} \lhd \mathbf{A}_n$ and suppose that $\mathbf{N}$ is nontrivial. We have to prove that $\mathbf{N} = \mathbf{A}_n$. We use the following fact:

Fact. $\mathbf{A}_n$ is generated by the set of all 3-cycles, if $3 \le n$.

The verification of the fact is left as an entertainment for graduate students.

Let us first see that $\mathbf{N}$ has at least one 3-cycle. Each element of $\mathbf{A}_n$ is a permutation of $\{0,1,2,3,4,\ldots,n-1\}$. A permutation might have fixed points. (The identity fixes all n points.) Let $\alpha \in A_n$ fix as many points as possible while still being different from the identity permutation. Consider the decomposition of α into a product of disjoint cycles. There are three cases.

Let us first suppose that the longest cycle in the decomposition has length 2. As α is even, it does no harm to suppose

$$\alpha = (0,1)(2,3)\cdots.$$

Let $\beta = (2,3,4) \in A_n$. Notice that $\beta\alpha\beta^{-1} \in N$ by normality. So $\beta\alpha\beta^{-1}\alpha^{-1} \in N$ as well. Direct computations show that 0 and 1 are fixed points of $\beta\alpha\beta^{-1}\alpha^{-1}$. Also direct computation shows that 2 is not fixed, so that $\beta\alpha\beta^{-1}\alpha^{-1}$ is not the identity. The only points moved by β are 2, 3, and 4. Now any point that is fixed by both α and β is fixed by $\beta\alpha\beta^{-1}\alpha^{-1}$. It might be that 4 is a fixed point of α. In that event, it would not be fixed by $\beta\alpha\beta^{-1}\alpha^{-1}$. But in any event, $\beta\alpha\beta^{-1}\alpha^{-1}$ has at least one more fixed point than α. This is contrary to the choice of α, so we reject this case.

Second, suppose that α is itself a 3-cycle. Well, this is what we want.

The third case remains: the decomposition of α has a cycle of length at least three but α is not itself a 3-cycle. So we consider that

$$\alpha = (0,1,2,\ldots)\ldots.$$

Since α is even it cannot move just 4 points. That would make α a 4-cycle and 4-cycles are odd. So α must move at least two points in addition to 0, 1, and 2. Call these points 3 and 4. Again let $\beta = (2,3,4)$ and consider $\beta\alpha\beta^{-1}\alpha^{-1}$. For any point d fixed by α, we have $4 < d < n$. So d is also fixed by $\beta\alpha\beta^{-1}\alpha^{-1}$. But direct computation still shows that 1 is also fixed by the latter permutation, while 2 is not. So that composite permutation has more fixed points than α. We have to reject this case.

Our conclusion is that $\alpha \in N$ is a 3-cycle. Say $\alpha = (0,1,2)$.

Now, suppose that i, j, k, ℓ, and m are distinct elements of $\{0,1,2,3,4,\ldots\}$. Let γ be the permutation so that

$$\begin{aligned}
\gamma(0) &= i\\
\gamma(1) &= j\\
\gamma(2) &= k\\
\gamma(3) &= \ell\\
\gamma(4) &= m\\
&\vdots
\end{aligned}$$

Now either γ is even or $(\ell, m)\gamma$ is even. Let λ be the one of these that is even. Then direct calculation shows $\lambda\alpha\lambda^{-1} = (i, j, k)$. This means that N contains all the 3-cycles. That means $\mathbf{N} = \mathbf{A}_n$.

So we conclude that $\mathbf{A}_n$ is simple and that $\mathbf{S}_n$ is not solvable. □

So how can we ensure that a polynomial $f(x)$ has Galois group $\mathbf{S}_n$?

The $\mathbf{S}_p$ Criterion for Polynomials over $\mathbb{Q}$. *Let p be a prime number and let $f(x) \in \mathbb{Q}[x]$ be of degree p. If $f(x)$ is irreducible and $f(x)$ has exactly two non-real roots in $\mathbb{C}$, then the Galois group of $f(x)$ over $\mathbb{Q}$ is $\mathbf{S}_p$ and $f(x)$ is not solvable by radicals over $\mathbb{Q}$ if $p > 4$.*

Proof. Let $r_0, \ldots, r_{p-3}$ be the real roots of $f(x)$ and let r_{p-2} and r_{p-1} be the non-real complex roots. Let $\mathbf{E} = \mathbb{Q}[r_0, \ldots, r_{p-1}]$ be the splitting field of $f(x)$. By Kronecker we know that $[\mathbb{Q}[r_0] : \mathbb{Q}] = p$. By the Dimension Formula, we see that $p \mid [\mathbf{E} : \mathbb{Q}]$. By the Fundamental Theorem of Galois Theory, this means $p \mid |\mathrm{Gal}(\mathbf{E}/\mathbb{Q})|$. By Cauchy, this Galois group must have an element of order p. On the other hand, complex conjugation is an automorphism of $\mathbb{C}$ that fixes every real. In particular, all the coefficients of $f(x)$ are fixed by conjugation. So conjugation must permute the roots of $f(x)$. This entails that r_{p-1} is the complex conjugate of r_{p-2}. By restricting our group to the p-element set $\{r_0, \ldots, r_{p-1}\}$ we see that this subgroup of $\mathbf{S}_p$ has an element of order p and a transposition (inherited from complex conjugation). It is a fact for the entertainment of graduate students that for any prime p, that $\mathbf{S}_p$ is generated by any transposition and any element of order p. So the Galois group of $f(x)$ is isomorphic to $\mathbf{S}_p$. □

It is easy to devise polynomials of prime degree with integer coefficients that meet these criteria. For example, suppose we want $f(x)$ to be of degree 5. We want it to have 3 real roots and 2 non-real complex roots (which we know must be complex conjugates). So we could just pick any three real numbers r_0, r_1, and r_2 and any non-real complex number s and let

$$f(x) = (x - r_0)(x - r_1)(x - r_2)(x - s)(x - \bar{s}).$$

But we have two problems: the coefficients of $f(x)$ might not be rational and even if they were $f(x)$ might not be irreducible over $\mathbb{Q}$. For this approach to work, r_0, r_1, r_2, and s must at least be algebraic and $f(x)$ must be their common minimal polynomial. This is harder to arrange, but still possible.

But there is another approach advanced by Richard Brauer. The idea is to use the curve sketching techniques of freshman calculus. The graph of $f(x)$ should cross the X-axis exactly 3 times. This can be arranged if $f(x)$ has a unique relative maximum (and $f(x)$ has a positive value there) and a unique relative minimum (and $f(x)$ has a negative value there). This suggests looking at the derivative $f'(x)$. This derivative will have degree 4 and we want it to have 2 real roots. Let's try

$$f'(x) = 5x^4 - 5 \cdot 16 = 5(x^2 + 4)(x^2 - 4) = 5(x^2 + 4)(x - 2)(x + 2)$$

This means that $f(x) = x^5 - 5 \cdot 16x + c$ where c is the constant of integration but is subject to the following constraints:

$$\begin{gathered} 0 < f(-2) = -2^5 + 5 \cdot 2^5 + c = 2^7 + c \\ 0 > f(2) = 2^5 - 5 \cdot 2^5 + c = -2^7 + c \\ f(x) = x^5 - 5 \cdot 16x + c \text{ is irreducible over } \mathbb{Q}. \end{gathered}$$

The first two constraints reduce to $-2^7 < c < 2^7$. This is a comfortably sized range. With Eisenstein's Criteria in mind, pick $c = 5$ (and there are other obvious choices). We find

$$f(x) = x^5 - 80x + 5$$

is a polynomial of degree 5 that cannot be solved by radicals.

What about the possibility that the Galois group of our polynomial is only a subgroup of $\mathbf{S}_n$? To put it another way, what can we say about the solvable subgroups of $\mathbf{S}_n$ that are Galois groups? One thing that we can recall from Kronecker is that given any two roots of an irreducible polynomial, there will be an automorphism in the Galois group that takes one root to another. We could frame this property for any subgroup of $\mathbf{S}_n$. We will say that a subgroup $\mathbf{G}$ of $\mathbf{S}_n$ is **transitive** provided for any $i, j \in \{0, 1, 2, \ldots, n-1\}$

there is some $\sigma \in G$ so that $\sigma(i) = j$. Then, according to Kronecker, the Galois group of any separable irreducible polynomial of degree n will be (isomorphic to) a transitive subgroup of $\mathbf{S}_n$. What can we say about the transitive solvable subgroups of $\mathbf{S}_n$?

Theorem on the Transitive Solvable Subgroups of $\mathbf{S}_p$. *Let p be a prime number and let $\mathbf{G}$ be a transitive solvable subgroup of $\mathbf{S}_p$. Then every $\sigma \in G$, except the identity, has no more than one fixed point.*

The proof of this theorem relies on two lemmas that are of some interest in their own right.

Lemma on Normal Subgroups of Transitive Groups. *Let p be a prime number, let $\mathbf{G}$ be a transitive subgroup of $\mathbf{S}_p$, and let $\mathbf{N}$ be a nontrivial normal subgroup of $\mathbf{G}$. Then $\mathbf{N}$ is transitive.*

Proof. The group $\mathbf{N}$ induces a partition of $\{0,1,\dots,p-1\}$ into orbits. I contend that all the orbits have the same size. To see this let $\mathcal{O}$ and $\mathcal{Q}$ be any two orbits. Pick elements $a \in \mathcal{O}$ and $b \in \mathcal{Q}$ and, since $\mathbf{G}$ is transitive, pick $\beta \in G$ so that $\beta(a) = b$. Let $c \in \mathcal{O}$. Pick $\sigma \in N$ so that $\sigma(a) = c$. Then observe that

$$\beta(c) = \beta(\sigma(a)) = \beta(\sigma(\beta^{-1}(b))) = (\beta\circ\sigma\circ\beta^{-1})(b).$$

Since $\mathbf{N}$ is a normal subgroup of $\mathbf{G}$, we see that $\beta(c) \in \mathcal{Q}$. So β induces a map from $\mathcal{O}$ into $\mathcal{Q}$. A similar argument shows that β^{-1} induces a map from $\mathcal{Q}$ into $\mathcal{O}$. These induced maps invert each other, so the two orbits are the same size.

Since $\mathbf{N}$ is nontrivial, there must be a nontrivial orbit. So all orbits have the same size $k > 1$. But our set with p elements is partitioned into sets of size k. So $k \mid p$. Since p is prime, we have $k = p$. That is, there is only one orbit. This means $\mathbf{N}$ is transitive. □

Lemma on p-cycles in Solvable Transitive Subgroups of $\mathbf{S}_p$. *Let p be a prime number and $\mathbf{G}$ be a nontrivial transitive solvable subgroup of $\mathbf{S}_p$. The last nontrivial group in any composition series of $\mathbf{G}$ is a cyclic group of order p and every p-cycle in G belongs to this cyclic group.*

Proof. Let $\mathbf{G}$ be a solvable transitive subgroup of $\mathbf{S}_p$, where p is prime. Consider a composition series for $\mathbf{G}$.

$$\mathbf{G} = \mathbf{G}_0 \rhd \mathbf{G}_1 \rhd \cdots \rhd \mathbf{G}_{n-1} \rhd \mathbf{G}_n$$

so that $\mathbf{G}_n$ is trivial, $\mathbf{G}_{n-1}$ is not trivial, and $\mathbf{G}_k/\mathbf{G}_{k+1}$ is of prime order for all $k < n$. By the lemma above, $\mathbf{G}_k$ is transitive for each $k < n$. Notice that $\mathbf{G}_{n-1}$ is a cyclic group of prime order q. Let σ generate this group. Write the permutation σ as a product of disjoint cycles. These cycles must all have length q since q is prime. Moreover, every power of σ is the product of the powers of these disjoint cycles. So if there were more than one cycle in the decomposition of σ we would have that $\mathbf{G}_{n-1}$ could not be transitive. So there is only one cycle and it is of length p. So $p = q$ and we infer, with Lagrange's help, that $p \mid |G_k|$ for all $k < n$.

Now let τ be any p-cycle that belongs to G. Since $\mathbf{G}_n$ is trivial, we see that $\tau \notin G_n$. Pick k as large as possible so that $\tau \in G_k$. So $\tau \notin G_{k+1}$. Let $\mathbf{H}$ be the subgroup of $\mathbf{G}_k$ generated by τ. Then $|H| = p$. Every element of H other than the identity generates $\mathbf{H}$. So $H \cap G_{k+1}$ can contain only the identity element since $\tau \notin G_{k+1}$. Since $\mathbf{G}_k \rhd \mathbf{G}_{k+1}$ we see that $\mathbf{HG}_{k+1}$ is a subgroup of $\mathbf{G}_k$. We also know

$$|HG_{k+1}||H \cap G_{k+1}| = |H||G_{k+1}|.$$

So we find that $|HG_{k+1}| = p|G_{k+1}|$. Now $\mathbf{HG}_{k+1}$ is a subgroup of $\mathbf{S}_p$ and this last group has cardinality $p!$. So p^2 cannot divide the order of $\mathbf{HG}_{k+1}$. But this means that p cannot divide $|G_{k+1}|$. This forces $k+1 = n$. Therefore, $\tau \in G_{n-1}$, as desired. □

Proof of the Theorem on the Transitive Solvable Subgroups of $\mathbf{S}_p$.

Let $\tau \in G$ and suppose that $a, b \in \{0, 1, \ldots, p-1\}$ with

$$\tau(a) = a$$
$$\tau(b) = b$$

Our aim is to show that $a = b$ or that τ is the identity. Let σ be a p-cycle that generates $\mathbf{G}_{n-1}$. Now $\tau\sigma\tau^{-1} \in G$ and it must also be a p-cycle (since it has order p and p is prime). So $\tau\sigma\tau^{-1} \in G_{n-1}$ by the last lemma. With this in view, pick a positive $k < p$ so that

$$\tau\sigma\tau^{-1} = \sigma^k.$$

Since $\mathbf{G}_{n-1}$ is transitive, we can pick a natural number $\ell < p$ so that $\sigma^\ell(a) = b$. From the displayed equation we see that $\tau\sigma(a) = \sigma^k\tau(a) = \sigma^k(a)$. An easy induction reveals that $\tau\sigma^j(a) = \sigma^{jk}(a)$ for every natural number j. In particular,

$$\sigma^\ell(a) = b = \tau(b) = \tau\sigma^\ell(a) = \sigma^{\ell k}(a).$$

So we find $a = \sigma^{\ell k - \ell}(a)$. But the only permutation in G_{n-1} with a fixed point is the identity permutation. This means $p \mid \ell(k-1)$. Since p is prime and $0 \leq \ell, k < p$ and k is positive, we conclude that either $\ell = 0$ or $k = 1$. In the first alternative we have $b = \sigma^\ell(a) = \sigma^0(a) = a$, while in the second alternative we have that σ and τ commute. In that case,

$$\begin{aligned}
\tau(\sigma(a)) &= \sigma(\tau(a)) = \sigma(a) \\
\tau(\sigma^2(a)) &= \sigma^2(\tau(a)) = \sigma^2(a) \\
&\vdots \\
\tau(\sigma^j(a)) &= \sigma^j(a) \\
&\vdots
\end{aligned}$$

In this way we see that τ must fix each of the elements of $\{0, 1, \ldots, p-1\}$. That is τ is the identity. □

As a corollary we arrive at the following result.

Artin's Criteria for Unsolvability of Polynomials over $\mathbb{Q}$. *Every irreducible polynomial of prime degree with coefficients in $\mathbb{Q}$ that has at least two real roots and at least one nonreal root in $\mathbb{C}$ is not solvable by radicals.*

Proof. Let $\mathbf{E}$ be a subfield of $\mathbb{C}$ that is a splitting field of our polynomial. Since the coefficients of our polynomial are real (even rational) we see that complex conjugation, restricted to E, is a member of the Galois group. Because the polynomial has a root that is not real, we see that the restriction of complex conjugation is not merely the identity map. On the other hand, complex conjugation fixes two of the roots (since two of them are real). So the Galois group of the polynomial cannot be solvable, by the Theorem on Transitive Solvable Subgroups of $\mathbf{S}_p$. So by Galois' characterization, our polynomial is not solvable by radicals. □

Notice that the hypotheses of this theorem are weaker than those laid out in the $\mathbf{S}_p$ Criterion over $\mathbb{Q}$. On the other hand, the conclusion is also weaker: that the Galois group is not solvable rather than that the Galois group is actually $\mathbf{S}_p$.

In applying these new criteria it is enough to show that the graph of $f(x)$ on the $X \times Y$ plane crosses the X-axis at least twice, but not p times (provided $f(x)$ is irreducible of prime degree p).

Let us devise an example to which Artin's Criterion applies, but not the earlier criterion. We will find an irreducible polynomial $f(x)$ of degree 7 with 3 real roots and 4 nonreal roots. The graph of such a polynomial will cross the X-axis 3-times. One way to achieve this is to make sure the leading coefficient is positive and that the graph has one local maximum (where the function is positive) and one local minimum (where the function is negative). We hope to use Eisenstein to ensure that $f(x)$ is irreducible, so our polynomial will have integer coefficients. Given the curve-sketching nature of this idea, we will first create a suitable derivative $f'(x)$. This must have degree 6. Here is one to start with:

$$(x-3)(x+3)(x^2+3)(x^2+9) = x^6+3x^4-3^4x^2-3^5.$$

I have used a lot of 3's in the hope that this will make the eventual use of Eisenstein easier. Were this the derivative of our polynomial, we would know that the graph is increasing on $(-\infty,-3)$, that it is decreasing on $(-3,3)$ and that it is increasing again on $(3,\infty)$. The next step would be to integrate this polynomial, but that would introduce some fractional coefficients. To ease this, why not multiply the thing by $7\cdot 5$? So take

$$f'(x) = 7\cdot 5x^6+7\cdot 5\cdot 3x^4-7\cdot 5\cdot 3^4x^2-7\cdot 5\cdot 3^5.$$

Integrating gets us

$$f(x) = 5x^7+7\cdot 3x^5-7\cdot 5\cdot 3^3x^3-7\cdot 5\cdot 3^5x+c$$

where c is the constant of integration. I hope to chose c so that the local maximum (it is at $x=-3$) is positive, that the local minimum (it is at $x=3$) is negative, and finally, so that Eisenstein will tell us that $f(x)$ is irreducible. Sheer computation shows

$$f(-3) = 3^6\cdot 48+c \qquad \text{and } f(3) = -3^6\cdot 48+c.$$

Given the desire for $f(-3)>0$ and $f(3)<0$, it turns out that c must be selected so that

$$-3^6\cdot 48 < c < 3^6\cdot 48.$$

This is a very commodious range of choices. I take $c=3$ (but you might like 7 or even 21 better). This choice gives

$$f(x) = 5x^7+7\cdot 3x^5-7\cdot 5\cdot 3^3x^3-7\cdot 5\cdot 3^5x+3.$$

So Eisenstein applies with 3 as the chosen prime. We could make this more mysterious by saying

$$f(x) = 5x^7+21x^5-945x^3-8,435x+3.$$

This polynomial is not solvable by radicals.

The **Inverse Galois Problem** reverses the concern we have just been pursuing: start with a finite group **G** and ask whether it is (isomorphic to) a Galois group. Framed this broadly, the answer is always yes. But tighten the problem some—for instance ask whether **G** is a Galois group over the rationals—and you arrive at a problem that is still open here in 2015. It may well be that every finite group is a Galois group over the rationals. This is one of the great open problems in mathematics.

26.1 Problem Set 22

Algebra Homework, Edition 22
Tweny Third Week
Solvability by radicals and other things Galois may have known

Problem 109.
Let p be prime and let $\mathbf{H}$ be a subgroup of $\mathbf{S}_p$. Prove that if $\mathbf{H}$ has a transposition and an element of order p, then $\mathbf{H} = \mathbf{S}_p$. Provide an explicit counterexample when p is not prime.

Problem 110.
Prove that $x^5 - 2x^3 - 8x + 2$ is not solvable by radicals over the field $\mathbb{Q}$ of rational numbers.

Problem 111.
Let $\mathbf{F}$ be a finite field. Prove that the product of all the nonzero elements of $\mathbf{F}$ is -1. Using this, prove Wilson's Theorem:

$$(p-1)! \equiv -1 \pmod{p}$$

for every prime number p.

Problem 112.
Let $\mathbf{E}$ be the splitting field of $x^5 - 2$ over the field $\mathbb{Q}$ of rationals. Find the lattice (draw a picture) of all fields intermediate between $\mathbb{Q}$ and $\mathbf{E}$.

Problem 113.
Let $\mathbf{F}$ be a field of characteristic p, where p is a prime. Let $\mathbf{E}$ be a field extending F. Prove that E is a normal separable extension of $\mathbf{F}$ of dimension p if and only if $\mathbf{E}$ is the splitting field over $\mathbf{F}$ of an irreducible polynomial of the form $x^p - x - a$, for some $a \in F$.

Problem 114.
Let p be a prime number. Is $x^5 - 5px - p$ solvable by radicals over $\mathbb{Q}$?

Problem 115.
Prove that every polynomial with rational coefficients whose splitting field over $\mathbb{Q}$ has dimension 1225 is solvable by radicals.

LECTURE 27

Algebraic Closures of Real-Closed Fields

Here we want to obtain the result that the field of complex numbers is the algebraic closure of the field of real numbers. This assertion, traditionally called the Fundamental Theorem of Algebra, has a storied past and many proofs—indeed there are whole monographs devoted to the exposition of an array of proofs of this theorem. Many point of the doctoral dissertation of Gauss for the first fully correct proof. Sadly, even the proof in Gauss's dissertation also has a gap—let the dedicated graduate students take note!

The shortest proofs come by way of complex analysis: were $f(z)$ a rootless polynomial of positive degree then $\frac{1}{f(z)}$ would be analytic on the whole complex plane (i.e. it is holomorphic) and a simple argument shows it is bounded. So Liouville tells us that it must be constant—an impossibility. Of course, developing complex analysis to the point of Liouville's Theorem (or any of a number of other theorems of complex analysis that would serve) is not entirely immediate. These proofs have the added feature that they apply to complex functions other than polynomials of positive degree.

The approach we will take uses the apparatus of Galois theory, and has the advantage that it applies to fields other than the complex numbers.

The field of real numbers has the following three properties:

(a) Every polynomial of odd degree has a root.

(b) There is a set P of elements with the following properties:

   (i) $0 \notin P$.

   (ii) If $a, b \in P$, then $a+b, ab \in P$.

   (iii) For every nonzero element a of the field exactly one of $a \in P$ or $-a \in P$ holds.

(c) Every element of P has a square root in the field.

You should note that all these properties have an algebraic flavor. Property (a) follows by the familiar Intermediate Value Theorem of freshman calculus. The set P is just the set of positive reals. Property (c) again follows by the Intermediate Value Theorem.

Any field that has properties (a), (b), and (c) is called a **real closed field**. Of course, $\mathbb{R}$ is a real closed field, but there are other real closed fields, even ones that are countable. Indeed, the curious graduate student might want to start with the field $\mathbb{Q}$ of rational numbers and begin adding elements (of the reals) to obtain in the most parsimonious way a real closed field.

The properties stipulated in (b) allow us to define a linear ordering of the field:

$$a < b \overset{\text{def}}{\Longleftrightarrow} b - a \in P.$$

The demonstration that this defines a linear order that has the expected properties with respect to + and · is left in the hands of the graduate students. We could, of course, reverse this process: start with a well-behaved linear order and take $P = \{a \mid a > 0\}$ and show that P has the attributes given in (b) and (c).

To get a better grip on this notion, the eager graduate students should try proving that in a real closed field the square of any nonzero element is positive, that 1 is positive, and that the field must have characteristic 0.

The notion of a real closed field was propounded by Emil Artin around 1924 as a means to bring algebraic methods into play in what had been a largely analytic enterprise: the investigation of the real and complex numbers. The theorems here are taken mostly from two papers of Emil Artin and Otto Schreier which appeared in 1926 and 1927. Artin's famous solution to Hilbert's Seventeenth Problem, published also in 1927, was based on theory developed by Artin and Schreier in these two papers.

The proof I give below is the work of Artin and Schreier and uses Galois Theory and Sylow's Theorem. Artin and Schreier also provided a second argument that lifts a 1795 proof of Laplace of the Fundamental Theorem of Algebra to the case of real closed fields. Laplace's proof depended on Kronecker's Theorem, which was unknown at the time. In 1816 Gauss published a proof that filled this gap in Laplace's proof by an analysis of symmetric polynomials, circumventing the still unknown result of Kronecker.

The Artin-Schreier Fundamental Theorem of Algebra for Real Closed Fields. *If* $\mathbf{R}$ *is a real closed field, then* $\mathbf{R}\left[\sqrt{-1}\,\right]$ *is algebraically closed.*

Proof. First notice that $\left(\sqrt{-1}\,\right)^2 = -1$ and -1 is not positive. This means $\sqrt{-1} \notin R$. So x^2+1 is irreducible over $\mathbf{R}$ and $\mathbf{R}\left[\sqrt{-1}\,\right]$ is the splitting field of x^2+1. Let $\mathbf{C} = \mathbf{R}\left[\sqrt{-1}\,\right]$. According to Kronecker $[\mathbf{C} : \mathbf{R}] = 2$. Of course the members of C have the form

$$a + b\sqrt{-1}$$

where $a, b \in R$. Conjugation has its usual definition and it is an automorphism of $\mathbf{C}$ that fixes each element of R.

Now let $f(x) \in \mathbf{C}[x]$. By $\bar{f}(x)$ we mean the polynomial obtained from $f(x)$ by applying conjugation to each of the coefficients. Then $f(x)\bar{f}(x) \in R[x]$ follows easily from the description of the coefficients of the product of two polynomials together with the fact that conjugation is an automorphism of $\mathbf{C}$.

Observe that $f(x)$ has a root in C if and only if $f(x)\bar{f}(x)$ has a root in C. So it is enough for us to prove that every monic polynomial in $\mathbf{R}[x]$ of positive degree has a root in C. We already know this for polynomials of odd degree—they even have roots in R.

We use the following fact.

Contention. Every element of C has a square root in C.

Proof. Let $a + b\sqrt{-1}$ be an arbitrary element of C. In case $b = 0$, this element will belong to R and, since every positive element of R already has a square root in R, it is easy to see that every element of R has a square root in C. So we consider that $b \neq 0$. Then direct computation shows that $c + d\sqrt{-1}$ is a square root of $a + b\sqrt{-1}$, where

$$c = \frac{b}{2d} \text{ and } d^2 = \frac{-a + \sqrt{a^2+b^2}}{2}.$$

Notice that $\frac{-a+\sqrt{a^2+b^2}}{2}$ is a positive member of R. These constraints came about by twiddling with the quadratic formula. □

This entails that there is no extension $\mathbf{E}$ of $\mathbf{C}$ with $[\mathbf{E}:\mathbf{C}] = 2$ since every polynomial of degree 2 in $\mathbf{C}[x]$ is reducible using the quadratic formula.

Now let $f(x) \in \mathbf{R}[x]$ be a monic polynomial. Let $\mathbf{E}$ be the splitting field of $f(x)(x^2+1)$ over $\mathbf{R}$. We can suppose that $\mathbf{E}$ extends $\mathbf{C}$. Since our characteristic is 0, we know that $\mathbf{E}$ is a Galois extension of $\mathbf{R}$. Pick natural numbers ℓ and m, with m odd, so that

$$|\operatorname{Gal}(\mathbf{E}/\mathbf{R})| = 2^\ell m.$$

By Sylow, $\operatorname{Gal}(\mathbf{E}/\mathbf{R})$ has a subgroup $\mathbf{H}$ with $|H| = 2^\ell$. Let $\mathbf{K} = \operatorname{Inv}\mathbf{H}$. Then $[\mathbf{E}:\mathbf{K}] = 2^\ell$ and $[\mathbf{K}:\mathbf{R}] = m$, by the Fundamental Theorem of Galois Theory. Since $\mathbf{R}$ has no proper extension of odd dimension (every polynomial of odd degree has a root—so you get a grumble out of Kronecker and the Dimension Formula), we must have $m = 1$ and $\mathbf{K} = \mathbf{R}$. But then $[\mathbf{E}:\mathbf{R}] = 2^\ell$. But recall that $\mathbf{E}$ extends $\mathbf{C}$. So

$$2^\ell = [\mathbf{E}:\mathbf{R}] = [\mathbf{E}:\mathbf{C}][\mathbf{C}:\mathbf{R}] = [\mathbf{E}:\mathbf{C}]2.$$

In this way we find $[\mathbf{E}:\mathbf{C}] = 2^{\ell-1}$. If $\ell = 1$ then we find $\mathbf{E} = \mathbf{C}$, and we have reached the conclusion we desire. On the other hand, if $\ell > 1$, we see that $\operatorname{Gal}\mathbf{E}/\mathbf{C}$ is a group of cardinality $2^{\ell-1}$. By Sylow, there is a subgroup $\mathbf{N}$ of this Galois group so that $|N| = 2^{\ell-2}$. So $[\operatorname{Gal}\mathbf{E}/\mathbf{C}:\mathbf{N}] = 2$. Now every subgroup of index 2 must be a normal subgroup. The fixed field of $\mathbf{N}$ must be a (normal) extension of $\mathbf{C}$ with dimension 2. But we know that $\mathbf{C}$ has no extensions of dimension 2. So we reject the possibility that $\ell > 1$.

This means every polynomial over $\mathbf{R}$ of positive degree has a root in C. So our proof is complete.

The use of Sylow's Theorem (unknown until the late 1800's) above and of the Fundamental Theorem of Galois Theory to produce the fixed field of $\mathbf{N}$ can be avoided by following the line of reasoning proposed by Laplace in 1795. Here is how.

We still want to show that every polynomial $f(x)$ of positive degree with coefficients in R has a root in C. Let ℓ by the natural number so that the degree of $f(x)$ is $n = 2^\ell m$ where m is odd. Call this number ℓ the *2-index* of $f(x)$. Our proof is by induction on the 2-index. In the base step, $f(x)$ is a polynomial of odd degree, so it even has a root in R. For the inductive step, suppose $f(x)$ has 2-index $k+1$. Let $r_0, \dots, r_{n-1}$ be the roots of $f(x)$ in $\mathbf{E}$. For each $a \in R$ define

$$g_a(x) = \prod_{i<j<n} \big(x - (r_i + r_j + ar_ir_j)\big)$$

Notice that the degree of $g_a(x)$ is $\binom{n}{2}$. But

$$\binom{n}{2} = \frac{n(n-1)}{2} = \frac{1}{2}2^{k+1}m(2^{k+1}m-1) = 2^k m(2^{k+1}m-1)$$

so that the 2-index of each $g_a(x)$ is k. But observe that the coefficients of $g_a(x)$ must be fixed by every automorphism in $\operatorname{Gal}(\mathbf{E}/\mathbf{R})$. So each $g_a(x) \in \mathbf{R}[x]$. (This uses a little bit of Galois theory to say that the fixed field of $\mathbf{E}$ is actually $\mathbf{R}$.) So for each $a \in R$ we see by the induction hypothesis that

$$r_i + r_j + ar_ir_j \in C$$

for some choice of the natural numbers i and j with $i < j < n$. Now there are only finitely many ways to pick such i and j but infinitely choices for a. As every pigeon knows, we must have two distinct member of R, say a and b, so that for some choice of i and j

$$r_i + r_j + ar_ir_j \qquad \text{and} \qquad r_i + r_j + br_ir_j \qquad \text{both belong to } C.$$

Subtracting these and dividing away the nonzero $b - a$, we find first that r_ir_j and then $r_i + r_j$ also belong to C. But everyone can see that

$$(x - r_i)(x - r_j) = x^2 - (r_i + r_j)x + r_ir_j,$$

which is a polynomial of degree 2 with coefficients in C. But all polynomials in $\mathbf{C}[x]$ of degree 2 have roots in C. So $r_i \in C$ and it is a root of $f(x)$.

Laplace still needed the (much delayed) aide of Kronecker to obtain the splitting field **E**, but the little bit of Galois Theory used here can be finessed. □

Artin and Schreier also proved the converse.

Artin and Schreier's Characterization of Real Closed Fields. *A field* $\mathbf{R}$ *is a real closed field if and only if* x^2+1 *has no root in* $\mathbf{R}$ *and* $\mathbf{R}[\sqrt{-1}]$ *is algebraically closed.*

Proof. We only have to prove one direction. So suppose x^2+1 has no root in $\mathbf{R}$ and that $\mathbf{R}[\sqrt{-1}]$ is algebraically closed. First, observe that if $a,b \in R$ then there is some $c \in R$ so that $c^2 = a^2+b^2$. This follows since the analog of complex conjugation in $\mathbf{R}[\sqrt{-1}]$ is an automorphism whose set of fixed points is just R (an entertainment for graduate students!). Now using the algebraic closedness, pick $u \in \mathbf{R}[\sqrt{-1}]$ with $u^2 = a+bi$. Then

$$a^2+b^2 = (a+bi)(a-bi) = (a+bi)\overline{(a+bi)} = u^2\overline{u^2} = (u\overline{u})^2.$$

But $u\bar{u} \in R$ since it is fixed by this analog of complex conjugation. So take $c = u\bar{u}$. So we see that in $\mathbf{R}$ the sum of two squares is again a square. It follows that the sum of any finite number of squares in again a square. Now -1 cannot be a square in $\mathbf{R}$ since x^2+1 has no root in R. This also means that 0 cannot be the sum of a finite number of nonzero squares. Let us take P to be the set of all those members of R that can be written as a sum of nonzero squares, which is the same as the set of those members of R that are themselves nonzero squares. In the definition of real closed fields there are four stipulations our set P must satisfy. They are all easy (aren't they?).

So it only remains to show that every polynomial in $\mathbf{R}[x]$ of odd degree has a root in R. Now every polynomial of odd degree must have an irreducible factor of odd degree. Such an irreducible polynomial must have a root r in $\mathbf{R}[\sqrt{-1}]$ since that field is algebraically closed. But this is a field of dimension 2 over $\mathbf{R}$. Consider the Dimension Formula and Kronecker's Theorem. The degree of our irreducible polynomial must divide 2. The only odd number that divides 2 is 1. So our irreducible polynomial has degree 1. That means it has a root in R.

In this way, we see that $\mathbf{R}$ is a real closed field. □

Actually, Artin and Schreier went on to prove that if $\mathbf{R}$ is any field so that $[\mathbf{A} : \mathbf{R}]$ is finite, where $\mathbf{A}$ is algebraically closed, then $\mathbf{R}$ is a real closed field. This is an intriguing result: given a field $\mathbf{F}$ and its algebraic closure $\mathbf{A}$ there are only three possibilities: $\mathbf{A}$ is infinite dimensional over $\mathbf{F}$ or the dimension is just 2 (and $\mathbf{F}$ is a real closed field) or the dimension is just 1 (the field $\mathbf{F}$ is algebraically closed already). Why not look into this matter a bit on your own?

You can see from the proof of the Fundamental Theorem of Algebra for Real Closed Fields, that the properties of the set P were used only to establish that every polynomial in $\mathbf{C}[x]$ of degree 2 has a root in C. After some tampering, you can see that the following statement can also be proven:

> Every field of characteristic 0 in which every polynomial of degree 2 has a root and in which every polynomial of odd degree has a root, is algebraically closed.

In essence, this is the result of Artin and Schreier (and of Gauss) with the "real" part stripped out. In 2007, Joseph Shipman proved a wide ranging extension of this result, namely:

> Every field in which every polynomial of prime degree has a root is algebraically closed.

27.1 PROBLEM SET 23

ALGEBRA HOMEWORK, EDITON 23
TWENTY FOURTH WEEK
INTERMEDIATE FIELDS AND REAL CLOSED FIELDS

PROBLEM 116.
Let **E** be the splitting field over $\mathbb{Q}$ of $x^4 - 2$. Determine all the fields intermediate between **E** and $\mathbb{Q}$. Draw a diagram of the lattice of intermediate fields.

PROBLEM 117.
Prove that the field of real numbers has only one ordering that makes it into an ordered field. In contrast, prove that $\mathbb{Q}[\sqrt{2}]$ has exactly two such orderings.

PROBLEM 118.
Let **R** be a real closed field and let $f(x) \in \mathbf{R}[x]$. Suppose that $a < b$ in **R** and that $f(a)f(b) < 0$. Prove that there is some $c \in R$ with $a < c < b$ such that c is a root of $f(x)$.

PROBLEM 119.
Let **R** be a real closed field, let $a_0 + a_1x + \cdots + a_{n-1}x^{n-1} + x^n = f(x) \in R[x]$, and put $M = |a_0| + |a_1| + \cdots + |a_{n-1}| + 1$. Prove that every root of $f(x)$ which belongs to R belongs to the interval $[-M, M]$.

PROBLEM 120.
Prove that every element of a finite field can be written as the sum of two squares.

PROBLEM 121.
Let **F** be a field. Prove that the following are equivalent.

(a) **F** is not algebraically closed but there is a finite upper bound on the degrees of the irreducible polynomials in $\mathbf{F}[x]$.

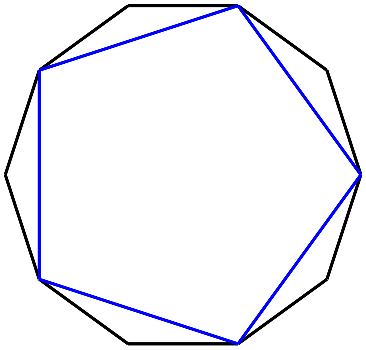
(b) **F** is a real closed field.

LECTURE 28

Gauss on Constructing Regular Polygons by Straightedge and Compass

There were two compass-and-straightedge problems from the ancient Greeks left unresolved earlier in our development: squaring the circle and determining which regular polygons are constructible. Squaring the circle is handled in a different lecture where it is shown that π is not an algebraic number, and so the number $\sqrt{\pi}$, which is the length of side of a square whose area is the same as the unit circle, is not constructible.

Here we take up the construction of regular polygons.



The drawing above shows how a regular pentagon and a regular decagon are related. Recalling that it is easy to bisect line segments, it is clear that the regular pentagon is constructible if and only if the regular decagon is constructible. Here there is nothing special about the numbers 5 and 10. This idea holds for n and $2n$ where n is any natural number larger than 2. Noting that squares are constructible, this reduces the problem of determining which regular polygons are constructible to the constructibility of those with an odd number of sides.

Euclid knew how to construct regular polygons with $3, 5$, and 15 sides. But the ancient geometers had no constructions for $7, 9, 11$, or 13. It turns out that none of those four regular polygons are constructible. When Gauss was 19 in 1796, he discovered construction for the case of 17—for the regular heptadecagon.

We carried out our earlier development of constructible numbers before we developed Galois theory. We showed there that a (real) number r was constructible if and only if it is captured in some square root tower over $\mathbb{Q}$. We can now do a little better.

A Characterization of the Field of Constructible Reals. *Let r be a real number. The number r is constructible if and only if r is algebraic over $\mathbb{Q}$ and $[\mathbf{K}:\mathbb{Q}]$ is a power of 2, where $\mathbf{K}$ is the splitting field of the minimal polynomial of r.*

Proof. First suppose that r is constructible. By the Basic Theorem for the Field of Constructible Reals, which is our earlier result, we know that $\mathbb{Q}[r]$ is engulfed in a square root tower over $\mathbb{Q}$. Our difficulty is that $\mathbf{K}$ is probably larger than $\mathbb{Q}[r]$. We apply the same devise to the square root tower that we applied to the radical tower to get Galois' Criterion for Solvability by Radicals. In particular, Lemma 25.0.3 works here—just note that the prime $p = 2$ in this case. Then $\mathbf{K}$ will be engulfed by this enhanced square root tower.

For the converse, we see that $\mathbf{K}$ is a Galois extension of $\mathbb{Q}$. By the Fundamental Theorem of Galois Theory, we see that the Galois group has order 2^r for some positive natural number r. So the Galois group is a p-group (with $p = 2$) and that makes it solvable. A composition series for the Galois group gives us a normal series with quotient of size 2. By way of the Fundamental Theorem of Galois Theory this normal series is associate with a series of intermediate fields

$$\mathbb{Q} = \mathbf{L}_0 \subseteq \mathbf{L}_1 \subseteq \cdots \subseteq \mathbf{L}_r = \mathbf{K}$$

where $\mathbf{L}_{k+1}$ is a normal extension of $\mathbf{L}_k$ and $[\mathbf{L}_{k+1}:\mathbf{L}_k] = 2$, for each $k < r$. We would be finished if we could see that this tower of fields is a square root tower. So pick $k < r$ and let $u \in L_{k+1}$ with $u \notin L_k$. Since there are no integers between 1 and 2, we see that $\mathbf{L}_{k+1} = \mathbf{L}_k[u]$. So the minimal polynomial of u over $\mathbf{L}_k$ has the form $x^2 + 2bx + c$. Now just complete the square! $x^2 + 2bx + c = (x+b)^2 + c - b^2$. Let $v = u + b$. Then $\mathbf{L}_k[u] = \mathbf{L}_k[v]$ and v is a root of $x^2 + (c - b^2)$. This means that $\mathbf{L}_{k+1}$ is obtained from $\mathbf{L}_k$ by adding a square root of some element of L_k. □

We could have easily reframed this characterization in terms of constructible complex numbers. A regular n-sided polygon is constructible if and only if a primitive n^{th} root of unity is constructible. Considering n to be odd, we could first factor n into a product of (odd) primes. Then we could tackle the question of which regular p-sided polygons are constructible for various odd primes p. Gauss discovered that only certain kinds of odd primes work out, namely the Fermat primes. These are primes that are natural numbers of the form

$$2^{2^k} + 1.$$

In the 1650's Fermat conjectured that each number of this form is prime. However, Leonard Euler refuted this conjecture in 1732 by showing

$$2^{32} + 1 = 641 \cdot 6700417.$$

Eisenstein raised the problem of showing that the number of Fermat primes is infinite. The only Fermat primes known in 2015 were the ones known to Fermat, namely the first 5:

$$3, 5, 17, 257, \text{ and } 65,537.$$

Fact. Let m be a natural number. If $2^m + 1$ is a prime number, then $m = 2^n$ for some natural number n.

Proof. Let us suppose, to the contrary, that b is an odd prime that divides m. Pick the natural number c so that $m = bc$. Now since b is odd, it is easy to check that

$$x^b + 1 = (x+1)(x^{b-1} - x^{b-2} + \cdots - x^1 + 1).$$

Plugging 2^c in for x we find

$$2^m + 1 = (2^c)^b + 1 = (2^c + 1)(2^{c(b-1)} - \cdots + 1).$$

In this way we see that $2^m + 1$ is not prime since it is divisible by $2^c + 1$. □

So one could say, more easily, that a prime number p is a Fermat prime provided there is a natural number m so that $p = 2^m + 1$.

Gauss's Characterization of Constructible Regular Polygons. *Let n be a natural number larger than 2. A regular n-sided polygon is constructible if and only if n factors as $2^e p_0 p_1 \dots p_{r-1}$ where e is a natural number and $p_0, p_1, \dots, p_{r-1}$ are distinct Fermat primes.*

To prove this theorem, we need to know more about primitive roots of unity and their minimal polynomials. Fix n. We know the n^{th} roots of unity constitute a subgroup of the multiplicative group $\mathbb{C}^\times$ of the field of complex numbers. This subgroup has n elements, a finite number, and is therefore cyclic. So this group is isomorphic with the group $\langle \mathbb{Z}_n, +, -, 0\rangle$. The single elements that can generate this last group we know to be those natural numbers less than n that are relatively prime to n. (Remember how nice cyclic groups are?) So the number of primitive n^{th} roots of unity is $\varphi(n)$, Euler's totient function: the number of natural numbers less than n that are relatively prime to n. So consider the function

$$\lambda_n(x) := \prod_{\zeta \text{ is primitive}} (x - \zeta)$$

where the product is over all the primitive n^{th} roots of unity. Let $\mathbf{E}$ be the splitting field of $x^n - 1$ over the rationals. Of course $\mathbf{E} = \mathbb{Q}[\zeta]$, where ζ is any primitive root of unity. Any automorphism in the Galois group of $\mathbf{E}$ over $\mathbb{Q}$ must take one primitve n^{th} root of unity to another. So such an automorphism must fix each coefficient of $\lambda_n(x)$. Since the fixed field of this Galois group is $\mathbb{Q}$, we see that $\lambda_n(x)$ has rational coefficients. The polynomials of the form $\lambda_n(x)$ are called cyclotomic polynomials. They are monic.

Now suppose that ξ is an n^{th} root of unity. The subgroup of the group of all n^{th} roots of unity that is generated by ξ must have an order that divides n, according to Lagrange. Say that order is d. Then ξ is a primitive d^{th} root of unity and $d \mid n$. From this we see

$$x^n - 1 = \prod_{d|n} \lambda_d(x) = \lambda_n(x) \prod_{d|n,\, d<n} \lambda_d(x).$$

The diligent graduate student will spot here the key to showing by induction on n, that the cyclotomic polynomials actually have integer coefficients. The next theorem is due to Gauss.

Theorem on the Irreducibility of Cyclotomic Polynomials. *Every cyclotomic polynomial is irreducible over $\mathbb{Q}$.*

Proof. Suppose that $\lambda_n(x) = f(x)g(x)$ in $\mathbb{Z}[x]$, where $f(x)$ and $g(x)$ are monic and $f(x)$ is irreducible in $\mathbb{Z}[x]$ (and hence irreducible in $\mathbb{Q}[x]$). Let ξ be a root of $f(x)$ and suppose that p is any prime number that does not divide n. Then ξ^p is a primitive n^{th} root of unity (what else can its order be?). This means that ξ^p is a root of $\lambda_n(x)$. So there are two cases.

Case: ξ^p is a root of $f(x)$ for all primes p not dividing n and all roots ξ of $f(x)$.
In this case, it must be that if r is a natural number relatively prime to n, then ξ^r is a root of $f(x)$. Since every primitive n^{th} root of unity looks like ξ^r where r and n are relatively prime, we see that $(x - \xi)$ divides $f(x)$ for every primitive n^{th} root ξ of unity. But this means $\lambda_n(x) \mid f(x)$. Since $\lambda_n(x) = f(x)g(x)$ and both $f(x)$ and $\lambda_n(x)$ are monic, we get $\lambda_n(X) = f(x)$. In short, $\lambda_n(x)$ is irreducible.

Case: ξ^p is not a root of $f(x)$ for some prime p not dividing n and some root ξ of $f(x)$.
Well, we must reject this case.

Recalling that $\lambda_n(x) = f(x)g(x)$, we see that ξ^p is a root of $g(x)$. This means that ξ is a root of $g(x^p)$. Now $f(x)$ is the minimal polynomial of ξ so that $f(x) \mid g(x^p)$. Let us say that $g(x^p) = f(x)h(x)$. Recall that there is also some $k(x) \in \mathbb{Z}[x]$ so that $x^n - 1 = \lambda_n(x)k(x) = f(x)g(x)k(x)$.

We know the map $\rho \colon \mathbb{Z} \to \mathbb{Z}_p$ that sends each integer to its residue modulo p is a homomorphism. Using the map extension property for polynomial rings this map extends to a homomorphism from $\mathbb{Z}[x]$ to $\mathbb{Z}_p[x]$.

To conserve notation, let us use ρ for the extension as well. Now $\mathbb{Z}_p^\times$, the multiplicative group of nonzero elements of $\mathbb{Z}_p$ has $p-1$ elements. So $a^{p-1}=1$ for every nonzero element of $\mathbb{Z}_p$. This means that $a^p=a$ for every element of $\mathbb{Z}_p$. A bit of twiddling (do it, why not) shows $\rho(f(x))\rho(h(x)) = \rho(g(x^p) = (\rho(g(x)))^p$. So $\rho(f(x))$ and $\rho(g(x))$ must have an irreducible factor. This means that $\rho(x^n-1) = x^n - \rho(1)$ must have multiple roots. On the other hand, the derivative of $x^n-\rho(1)$ (in $\mathbb{Z}_p[x]$) is $\rho(n)x^{n-1}$ and $\rho(n)$ is not zero since p does not divide n. By our criterion for multiple roots, we see that $x^n-\rho(1)$ does not have multiple roots, a contradiction. So we must reject this case. □

Proof of Gauss's Characterization of Constructible Regular Polygons. We need only consider the case when n is odd. Let $n = p_1^{e_1}\dots p_{r-1}^{e_{r-1}}$, where $p_1,\dots,p_{r-1}$ are distinct odd primes and $e_1,\dots,e_{r-1}$ are positive natural numbers. Recalling that φ is Euler's totient function, we get

$$\varphi(n) = \varphi(p_1^{e_1})\cdots\varphi(p_{r-1}^{e_{r-1}}) = p_1^{e_1-1}(p_1-1)\cdots p_{r-1}^{e_{r-1}-1}(p_{r-1}-1).$$

Now $\varphi(n)$ is the degree of $\lambda_n(x)$. But now we know that $\lambda_n(x)$ is irreducible and so it must be the minimal polynomial of a primitive n^{th} root of unity. Hence $\varphi(n)$ is the dimension of its splitting field. So we know that a regular n-sided polygon is constructible if and only if $\varphi(n)$ is a power of 2. From the equation displayed above, we see this is equivalent to saying that $e_1=\cdots=e_{r-1}=1$ and each p_i-1 is a power of 2—that is each $p_i = 2^{m_i}+1$ for some natural number m_i. □

There is one point above that might have made you nervous. We seem to need that the Euler totient function φ respects multiplication. It really doesn't. But it is true that if n and m are relatively prime natural numbers then $\varphi(nm)=\varphi(n)\varphi(m)$. I leave the demonstration of this in the trustworthy hands of every graduate student who plans to pass those PhD exams.

It would be striking but it is conceivable that the five known Fermat primes exhaust the supply of Fermat primes. In that case, Gauss's theorem would tell us that there are only 32 odd numbers n so that the regular n-sided polygon is constructible by straight-edge and compass.

LECTURE 29

The Lindemann-Weierstrass Theorem on Transcendental Numbers

There are only countably many real numbers that are algebraic over the field of rational numbers. So almost every real number is transcendental. On the other hand, the real numbers that we can describe, in one manner or another, appear to be preponderantly algebraic. This is not too surprising, since even being generous with the notion of description, leads to only countably many descriptions (all of which can be typed in at a keyboard, say) and hence to only countably many describable reals. Within this smaller set it may be that algebraic numbers have a more substantial presence.

In 1844, Liouville cooked up the earliest examples of describable reals that are transcendental. These designer reals are likely interesting only for the property that they are transcendental. It wasn't until 1873 that Charles Hermite proved that e is transcendental. In 1882, Ferdinand Lindemann built on Hermite's methods to prove that π is transcendental. In 1885, Karl Weierstrass reframed Lindemann's proof to obtain the general result that is the topic of this lecture.

Lindemann's finding that π is transcendental settled the millennia-old problem of whether the circle could be squared using just straightedge and compass—that is whether a square could be constructed with the same area as the unit circle. Lindemann's theorem is one of the great victories of mathematics.

29.1 Formulation of the Lindemann-Weierstrass Theorem

To formulate the theorem, we need a new notion. Let $\mathbf{E}$ be a field extending $\mathbf{F}$. Let $u_0,\dots,u_{n-1} \in E$ be distinct. We say that these elements are **algebraically independent over F** provided that whenever $f(x_0,\dots,x_{n-1}) \in \mathbf{F}[x_0,\dots,x_{n-1}]$ such that $f(u_0,\dots,u_{n-1}) = 0$ then it must be that $f(x_0,\dots,x_{n-1})$ is the zero polynomial.

The Lindemann-Weierstrass Theorem. *If $u_0,\dots,u_{n-1}$ are distinct complex numbers that are algebraic over the rationals and are also linearly independent over the rationals, then the complex exponentials*

$$e^{u_0}, e^{u_1}, \dots, e^{u_{n-1}}$$

are algebraically independent over the field of complex numbers that are algebraic over the rationals.

Recall $e^{s+i\theta} = e^s(\cos\theta + i\sin\theta)$ for all real numbers s and θ. Or you can define the complex exponential function in a number of other ways familiar from complex analysis.

Before trying to prove anything like the Lindemann-Weierstrass Theorem, here are two corollaries.

Hermite says, "e is transcendental".

Proof. Let $u = 1$. It is immediate that 1 is algebraic and that $\{1\}$ is linearly independent over $\mathbb{Q}$. So the theorem says e^1 is algebraically independent over $\mathbb{Q}$. This means that e is not the root of any polynomial in $\mathbb{Q}[x]$ of positive degree. So e is transcendental. □

Lindemann says, "π is transcendental".

Proof. We all know that $e^{i\pi} = -1$. So $e^{i\pi}$ is algebraic. This means that $\{e^{i\pi}\}$ is not algebraically independent over $\mathbb{Q}$. By the theorem $i\pi$ cannot be algebraic since $\{i\pi\}$ is certainly linearly independent over $\mathbb{Q}$. But i is algebraic and we know that the product of algebraic numbers is algebraic, so π is not algebraic. That is, π is transcendental. □

Here is a closely related theorem.

The Lindemann-Weierstrass Theorem, Alternate Version. *If $u_0, \ldots, u_{n-1}$ are distinct algebraic numbers, then the complex exponentials $e^{u_0}, \ldots, e^{u_{n-1}}$ are linearly independent over the field of complex numbers that are algebraic over $\mathbb{Q}$.*

Proof of the Lindemann-Weierstrass Theorem from the Alternate Version. Let $u_0, \ldots, u_{n-1}$ be distinct complex algebraic numbers that are linearly independent over $\mathbb{Q}$. Let $\langle k_0, \ldots, k_{n-1}\rangle$ and $\langle \ell_0, \ldots, \ell_{n-1}\rangle$ be two different sequences of natural numbers. Then

$$\prod_{i<n} (e^{u_i})^{k_i} = e^{\sum_{i<n} k_i u_i} \text{ and } \prod_{i<n} (e^{u_i})^{\ell_i} = e^{\sum_{i<n} \ell_i u_i}.$$

Notice that $\sum_{i<n} k_i u_i \neq \sum_{i<n} \ell_i u_i$ by the linear independence of the u_i's. So the corresponding products are also different. Now suppose we are given m distinct sequences of natural numbers. This would result in m pairwise distinct products of the form above. By the Alternate Version, these products will be linearly independent over the algebraic numbers. But this is just another way of saying that the complex exponentials $e^{u_0}, \ldots, e^{u_{n-1}}$ are algebraically independent over the algebraic numbers. □

The graduate student in the mood for a challenge should figure out how to derive this alternate version from the Lindemann-Weierstrass Theorem itself. In any case, a proof of the Alternate Version, which omits mention of algebraic independence, would give us a proof of the Lindemann-Weierstrass Theorem. The Alternate Version is also easy to apply. The graduate students should try their hands at establishing the transcendence of e and π using this Alternate Version.

By way of preparation for proving the Alternate Version of the Lindemann-Weierstrass Theorem, we need to develop one more notion.

29.2 Algebraic Integers

Here is a question that may seem, at first glance, to be so obvious that it barely needs to be asked:

> "How can one pick out the subset of integers from the set of rationals?"

Of course, we saw how to *build* the rationals from the integers—we even know how to do this fraction field trick with any integral domain. However, the question above asks for a reversal of this trick. We could invent an infinite process that collects the integers: first throw in 0, then throw in 1 and -1, then $1+1$ and $(-1)+(-1), \ldots$. But is there another way, a finite elementary way?

We look at one ingenuous way here. Let **E** be a field and let **R** be a subring of **E**. We say that $u \in E$ is **integral** over **R** provided u is a root of a monic polynomial belonging to $\mathbf{R}[x]$. When **R** is actually a subfield of **E** the integral elements correspond to the elements that are algebraic over **R**. When $\mathbf{E} = \mathbb{C}$ and $\mathbf{R} = \mathbb{Z}$ we refer to the integral elements as **algebraic integers**.

Fact. A complex number u is an algebraic integer if and only if u is algebraic over $\mathbb{Q}$ and the minimal polynomial of u over $\mathbb{Q}$ actually belongs to $\mathbb{Z}[x]$.

Proof. Let $m(x)$ be the minimal polynomial of u over $\mathbb{Q}$. It is evident that if $m(x) \in \mathbb{Z}[x]$, then u is an algebraic integer.

So now suppose that u is an algebraic integer and pick $f(x) \in \mathbb{Z}[x]$ so that $f(x)$ is monic and u is a root of $f(x)$. Then we have that $m(x) \mid f(x)$ in $\mathbb{Q}[x]$. Now $f(x)$ factors (uniquely) into irreducible monic factors in $\mathbb{Z}[x]$. We also know that every irreducible in $\mathbb{Z}[x]$ is irreducible in $\mathbb{Q}[x]$. This means our factorization of $f(x)$ in $\mathbb{Z}[x]$ is also a factorization of $f(x)$ in $\mathbb{Q}[x]$ into irreducibles. So $m(x)$ must be an associate (over $\mathbb{Q}$) of one of the factors of $f(x)$. But the factors of $f(x)$ as well as $m(x)$ are monic. This forces the unit involved in the association to be 1 and so $m(x)$ must be one of the irreducible factors of $f(x)$, which were all in $\mathbb{Z}[x]$. So $m(x) \in \mathbb{Z}[x]$, as desired. □

Theorem on Rational Algebraic Integers. *A rational number is an algebraic integer if and only if it is an integer. A complex number u is algebraic if and only if there is some $b \in \mathbb{Z}$ with $b \neq 0$ such that bu is an algebraic integer.*

Proof. Evidently, every member of $\mathbb{Z}$ is an algebraic integer. For the converse, suppose that u is an algebraic integer that happens to be rational. Then $x - u$ is the minimal polynomial of u over $\mathbb{Q}$. By the fact above, it belongs to $\mathbb{Z}[x]$. Hence, $u \in \mathbb{Z}$.

Now let u be a complex number that is algebraic over $\mathbb{Q}$. Let $m(x) \in \mathbb{Q}[x]$ be the minimal polynomial of u. Let $b \in \mathbb{Z}$ be the product of the denominators of the coefficients of $m(x)$. Suppose

$$m(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_{n-1}x^{n-1} + x^n.$$

Let

$$f(x) = b^n a_0 + b^{n-1}a_1 x + b^{n-2}a_2 x^2 + \cdots + b a_{n-1}x^{n-1} + x^n.$$

Then $f(x) \in \mathbb{Z}[x]$ and $f(x)$ is monic. Every graduate student will check that $f(bu) = b^n m(u) = 0$. So bu is an algebraic integer. For the converse, if bu is a root of a monic $g(x) \in \mathbb{Z}[x]$. Then u is a root of $g(bx)$, which is certainly a polynomial in $\mathbb{Z}[x] \subseteq \mathbb{Q}[x]$, even if it is not monic. So u is algebraic over $\mathbb{Q}$. □

Let **E** be any algebraic extension of $\mathbb{Q}$. We make the harmless assumption that $\mathbb{C}$ is an extension of **E**. The following characterization of algebraic integers proves useful.

Theorem Characterizing Algebraic Integers. *Let* **E** *be an algebraic extension of* $\mathbb{Q}$ *and let* $u \in E$*. The element u is an algebraic integer if and only if* $\mathbb{Z}[u]$ *is finitely generated as a* $\mathbb{Z}$*-module. Moreover, if* $\mathbb{Z}[u]$ *is finitely generated as a* $\mathbb{Z}$*-module and* $w \in \mathbb{Z}[u]$*, then w is an algebraic integer.*

Proof. First, suppose that u is an algebraic integer. Let $m(x) \in \mathbb{Z}[x]$ be its minimal polynomial. Say it is of degree n. Then since $m(u) = 0$ and $m(x)$ is monic, we can express u^n as a linear combination of $1, u, u^2, \ldots, u^{n-1}$ using coefficients from $\mathbb{Z}$. The same applies to u^m for all $m \geq n$. But $1, u, u^2, \ldots, u^n, u^{n+1}, \ldots$ generate $\mathbb{Z}[u]$ as a $\mathbb{Z}$-module. Since we can make do with just the first n powers of u, we find that $[u]$ is finitely generated.

For the converse and the "moreover" part of the theorem, suppose $\mathbb{Z}[u]$ is finitely generated as an $\mathbb{Z}$-module and that $w \in \mathbb{Z}[u]$. Let $v_0, \ldots, v_{n-1}$ be such a generating set. Then notice that $wv_i \in \mathbb{Z}[u]$ for each $i < n$. This means that we can find integers $a_{i,0}, a_{i,1}, \ldots, a_{i,n-1}$ so that $uwv_i = a_{i,0}v_0 + a_{i,1}v_1 + \cdots + a_{i,n-1}v_{n-1}$. We get a whole system of equations that we can represent in matrix form:

$$w\begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{n-1} \end{pmatrix} = \begin{pmatrix} a_{0,0} & a_{0,1} & \ldots & a_{0,n-1} \\ a_{1,0} & a_{1,1} & \ldots & a_{1,n-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n-1,0} & a_{n-1,1} & \ldots & a_{n-1,n-1} \end{pmatrix}\begin{pmatrix} v_0 \\ v_1 \\ \vdots \\ v_{n-1} \end{pmatrix}$$

We write this more compactly as

$$w\bar{v} = A\bar{v}.$$

So a bit of rearrangement yields

$$(wI - A)\,\bar{v} = 0.$$

Now $\bar{v}$ cannot be the zero vector. So it must be that $\det(wI - A) = 0$. But $\det(xI - A)$ is the characteristic polynomial of the integer matrix A—it is a monic polynomial with integer coefficients and w is a root of it. So w is an algebraic integer, as desired. □

While this theorem was framed with reference to the ring $\mathbb{Z}$ and its field of fractions $\mathbb{Q}$ and an algebraic extension **E** of that field of fractions. The proof works quite generally.

Graduate students know the routine that will replace the single element u by a finite system of elements $u_0, \ldots, u_{m-1}$ in this characterization.

Now consider the set $O_{\mathbf{E}}$ of all algebraic integers belonging to E. Consider algebraic integers u and v. Then $\mathbb{Z}[u, v]$ will be finitely generated as a $\mathbb{Z}$-module. (You did that routine bit, right?) So each element of $\mathbb{Z}[u, v]$ is an algebraic integer. This means, $-u, u + v$, and uv are all algebraic integers. In other words, $\mathbf{O_E}$ is a subring of **E**.

29.3 Proving the Lindemann-Weierstrass Theorem

There is one more useful idea before we get back to the Lindemann-Weierstrass Theorem.

Suppose that **E** is a Galois extension of $\mathbb{Q}$. We devise a ring, $\mathbf{R_E}$, which, while it may strike you as peculiar, has a number of useful features.

The elements of $\mathbf{R_E}$ are just those functions $f\colon E \to E$ that output the value 0 except at finitely many inputs. The function that is constantly 0 is one such function and it will serve as the zero of our ring. Moreover, in our ring we add just the way we added functions in calculus:

$$(f + g)(u) := f(u) + g(u).$$

It is the multiplication in our ring that is peculiar. In the first place, the identity element of this ring will be the function $\mathbb{1}$ defined via

$$\mathbb{1}(u) := \begin{cases} 1 & \text{if } u = 0 \\ 0 & \text{if } u \neq 0 \end{cases}$$

Here is how to define the strange product:

$$(fg)(u) := \sum_{v+w=u} f(v)g(w),$$

and we can get away with this definition since the sum above is finite, there being only finitely many values of v where $f(v)$ is different from 0. At any rate, we have a system consisting of the nonempty set $R_{\mathbf{E}}$ that is equipped with something we called addition, something we called multiplication, a negation (well, what else could it be?), and a couple of elements called zero and one. It at least has a chance to be a ring. The following lemma expresses some basic properties of this construction.

Lemma. *Let* $\mathbf{E}$ *be a Galois extension of* $\mathbb{Q}$ *and let* $\mathbf{G}$ *be its Galois group. Then* $\mathbf{R_E}$ *is an integral domain and for all* $\sigma \in \mathbf{G}$ *and all* $f \in \mathbf{R_E}$ *both* $\sigma \circ f$ *and* $f \circ \sigma$ *are elements of* $\mathbf{R_E}$*.*

Proof. The graduate students will know the way to check that $\mathbf{R_E}$ is indeed a commutative ring in which the zero and the one are distinct. To see that the product of two nonzero elements is again nonzero, we impose a strict linear ordering $\sqsubset$ on $\mathbf{E}$ as follows:

$$a + bi \sqsubset c + di \text{ if and only if } a < c \text{ or both } a = c \text{ and } b < d,$$

where a, b, c, and d are real. This is usually called the lexicographic order. It respects addition. Now suppose that f and g are nonzero elements of $\mathbf{R_E}$. There are only finitely many (but at least one) elements of E that are assigned nonzero values by f. Let u be the one that is largest with respect to $\sqsubset$. Likewise, let v be the largest element, with respect to $\sqsubset$ that g assigns a nonzero value. But then $(fg)(u+v) = f(u)g(v) \neq 0$. So $\mathbf{R_E}$ is an integral domain.

Now let f be an element of $\mathbf{R_E}$, let σ be an element of the Galois group. Plainly, $\sigma \circ f \colon E \to E$ and since σ is an automorphism it sends 0 to 0. So $\sigma \circ f$ can assign nonzero values to only finitely many inputs (in fact, the same inputs assigned 0 to f). This means $\sigma \circ f$ belongs to $\mathbf{R_E}$. To handle $f \circ \sigma$ notice that for each $u \in E$ our polynomial of minimal degree has integer coefficients. Hence, σ fixes all the coefficients of this minimal polynomial. Thus, $\sigma(u)$ must be a root of the same minimal polynomial. Therefore $\sigma(u)$ is again in E. So $f \circ \sigma \colon \mathbf{E} \to \mathbf{E}$. And once again you can see, if you look hard enough, that this composite function only assigns finitely many inputs nonzero values. □

Proof of the Lindemann-Weierstrass Theorem, alternate Version. The proof is by way of contradiction. At the outset we assume there is that following kind of set-up: A system $\langle u_0, \ldots, u_{n-1} \rangle$ of distinct algebraic numbers and a system $\langle c_0, \ldots, c_{n-1} \rangle$ of algebraic numbers, not all 0, so that

$$c_0 e^{u_0} + \cdots + c_{n-1} e^{u_{n-1}} = 0.$$

We will call such a set-up a *contradictory set-up*. The theorem we are trying to prove asserts that there are no contradictory set-ups.

The first part of this proof involves devising contradictory set-ups with better and better properties. Once we have too many good properties we will have our contradiction—it will be too good to be true.

Suppose we have a contradictory set-up. Each u_i in this set-up is the root of some polynomial of minimal degree that belongs to $\mathbb{Z}[x]$ (it may not be monic). Any other root of this polynomial will also be algebraic. Let us expand our set-up by including all these other roots and do it for each u_i. Then this expanded set-up is again contradictory.

So let us replace our original contradictory set-up with this expanded one. (So we won't have to introduce new elements with new indices.) Now the elements $u_0, \ldots, u_{n-1}, c_0, \ldots, c_{n-1}$ are all algebraic. Let $\mathbf{E}$ be a Galois extension of $\mathbb{Q}$ that includes all these elements. Denote the Galois group of this extension by $\mathbf{G}$.

Define the map $\Psi \colon \mathbf{R_E} \to \mathbb{C}$ via

$$\Psi(f) = \sum_{u \in E} f(u) e^u.$$

I leave it in the trustworthy hands of the graduate students to prove that this map is a homomorphism. (The only challenging bit concerns the preservation of the peculiar multiplication—just remember $e^{v+w} = e^v e^w$.)

By letting $f \in R_{\mathbf{E}}$ be the function that assigns to each u_i the value c_i and assigns 0 to every thing else, we see that f belongs to the kernel of Ψ, which is therefore nontrivial. (Indeed, the kernel of Ψ is the place to look when devising contradictory set-ups.) Now let

$$g = \prod_{\sigma \in G} (\sigma \circ f).$$

Observe that each $\sigma \circ f$ is a nonzero element of $R_{\mathbf{E}}$. Since $\mathbf{R_E}$ is an integral domain, we see that g is also nonzero. On the other hand, since the identity permutation is one of the choices for σ, we see that f itself is one of the factors that make up g. So g is in the kernel of Ψ. But we also see from the equation displayed above, that $\tau \circ g = g$ for every $\tau \in \mathbf{G}$. This entails that $\tau(g(v)) = g(v)$ for all $\tau \in \mathbf{G}$ and all $v \in E$. As $\mathbb{Q}$ is the fixed field of $\mathbf{G}$ by the Fundamental Theorem of Galois Theory, we have that $g(v) \in \mathbb{Q}$ for all $v \in E$, This gives us

$$0 = \Psi(g) = \sum_{u \in E} g(u)e^u$$

So by using those algebraic numbers that g gives nonzero values, we have devised a contradictory set-up where the linear combination has *rational* coefficients. By multiplying this linear combination by an integer to cancel denominators, we obtain a contradictory set-up in which the linear combination has *integer* coefficients. For convenience, we redefine g so that it produces these integers, rather than the original rationals.

Once more it is harmless to suppose our new system $u_0, \ldots, u_{n-1}$ of algebraic integers is complete in the sense that if v is the image under σ of some algebraic number on our list, where $\sigma \in \mathbf{G}$, then v is also on the list.

Observe that $\mathbf{G}$ partitions the set E into orbits.

Now define

$$h = \prod_{\sigma \in G} (g \circ \sigma).$$

You see that h is not the zero of $\mathbf{R_E}$ and that h belongs to the kernel of Ψ and that all the values assigned by h are actually integers. But this time $h \circ \tau = h$ for all $\tau \in \mathbf{G}$. That is $h(v) = h(w)$, if v and w lie in the same orbit. So we find

$$0 = \sum_{u \in E} h(u)e^u,$$

where all the coefficients are integers and any two algebraic numbers that lie in the same orbit produce the same coefficient.

We are once more in position to revise our contradictory set-up. We replace our system of algebraic numbers with those algebraic numbers which h assigns nonzero values, being sure to add all their images by means of automorphisms in $\mathbf{G}$. So now we have a contradictory set-up in which the coefficients of the linear combination are all integers and those coefficients that are associated with algebraic numbers in the same orbit are themselves the same.

We need one more revision. We want 0 to be among the algebraic numbers in our contradictory set-up. Pick an algebraic number w so that $h(w) \neq 0$. Now define $m \colon E \to E$ by

$$m(-v) = \begin{cases} h(w) & \text{if } w \text{ and } v \text{ lie in the same orbit} \\ 0 & \text{otherwise.} \end{cases}$$

Because the orbits here are finite, we find m belongs to $R_{\mathbf{E}}$. Finally, put $t = mh$. Since h belongs to the kernel of Ψ so does the function t. Moreover, t is not the zero function, since neither m nor h are and we know $\mathbf{R_E}$ is an integral domain. Now observe

$$t(0) = (mh)(0) = \sum_{u+v=0} m(u)h(v) = \sum_{v \in E} m(-v)h(v) = a(h(w))^2,$$

where a is just the number of elements in the orbit of w. To retain that nice feature about orbits, we have to show that $t \circ \sigma = t$ for all σ in the Galois group. So let σ be an element of the Galois group and $u \in E$. Then

$$t(\sigma(u)) = (mh)(\sigma(u)) = \sum_{v+z=\sigma(u)} m(v)h(z) = \sum_{\sigma^{-1}(v)+\sigma^{-1}(z)=u} m(v)h(z).$$

But since $v = \sigma(\sigma^{-1}(v))$ and likewise for z, the last sum becomes

$$\sum_{\sigma^{-1}(v)+\sigma^{-1}(z)=u} m(\sigma(\sigma^{-1}(v)))h(\sigma(\sigma^{-1}(z))) = \sum_{q+r=u} m(\sigma(q))h(\sigma(r)) = t(u).$$

So we find indeed that $t \circ \sigma = t$ for all members σ of the Galois group.

Now as a last revision, consider the contradictory set-up where the system of u_i's consists of those algebraic numbers to which t assigns nonzero values, as well as all the algebraic numbers that lie in orbits including one of the algebraic numbers that t assigns a nonzero value. The integer 0 will be among these u_i's and the associated linear combination will be $\Psi(t) = 0$ and it will have two additional properties: all its coefficients will be integers and the coefficients associated to u_i's that belong to the same orbit will be the same. Also observe that the coefficient associate with 0 will be $t(0)$, which is not 0.

The Galois group $\mathbf{G}$ partitions $\{u_0,\ldots,u_{n-1}\}$ into orbits. Say there are m of them: $U_0,\ldots,U_{m-1}$. Then the linear combination in our contradictory set-up looks like

$$0 = c_0 + c_1 \sum_{u\in U_0} e^u + \cdots + c_{m-1} \sum_{u\in U_{m-1}} e^u$$

where each c_j is an integer and we know $c_0 \neq 0$.

The last stage in our proof is to show that there is a polynomial $q(x)$ with integer coefficients that very closely approximates e^x at each of the u_i's. (If you have seen such a polynomial approximation, then it is here that you might suspect the hand of Weierstrass.)

Lemma. *For any large enough prime number p there is an integer k not divisible by p and a polynomial $q(x)$ all of whose coefficients are integers divisible by p such that*

$$\left|ke^u - q(u)\right| < \frac{1}{p} \qquad \textit{for every } u \textit{ listed in our contradictory set-up except } u = 0,$$

and $q(u)$ is an algebraic integer for each u in our contradictory set-up except $u = 0$.

The threshold for that "large enough" depends on the sequence $u_o,\ldots,u_{n-1}$ in our ultimate contradictory set-up. This lemma is suppose to allow us to take that last linear combination and replace all those e^u's by corresponding $q(u)$'s while keeping the whole mess close to 0. The upshot is intended to be a positive integer strictly less than 1, certainly a contradiction, as we desired.

Since the proof of the lemma above runs to a couple of pages, let's finish the proof of the Alternate Version of the Lindemann-Weierstrass Theorem (and hence the Lindemann-Weierstrass Theorem) first.

$$\begin{aligned}
0 &= c_0 + c_1 \sum_{u\in U_0} e^u + \cdots + c_{m-1} \sum_{u\in U_{m-1}} e^u \\
0 &= kc_0 + c_1 \sum_{u\in U_0} ke^u + \cdots + c_{m-1} \sum_{u\in U_{m-1}} ke^u \\
\text{Let } z &:= kc_0 + c_1 \sum_{u\in U_0} q(u) + \cdots + c_{m-1} \sum_{u\in U_{m-1}} q(u) \qquad (*)
\end{aligned}$$

Subtracting we get

$$-z = \quad c_1 \sum_{u\in U_0} (ke^u - q(u)) + \cdots + c_{m-1} \sum_{u\in U_{m-1}} (ke^u - q(u))$$

Let λ be the maximum of the integers $c_1,\ldots,c_{m-1}$. Then using the triangle inequality and the lemma, we get

$$|z| \le \frac{n\lambda}{p} \qquad \text{recall } n \text{ is the number of } u\text{'s in our contradictory set-up}$$

$$|z| < 1 \qquad \text{since the lemma allows us to pick arbitrarily large } p\text{'s.}$$

On the other hand, each of the sums in $(*)$, where z was defined, is left fixed by each automorphism in our Galois group. The fixed field of this Galois group is $\mathbb{Q}$. So each of these sums is an algebraic integer that is also rational. The only rational algebraic integers are the ordinary integers. Also, because each coefficient of $q(x)$ is divisible by p, we see that each of the sums is, as well. But kc_0 is *not* divisible by p, provided we take p large enough that it cannot divide c_0. This means that z will be an integer that is not divisible by p. So the integer z is not 0. This means that $|z|$ is a positive integer that is less than 1. Our contradiction, at last! □

Proof of our Lemma. It is possible to say immediately what the threshold for the "any large enough prime number" is and also what the value of k is, in terms of p. However, then these values would seem to be without motivation. Instead, we allow the values to become visible during the course of the proof. So for now, just image p is some very large prime number.

We will make the polynomial $q(x)$ from the polynomials of minimal degree of our various u's. Let $F(x)$ be the product of all the distinct such minimal polynomials, except the minimal polynomial of 0. Then $F(x)$ will be a polynomial with integer coefficients, its constant term won't be 0, and it will have all our u's as roots, with the exception of 0. We reserve s for the degree of $F(x)$. The polynomial $F(x)$ won't do for $q(x)$ because we have no reason to think of it as a good approximation to e^x and also because its coefficients may well not be divisible by p, a property we need. We have another trouble—$q(u)$ is suppose to turn into a algebraic integer. Pick an integer d so that du^k is an algebraic integer for each $k \le s+1$ and for each u from our contradictory set-up

It is suggestive to recall the Taylor series for the exponential function:

$$e^x = \frac{x^0}{0!} + \frac{x^1}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

Let us try

$$G(x) = \frac{d^p x^{p-1}}{(p-1)!}(F(x))^p = \frac{(dxF(x))^{p-1}\,dF(x)}{(p-1)!}$$

This, at least, has the look of a term of the Taylor series. We make this into a series:

$$H(x) := G^{(0)}(x) + G^{(1)}(x) + G^{(2)}(x) + \cdots$$

where we are taking formal derivative of the polynomial $G(x)$. Notice that $H(x)$ is really a polynomial, since $G(x)$ is a polynomial and eventually all those derivative will vanish. The graduate students are welcome to tackle those derivatives with the help of the Product Rule and the Chain Rule!

To get a better idea, let

$$(F(x))^p = b_0 + b_1 x + b_2 x^2 + \cdots + b_{r-1}x^{r-1} + b_r x^r.$$

Recall we know that $F(x)$ has integer coefficients. So

$$x^{p-1}(F(x))^p = b_0 x^{p-1} + b_1 x^p + b_2 x^{p+1} + \cdots + b_{r-1}x^{p+r-2} + b_r x^{r+p-1}.$$

So you can see that $G^{(j)}(0) = 0$ for all $j < p-1$ and $G^{(p-1)}(0) = d^p b_0$. A little bit of diligence gives

$$G^{(p)}(x) = d^p\left[p!\binom{p+1}{1}b_1 + p!\binom{p+2}{2}b_2 x + p!\binom{p+3}{3}b_3 x^2 + \cdots\right]$$

So all the coefficients of $G^{(p)}(x)$ are integers divisible by p. The same holds for the coefficients of $G^{(j)}(x)$ when $j \ge p$.

So let us try

$$q(x) := G^{(p)}(x) + G^{(p+1)}(x) + \cdots$$

and we will take

$$k := H(0) = G^{(0)}(0) + G^{(1)}(0) + \cdots = G^{(p-1)}(0) + G^{(p)}(0) + \cdots$$

We have arranged that $F(u) = 0$ for each nonzero u's in our contradictory set-up. If you were virtuous with the Product Rule and the Chain Rule above, you will agree that $G^{(j)}(u) = 0$ for each of our nonzero u's as long as $j < p$. [In case, you were wondering, this is the reason we raised $F(x)$ to the p^{th} power when we defined $G(x)$.] But also, examining the second form given when $G(x)$ was defined you will see that when one of our u's is plugged in for x the result will be algebraic integers even before all the terms in the polynomials are added together. Due diligence with the Product Rule and the Chain Rule should convince you that the same applies to each derivative of $G(x)$. Hence $q(u)$ will be an algebraic integer whenever u is an algebraic number from our contradictory set-up.

We must also ensure that $H(0)$ is not divisible by p. Recall that $G^{(p-1)}(0) = d^p b_0$, where b_0 is the p^{th} power of the product of the constant terms of those polynomials of minimal degree. On the other hand, $G^{(j)}(0)$ is divisible by p whenever $j \geq p$. So we can ensure that $H(0)$ is *not* divisible by p just by insisting that p is a prime larger than any that divide any of the constant terms of those minimal polynomials or that divide d.

What remains of the proof of the lemma is to contend with the inequality. Reformulate it as

$$\left|H(0)e^u - H(u)\right| < \frac{1}{p} \quad \text{for all of our } u\text{'s.}$$

Establishing this would complete the proof of the lemma. Rewriting a bit we get

$$|e^u||e^{-u}H(u) - H(0)| < \frac{1}{p} \quad \text{for all of our } u\text{'s.} \tag{$\star$}$$

Notice

$$\begin{aligned} H(x) &= G(x) + G^{(1)}(x) + G^{(2)}(x) + G^{(3)}(x) + \cdots \\ H^{(1)}(x) &= \phantom{G(x) + {}} G^{(1)}(x) + G^{(2)}(x) + G^{(3)}(x) + \cdots \end{aligned}$$

Now let $\varphi(x) := e^{-x}H(x)$. Observe that $\varphi(0) = H(0)$, so

$$e^{-u}H(u) - H(0) = \varphi(u) - \varphi(0).$$

This last is reminiscent of the one side of the Mean Value Theorem from calculus on the reals. Still, an argument for the Mean Value Theorem in several variables (but applied on the complex plane) yields

$$|\varphi(u) - \varphi(0)| \leq |\varphi'(v)||u - 0| \quad \text{for some } v \text{ on the line seqment joining } u \text{ and } 0.$$

So we see that ($\star$) becomes

$$|ue^u||\varphi'(v)| < \frac{1}{p} \quad \text{for each } u, \text{ where } v \text{ is between } u \text{ and } 0. \tag{$\star\star$}$$

So consider the derivative.

$$\begin{aligned} \varphi'(x) &= -e^{-x}H(x) + e^{-x}H^{(1)}(x) \\ &= -e^{-x}\left[H(x) - H^{(1)}(x)\right] \\ &= -e^{-x}G(x). \end{aligned}$$

Hence

$$|\varphi'(x) = |e^{-x}||G(x)|.$$

Let ρ be the maximum of the $|u|$ as u ranges through the finitely many algbraic integers in our contradictory set-up and let D be the disk in the complex plane about 0 with radius ρ.

$$|\varphi'(x)| \le e^{\rho}|G(x)| \quad \text{for all } x \in D.$$

Putting this together, we see that it is enough to establish

$$\rho e^{2\rho}|G(x)| < \frac{1}{p} \quad \text{for each } x \in D. \tag{$\star\star\star$}$$

Now recall

$$G(x) = \frac{d^p x^{p-1}}{(p-1)!}\,(F(x))^p$$

where $F(x)$ was the product of the minimal polynomials of the nonzero u's. So we can rewrite $|\rho e^{\rho} G(x)|$ as $\frac{1}{(p-1)!}\,(|K(x)|)^p$, where $K(x)$ is a function that depends on our u's, but not on the prime p. Let μ be the maximum value of $|K(x)|$ on the compact set D.

So finally, the proof of the lemma will be complete if you can prove that for any large enough prime p that following holds:

$$\frac{\mu^p}{(p-1)!} < \frac{1}{p}.$$

Another way to write this is

$$\frac{p\mu^p}{(p-1)!} < 1.$$

Well, does this hold when p is large enough? What do you think? □

29.4 Problem Set 24

Algebra Homework, Editon 24

Twenty Fifth Week

Transcendental numbers, constructible number, and other puzzles

Problem 122.
Prove that $\ln u$ and $\sin u$ are transcendental over the field of rational numbers, whenever u is a positive algebraic real number (and in the case of $\ln u$ that $u \neq 1$).

Problem 123.
Let $\mathbf{F}, \mathbf{K}$, and $\mathbf{L}$ be fields so that $\mathbf{K}$ is a finite separable extension of $\mathbf{F}$ and $\mathbf{L}$ is a finite separable extension of $\mathbf{K}$. Prove that $\mathbf{L}$ is a finite separable extension of $\mathbf{F}$.

Problem 124.
Prove that no finite field is algebraically closed.

Problem 125.
Archimedes studied cylinders circumscribed around spheres. We say that such a cylinder is **constructible** provided the radius of the sphere is a constructible real number. So the cylinder circumscribed around a sphere of radius 1 is constructible. Call this cylinder the **unit cylinder**. Let $\mathbf{C}$ be a cylinder circumscribed around a sphere so that the volume of $\mathbf{C}$ is twice as larger as the volume of the unit cylinder. Explain in detail why $\mathbf{C}$ is not constructible.

LECTURE 30

THE GALOIS CONNECTION BETWEEN POLYNOMIAL RINGS AND AFFINE SPACES

The unit circle in $\mathbb{R}\times\mathbb{R}$ is described by $x^2+y^2=1$. Likewise, the unit sphere in $\mathbb{R}\times\mathbb{R}\times\mathbb{R}$ is described by $x^2+y^2+z^2=1$. More generally, we could consider any field $\mathbf{F}$ and any system $\langle x_0,x_1,\dots,x_{n-1}\rangle$ of variables and any set Γ of equations $p(x_0,\dots,x_{n-1})=q(x_0,\dots,x_{n-1})$, where $p(x_0,\dots,x_{n-1})$ and $q(x_0,\dots,x_{n-1})$ belong to $\mathbf{F}[x_0,\dots,x_{n-1}]$. Such a set of equations describes a set of points in the **affine space** $\mathbf{F}^n$, namely the set of all n-tuples $\bar{a}=\langle a_0,\dots,a_{n-1}\rangle$ for which all the equations in Γ are true. Sets of points in such n-dimensional spaces are called **affine varieties**. Roughly speaking, affine varieties are geometric objects with what might be thought of as straightforward algebraic descriptions. They are the basic objects of study in algebraic geometry.

To conserve notation, we will write $f(\bar{x})$ and $\mathbf{F}[\bar{x}]$, meaning that $\bar{x}=\langle x_0,\dots,x_{n-1}\rangle$.

There is an immediate way to simplify the presentation of the notion of affine variety. Equations of the form $p(\bar{x})=q(\bar{x})$ can be replaced by equations of the form $p(\bar{x})-q(\bar{x})=0$. Or more simply, equations of the form $f(\bar{x})=0$, where $f(\bar{x})$ is a polynomial. In this way we can even construe Γ has a set of polynomials (rather than a set of equations) and the affine variety described by Γ will then be the points in n-dimensional space (over our field), that are simultaneous **zeros** of the polynomials in Γ.

So you can now see the two-place relation $\left\{\langle f(\bar{x}),\bar{a}\rangle \mid f(\bar{x})\in\mathbf{F}[\bar{x}] \text{ and } \bar{a}\in\mathbf{F}^n \text{ and } f(\bar{a})=0\right\}$ between the ring $\mathbf{F}[\bar{x}]$ and affine space $\mathbf{F}^n$. We can use this relation to establish a Galois connection between the ring of polynomials and the points in n-dimensional space. The polarities of this Galois connection are denoted as follows:

$$\mathbf{V}(\Gamma) := \{\bar{a} \mid \bar{a}\in\mathbf{F}^n \text{ and } f(\bar{a})=0 \text{ for all } f(\bar{x})\in\Gamma\}$$
$$\mathbf{I}(M) := \{f(\bar{x}) \mid f(\bar{x})\in\mathbf{F}[\bar{x}] \text{ and } f(\bar{a})=0 \text{ for all } \bar{a}\in M\}$$

for any $\Gamma\subseteq\mathbf{F}[\bar{x}]$ and any $M\subseteq\mathbf{F}^n$.

The closed sets on the geometrical side of this Galois connection will be precisely the affine varieties. What about the closed sets on the side of the polynomial ring? Well, it is straightforward to check that $\mathbf{I}(M)$ will be an ideal of the polynomial ring (why not check it now?). However, unlike with Galois' original connection (every subgroup was closed and every intermediate field was closed...), this time it may not be that every ideal is closed.

Let $\bar{a}$ be in $\mathbf{F}^n$. Suppose $f(\bar{x}) \in \mathbf{F}[\bar{x}]$ and let k be a positive integer. Observe that $\big(f(\bar{a})\big)^k = 0$ if and only if $f(\bar{a}) = 0$. This entails

$$\text{If } \big(f(\bar{x})\big)^k \in \mathbf{I}(M) \text{ for some positve integer } k, \text{ then } f(\bar{x}) \in \mathbf{I}(M),$$

for any $M \subseteq F^n$. So we see that the closed ideals must have this special property. This can be recast in an arbitrary commutative ring $\mathbf{R}$. Suppose J is an ideal of $\mathbf{R}$. It is straightforward to check that

$$\{u \mid u \in R \text{ and } u^k \in J \text{ for some positive integer } k\}$$

is always an ideal of $\mathbf{R}$ (this is a consequence of the Binomial Theorem, which holds in commutative rings). The ideal above, built from the ideal J, is called the **nilradical** ideal of J. I will denote it by $\operatorname{nilrad} J$ (although I am attracted to $\sqrt{J}$ as well). This is a kind of closure operator on ideals in the sense that

$$J \subseteq \operatorname{nilrad} J$$
$$\operatorname{nilrad} J = \operatorname{nilrad}\operatorname{nilrad} J$$

for all ideals J of our commutative ring. The ideal J is a **nilradical ideal** provided $J = \operatorname{nilrad} J$.

Returning to our Galois connection between the polynomial ring $\mathbf{F}[\bar{x}]$ and the affine space $\mathbf{F}^n$, we see that the closed sets of the side on the polynomial ring must all be nilradical ideals. This leaves open the question of whether some of the nilradical ideals might fail to be closed in the sense of our Galois connection. We would like to establish that the closed sets are precisely the nilradical ideals.

Hilbert's Nullstellensatz. *For any positive natural number n and any algebraically closed field* $\mathbf{F}$*, the closed subsets of* $\mathbf{F}[x_0,\ldots,x_{n-1}]$ *with respect to the Galois connection are precisely the nilradical ideals of this polynomial ring.*

Of course, this means that the polarities of our Galois connection establish a dual isomorphism between the lattice of nilradical ideals of our polynomial ring $\mathbf{F}[\bar{x}]$ and the affine varieties of $\mathbf{F}^n$. Indeed, at this point it is possible to write down a theorem that looks very much like the Fundamental Theorem of Galois Theory, except here we are concerned with connecting polynomial rings and their nilradical ideals with affine spaces and their affine varieties.

Of the many proofs available of Hilbert's Nullstellensatz, in the next section you will find one that traces its ancestory back to Oscar Zariski in the late 1940's. In the section after that, I sketch out a distinctly different proof due to Abraham Robinson in the 1950's which is framed from the perspective of model theory and hinges on an interesting theorem from 1910 of Ernst Steinitz concerning uncountable algebraically closed fields.

Before turning to these proofs of Hilbert's Nullstellensatz, it proves useful to look a little at nilradical ideals.

Around 1930 Wolfgang Krull found the following characterization of the nilradical of any ideal.

Theorem Characterizing the Nilradical. *Let* $\mathbf{R}$ *be a nontrivial commutative ring and let* J *be any proper ideal of* $\mathbf{R}$*. Then* $\operatorname{nilrad} J$ *is the intersection of all the prime ideals that contain* J*.*

Proof. To see that $\operatorname{nilrad} J$ is included in the intersection of all the prime ideals that include J, let $u \in \operatorname{nilrad} J$. What we need is to show that $u \in P$ for every prime ideal P such that $J \subseteq P$. So let P be such a prime ideal. Pick a positive natural number k so that $u^k \in J$. Since $J \subseteq P$, we have $u^k \in P$. But P is a prime ideal, so $u \in P$ as desired.

For the converse, suppose that $u \notin \operatorname{nilrad} J$. Now we must find a prime ideal P with $u \notin P$ but $J \subseteq P$. What we know is that $u^k \notin J$ for every positive integer k. That is

$$\{u^k \mid k \text{ is a positive integer}\} \text{ and } J \text{ are disjoint.}$$

Let $\mathcal{F}$ consist of those ideals that include J and are disjoint from $\{u^k \mid k \text{ is a positive integer}\}$. A routine application of Zorn's Lemma shows that $\mathcal{F}$ has a maximal member. Let P be a maximal member of $\mathcal{F}$. Evidently, $u \notin P$.

I contend that P is a prime ideal, as we desire. To see this suppose $v, w \in R$ with $vw \in P$. For contradiction, suppose that $v \notin P$ and $w \notin P$. Then the ideals $(P \cup \{v\})$ and $(P \cup \{w\})$ must contain positive powers of u, by the maximality of P in $\mathcal{F}$. So pick positive integers k and ℓ so that $u^k \in (P \cup \{v\})$ and $u^\ell \in (P \cup \{w\})$. This means we can also pick $s, t \in P$ and $q, r \in R$ so that

$$u^k = s + qv \text{ and } u^\ell = t + rw.$$

But then, $u^{k+\ell} = (s+qv)(t+rw) = (s+qv)t + srw + qrvw$. Each term in the latter sum belongs to P. But then $u^{k+\ell} \in P$, a contradiction. □

30.1 Oscar Zariski proves Hilbert's Nullstellensatz

Every proper ideal of the ring $\mathbf{F}[\bar{x}]$ extends to a maximal ideal and every maximal ideal in a prime ideal. From Hilbert's Nullstellensatz, we see that the variety $\mathbf{V}(I)$ of any proper ideal I cannot be the smallest variety, namely the empty set of points, which corresponds to the improper ideal—the whole of $\mathbf{F}[\bar{x}]$. In this way, we have for any algebraically closed field $\mathbf{F}$

Every proper ideal of $\mathbf{F}[\bar{x}]$ has a solution $\bar{a}$ in $\mathbf{F}^n$.

Of course, this is a consequence of Hilbert's Nullstellensatz. Sometimes it is called the *weak Nullstellensatz*. In 1930, George Yuri Rainich published a short paper under his birthname J. L. Rabinowitsch, with the observation that Hilbert's Nullstellensatz can be easily deduced from this consequence.

Here is how.

Let I be any ideal of $\mathbf{F}[\bar{x}]$. We already know that

$$\operatorname{nilrad} I \subseteq \mathbf{I}(\mathbf{V}(I)).$$

So we only need the reverse inclusion. In the event that $I = \mathbf{F}[\bar{x}]$, that inclusion is clear. So we consider the case when I is a proper ideal. Suppose that $f(\bar{x}) \in \mathbf{I}(\mathbf{V}(I))$. By Hilbert's Basis Theorem, we know that $\mathbf{F}[\bar{x}]$ is Noetherian. So I is finitely generated. Let us suppose that I is generated by $g_0(\bar{x}), \dots, g_{m-1}(\bar{x})$. So we know that if $g_i(\bar{a}) = 0$ for each $i < m$, then we have $f(\bar{a}) = 0$ as well. Let us introduce a new variable y. So the system

$$g_0(\bar{x}) = 0, \dots, g_{m-1}(\bar{x}) = 0, 1 - yf(\bar{x}) = 0$$

of equations has no solutions. Let J be the ideal of $\mathbf{F}[\bar{x}, y]$ generated by

$$\left\{g_0(\bar{x}), \dots, g_{m-1}(\bar{x}), 1 - yf(\bar{x})\right\}.$$

This ideal has no solution. So by the weak version of the Nullstellensatz we find that $J = \mathbf{F}[\bar{x}, y]$. This means that there are polynomials $h_i(\bar{x}, y) \in \mathbf{F}[\bar{x}, y]$ for $i \leq m$ so that

$$1 = h_0(\bar{x}, y)g_0(\bar{x}) + \dots + h_{m-1}(\bar{x}, y)g_{m-1}(\bar{x}) + h_m(\bar{x}, y)\left(1 - yf(\bar{x})\right).$$

Now $\mathbf{F}[\bar{x}, y]$ is an integral domain. Let $\mathbf{K}$ be its field of fractions. We consider that $\mathbf{K}$ is an extension of $\mathbf{F}[\bar{x}, y]$. So the elements of $\mathbf{K}$ are of the formal ratios $\frac{s(\bar{x}, y)}{t(\bar{x}, y)}$ of polynomials where $t(\bar{x}, y)$ is not the zero polynomial. We can use our map extension method to obtain a homomorphism from $\mathbf{F}[\bar{x}, y]$ into $\mathbf{K}$ that fixes each element of $\mathbf{F}$ and sends each x_i to itself, but sends y to $\frac{1}{f(x)}$. Applying this homomorphism to the equation displayed above, we find the following equation must hold in $\mathbf{K}$.

$$1 = h_0\left(\bar{x}, \frac{1}{f(\bar{x})}\right)g_0(\bar{x}) + \dots + h_{m-1}\left(\bar{x}, \frac{1}{f(\bar{x})}\right)g_{m-1}(\bar{x}) + h_m\left(\bar{x}, \frac{1}{f(\bar{x})}\right)\left(1 - \frac{1}{f(\bar{x})}f(\bar{x})\right).$$

The last term is 0 and we can drop it. Recalling that the $h_i(\bar{x}, y)$ are polynomials, we see that the only denominators appearing in the displayed equation are just certain powers of $f(\bar{x})$. So we can multiply the whole equations by a sufficiently high power of $f(\bar{x})$ to clear the denominators. We get

$$\left(f(\bar{x})\right)^q = h_0^*(\bar{x})g_0(\bar{x}) + \cdots + h_{m-1}^*(\bar{x})g_{m-1}(\bar{x})$$

where $h_i^*(\bar{x}) = \left(f(\bar{x})\right)^q h_i\left(\bar{x}, \frac{1}{f(\bar{x})}\right)$, for each $i < m$.

This demonstrates that $f(\bar{x}) \in \operatorname{nilrad} I$.

Knowing of this trick of Rabinowitsch (which can be found in the second volume of B. L. van der Wearden's *Moderne Algebra* 1931), Oscar Zariski set out to prove Hilbert's Nullstellensatz by devising a proof that every proper ideal has a solution—a proof that would require little in the way of additional ideas. He actually gave two new proofs in a paper that appeared in 1949.

Zariski based the proof given here on the following theorem, interesting in its own right.

Zariski's Theorem on Finite Integral Extensions. *Let* **F** *be a field and let* **D** *be an integral domain extending* **F** *that is generated by a finite set over* **F**. *If* **D** *is a field, then* **D** *is an algebraic extension of* **F**.

Here is how to deduce the weak version of the Nullstellensatz from Zariski's Theorem. Suppose that I is a proper ideal of $\mathbf{F}[\bar{x}]$. Let M be a maximal ideal that extends I. So $\mathbf{F}[\bar{x}]/M$ is a field. Moreover, it is generated, as a ring, over **F** by the elements $x_0 + M, \dots, x_{n_1} + M$. By Zariski's Theorem on Finite Integral Extensions, we conclude that $\mathbf{F}[\bar{x}]/M$ is (isomorphic to) an algebraic extension of **F**. But under the hypotheses of Hilbert's Nullstellensatz (and its weak version) **F** is algebraically closed. So $\mathbf{F}[\bar{x}]/M$ is isomorphic to **F**. But notice that $\langle x_0 + M, \dots, x_{n-1} + M\rangle$ is a solution for I since $I \subseteq M$ and M is the kernel of the quotient map from $\mathbf{F}[\bar{x}]$ onto $\mathbf{F}[\bar{x}]/M$. This means that I must also have a solution in **F**.

So it remains to prove Zariski's Theorem of Finite Integral Extensions.

Proof of Zariski's Theorem. The proof is by induction on the number of generators (over **F**) of the integral domain **D**.

Base Step: D = F

There is nothing to prove.

Inductive Step

Here we assume that every integral domain generated, as a ring, over any field by n additional elements that happens itself to be a field, is an algebraic extension over the ground field. Suppose that **D** is generated over **F** by the $n+1$ nonzero elements $b_0, \dots, b_n$ and **D** is a field. This means that the elements of D arise as values, at $\langle b_0, \dots, b_n\rangle$, of polynomials from $\mathbf{F}[x_0, \dots, x_n]$. For this reason we denote **D** by $\mathbf{F}[b_0, \dots, b_n]$. The smallest subfield of **D** that includes $F \cup \{b_0\}$ is denoted by $\mathbf{F}(b_0)$. This field actually consists of the values at b_0 of all the rational functions with coefficients from **F** (with the exception of those whose denominators evaluates to 0).

Evidently, **D** is generated over $\mathbf{F}(b_0)$ by the n elements $b_1, \dots, b_n$. By the induction hypothesis, **D** is an algebraic extension of $\mathbf{F}(b_0)$.

To complete the proof, we need to see that $\mathbf{F}(b_0)$ is an algebraic extension of **F**. That is, we have to show that b_0 is algebraic over **F**.

We already know that if b_0 is algebraic over **F**, then $\mathbf{F}[b_0]$ is a field and so $\mathbf{F}[b_0] = \mathbf{F}(b_0)$. In the event that b_0 is *not* algebraic over **F** the eager graduate student can work out that the homomorphism from $\mathbf{F}[x]$ onto $\mathbf{F}[b_0]$ that sends x to b_0 must be an isomorphism. So in either case, $\mathbf{F}[b_0]$ will be a unique factorization domain. In $\mathbf{F}[x]$ for any nonzero nonunit polynomial (i.e. a polynomial of positive degree) $f(x)$, it is clear that $f(x)$ and $f(x)+1$ are relatively prime. This means that if b_0 fails to be algebraic over **F**, then there is no nonzero element of $\mathbf{F}[b_0]$ that is divisible by every irreducible element of $\mathbf{F}[b_0]$. What Zariski does is produce a nonzero element of $\mathbf{F}[b_0]$ that is divisible by every irreducible element of $\mathbf{F}[b_0]$. Then the conclusion that b_0 is algebraic over **F** follows.

Roughly speaking, Zariski's argument below amounts to clearing a lot of denominators and making a crucial finiteness observation.

As each b_i, with $1 \le i \le n$, is algebraic over the field $\mathbf{F}(b_0)$, we see that b_i will have a minimal polynomial over this field. The coefficients of these minimal polynomials will be ratios of polynomials with coefficients in F that have been evaluated at b_0. We can clear the denominators of these minimal polynomials, to obtain polynomials in x of positive degree whose coefficients lie in the ring $\mathbf{F}[b_0]$. Let $d(b_0)$ be the product of the leading coefficients of all these modified minimal polynomials. So $d(b_0)$ is not 0 and $d(b_0) \in F[b_0]$.

Consider b_1. Let r be the degree of the minimal polynomial of b_1 over $\mathbf{F}(b_0)$. Notice that $d(b_0)b_1^r$ is a linear combination of powers of b_1 smaller than r using coefficients from $\mathbf{F}[b_0]$. The same can be said for $d(b_0)^2 b_1^{r+1}$. In fact, any large power of b_1 can be dealt with in this way. The same applies to large powers of $b_2, \ldots, b_n$.

Now let w be any element of $\mathbf{D} = \mathbf{F}[b_0, b_1, \ldots, b_n]$. This means there is a polynomial in $\mathbf{F}[x_0, \ldots, x_n]$ so that

$$w = p(b_0, b_1, \ldots, b_n).$$

A particular term of this polynomial looks like $cb_0^{k_0} b_1^{k_1} \ldots b_n^{k_n}$, where $c \in F$. But then $cb_0^{k_0}$ is an element of $\mathbf{F}[b_0]$. So there is a positive natural number q, depending on w and probably pretty large, so that $d(b_0)^q w$ is a sum of terms of the form

$$s(b_0) b_1^{\ell_1} \ldots b_n^{\ell_n}$$

where each $s(b_0) \in \mathbf{F}[b_0]$ and each ℓ_k is smaller than the degree of the minimal polynomial of b_k over $\mathbf{F}(b_0)$. It is important to notice that there are only finitely many elements of the form $b_1^{\ell_1} \ldots b_n^{\ell_n}$.

As $\mathbf{F}[b_0, \ldots, b_n] = \mathbf{F}(b_0)[b_1, \ldots, b_n]$ is an extension of the field $\mathbf{F}(b_0)$ by finitely many algebraic elements it is finite dimensional over $\mathbf{F}(b_0)$. Let $\{v_0, v_1, \ldots, v_{m-1}\}$ be a basis, with $v_0 = 1$, for this extension.

We can express each of those finitely many elements of the form $b_1^{\ell_1} \ldots b_n^{\ell_n}$ as a linear combination of these basis vectors. The coefficients of the finitely many linear combinations arising in the way belong to $\mathbf{F}(b_0)$. Let $e(b_0) \in \mathbf{F}[b_0]$ be the product of all the denominators occurring in all these linear combinations. This means that $e(b_0) b_1^{\ell_1} \ldots b_n^{\ell_n}$ will be a linear combination of the v_k's *using coefficients from* $\mathbf{F}[b_0]$.

Putting things together, we see that if w is any element of $\mathbf{D}$, then there will be some positive natural number q, depending on w, so that $(e(b_0)d(b_0))^q\, w$ is a linear combination of the v_k's using coefficients from $\mathbf{F}[b_0]$.

Now let $g(b_0)$ be an irreducible element of $\mathbf{F}[b_0]$. So it is not zero. Let $w = \frac{1}{g(b_0)}$. Pick a natural number q so that $(e(b_0)d(b_0))^q\, w$ is a linear combination of the v_k's using coefficients from $\mathbf{F}[b_0]$. But $(e(b_0)d(b_0))^q\, w$ belongs to $\mathbf{F}(b_0)$. By the linear independence of the v_k's, it must be that there is some $s(b_0) \in \mathbf{F}[b_0]$ so that

$$(e(b_0)d(b_0))^q\, w = \frac{(e(b_0)d(b_0))^q}{g(b_0)} = s(b_0).$$

But this means

$$(e(b_0)d(b_0))^q = g(b_0)s(b_0).$$

But since $g(b_0)$ is irreducible, it is prime. So we discover that our arbitrarily chosen irreducible $g(b_0)$ in fact divides the element $e(b_0)d(b_0)$ of $\mathbf{F}[b_0]$. We know this cannot happen if $\mathbf{F}[b_0] \cong \mathbf{F}[x]$. So we find, as Zariski did, that $\mathbf{F}[b_0]$ is a field and that b_0 is algebraic over $\mathbf{F}$.

This completes the inductive step. □

30.2 Abraham Robinson proves Hilbert's Nullstellensatz

Abraham Robinson, in the early 1950's, provided a proof of the following theorem.

Theorem on the Power of Algebraic Closures. *Let* **F** *be any field. Let* Γ *be any finite system of polynomial equations and polynomial inequations with coefficients from* **F**. *If* Γ *has a solution in any field extending* **F**, *then* Γ *has a solution in the algebraic closure of* **F**.

Here is an example of such a system:

$$x^3+4xy^2+1=0, \quad xyz-x^2=1, \quad \text{and} \quad x+y+z\neq 17.$$

The reason why I have called this the Theorem on the Power of Algebraic Closures is simply that the definition of the algebraic closure of **F** amounts to saying that it is the smallest field extending **F** for which every system $\langle p(x)=0\rangle$, where $p(x)\in\mathbf{F}[x]$ has positive degree, has a solution. So the theorem asserts that this condition concerning systems of one equation in one unknown entails the much broader condition concerning finite systems of polynomial equations and inequations in many variables.

Before turning to the proof of this theorem, I want to show how it gives rise to a proof a Hilbert's Nullstellensatz.

Robinson's proof of Hilbert's Nullstellensatz. Let J be an ideal of $\mathbf{F}[\bar{x}]$. We already know that

$$\operatorname{nilrad} J\subseteq \mathbf{I}(\mathbf{V}(J)).$$

We need the reverse inclusion. So suppose that $f(\bar{x})\notin \operatorname{nilrad} J$. We have to show that

$$f(\bar{x})\notin \mathbf{I}(\mathbf{V}(J)).$$

That is, we must find some $\bar{a}\in\mathbf{V}(J)$ so that $f(\bar{a})\neq 0$. Accord to Hilbert's Basis Theorem, $\mathbf{F}[\bar{x}]$ is a Noetherian ring. Thus J is a finitely generated ideal. Let the polynomials $g_0(\bar{x}),\ldots,g_{m-1}(\bar{x})$ generate J. So what we need is an n-tuple $\bar{a}\in F^n$ so that

$$g_0(\bar{a})=0,\ldots,g_{m-1}(\bar{a})=0, \text{ and } f(\bar{a})\neq 0.$$

That is $\bar{a}$ is a solution to the following system of polynomial equations and inequations.

$$g_0(\bar{x})=0,\ldots,g_{m-1}(\bar{x}), \text{ and } f(\bar{x})\neq 0.$$

Call this system Γ. In order to invoke that Theorem on the Power of Algebraic Closures, we need only see that Γ has a solution in some field extending **F**.

Since $f(x)\notin\operatorname{nilrad} J$, by Krull's characterization of nilradicals, there is a prime ideal P so that $f(x)\notin P$ but $J\subseteq P$. Because P is prime, we know that $\mathbf{F}[\bar{x}]/P$ is an integral domain, call it **D**. Let η be the quotient map. Since all the nonzero elements of F are units, while P is a proper ideal, we see that η embeds **F** into **D**. For each $i<n$, let $b_i=\eta(x_i)$. Put $\bar{b}=\langle b_0,\ldots,b_{n-1}\rangle$. Since each $g_j(\bar{x})\in P$, we see that $g_j(\bar{b})=0$ in **D**. On the other hand, $f(\bar{b})\neq 0$ in **D**, since $f(\bar{x})\notin P=\ker\eta$. So $\bar{b}$ is a solution to Γ in **D** and **D** extends **F**. However, we only know that **D** is an integral domain and we want a field. However, every integral domain can be extended to a field **E** (namely its field of fractions). Now we can invoke the Theorem on the Power of Algebraic Closures to complete the proof. □

So it remains to prove the Theorem on the Power of Algebraic Closures. I give here a sketch of Robinson's argument.

Recall the following theorem from Lecture 20.

The Steinitz's Theorem on Isomorphisms between Algebraically Closed Fields. *Let* **F**, **A**, **E** *and* **B** *be algebraically closed fields so that* **F** *is a subfield of* **A** *and* **E** *is a subfield of* **B**. *Further suppose that* Φ *is an isomorphism from* **F** *onto* **E**. *If* **A** *and* **B** *have the same cardinality* κ *and* κ *is larger than the cardinality of* **F**, *then there is an isomorphism* Φ^* *from* **A** *onto* **B** *that extends* Φ. *Thus, any two uncountable algebraically closed fields of the same characteristic and the same cardinality are isomorphic.*

We require a couple of new notions and one more theorem that come from a branch of mathematical logic called model theory. We considered polynomials as certain formal strings of symbols (as opposed to treating them as certain kinds of functions, like they were treated in calculus). Mathematical logic expands on this idea. We can treat various mathematical statements as formal strings of symbols, but now we want to have symbols for some words commonly used in mathematics:

$\lor$	for "or"
$\land$	for "and"
$\implies$	for "implies"
$\neg$	for "not"
$\forall x$	for "for all x"
$\exists x$	for "there exists x"
$\approx$	for "equal"

We should also supply ourselves with symbols to denote the operations we use (like addition and multiplication), as well as symbols to name certain (maybe all) elements in sight. We also need an infinite supply of symbols to name variables. Then there will be a reasonable way to put these symbols together. For example,

$$\forall u \forall v \exists x [x^2 + ux + v \approx 0]$$

is a formulation of "Every monic polynomial of degree 2 has a root." and

$$\exists x_0 \exists x_1 \ldots \exists x_{n-1} \left[g_0(\bar{x}) \approx 0 \land \cdots \land g_{m-1}(\bar{x}) \approx 0 \land \neg f(\bar{x}) \approx 0 \right]$$

captures the assertion that a certain system of polynomial equations and inequations has a solution.

It is important to point out that the variables appearing in these formal expressions are meant to range over the *elements* of some domain of discourse—here usually a field—rather than, say, sets of such elements. Because of this restriction, we call such formalized expressions *elementary sentences.* Notice that in the last sentences displayed above, we had to have names for the elements of our field in order to write down the coefficients of the polynomials.

The new notion we need is the notion of *elementary extension.* In the context of fields it goes like this. Suppose **F** is a subfield of **K**. We will say that **K** is an elementary extension of **F**, provided every elementary sentence in the formal language of fields expanded by names of each element of F, that is true in **K** is also true in **F**.

The theorem we need from model theory was proved by Alfred Tarski in 1928, although he did not introduce the notion of elementary extension until after World War II. We only formulate the theorem for fields, but it is true for mathematical structures much more broadly.

Tarski's Upward Theorem. *Let* **F** *be any infinite field. Then* **F** *has an elementary extension of every greater size.*

While we do not give a proof of this theorem here, it should be noted that this theorem, in a more general guise, is customarily proven early on in any exposition of model theory. It is inviting to see that the standard methods of one branch of mathematics can play a key role in another branch.

Abraham Robinson actually proved a statement that is stronger than the Theorem on the Power of Algebraic Closures.

Robinson's Theorem on the Model Completeness of Algebraically Closed Fields. *Let* **F** *and* **E** *be algebraically closed fields. If* **E** *is an extension of* **F**, *then it is an elementary extension of* **F**.

Proof of the Theorem on the Power of Algebraic Closures. Suppose that **F** is a field and Γ is a finite system of polynomial equations and inequations that has a solution in a field **E** that extends **F**. Of course we can write down a sentence γ in the language of field enhanced by symbols to name every element of F, so that γ asserts that Γ has a solution. This means that γ is true in **E**. Let $\overline{\mathbf{E}}$ be the algebraic closure of **E**. So γ is also true in $\overline{\mathbf{E}}$ (in fact, the solution in E is still available). Let $\overline{\mathbf{F}}$ be the algebraic closure of **F**. We can regard $\overline{\mathbf{F}}$ as a subfield of $\overline{\mathbf{E}}$. So $\overline{\mathbf{E}}$ is an elementary extension of $\overline{\mathbf{F}}$. But this means the sentence γ is true in $\overline{\mathbf{F}}$. In other words, Γ has a solution in the algebraic closure of **F**. □

But how to prove that Model Completeness Theorem?

Proof of Robinson's Model Completeness Theorem of Algebraically Closed Fields. Let **F** and **E** be algebraically closed fields so that **E** extends **F**. Then both fields are infinite and have the same characteristic.

Now let κ be any uncountable cardinal at least as large as the sizes of **F** and **E**. Use Tarski's Upward Theorem to obtain elementary extensions $\mathbf{F}^*$ of **F** and $\mathbf{E}^*$ of **E** so that both of these elementary extensions are of size κ. Above we saw how to express "every polynomial of degree 2 has a root" and the generalization to polynomials of arbitrary degree is clear. So all these sentences will be true in **F** and in **E**. So they are also true in the elementary extensions. So these extensions are algebraically closed and of the same characteristic. By Steinitz, they are isomorphic by an isomorphism that fixes each element of F (i.e. that extends the identity map of **F**). Consider any sentence γ, perhaps using symbols to name elements of **F**, that is true in **E**. It must also be true in the elementary extension $\mathbf{E}^*$. So it is also true in $\mathbf{F}^*$, via the isomorphism. But then it must also be true in **F**, since $\mathbf{F}^*$ is an elementary extension of **F**. That is **E** is an elementary extension of **F**. □

Abraham Robinson called a class $\mathcal{K}$ of algebraic systems **model complete** provided whenever $\mathbf{A}, \mathbf{B} \in \mathcal{K}$ and **B** extends **A**, then **B** is an elementary extension of **A**.

Within a couple of years of Robinson's proof of the Nullstellensatz, Abraham Seidenberg gave a proof of the Theorem on the Power of Algebraic Closures using elimination theory. At around the same time, Robinson described how to use Hilbert's Nullstellensatz to prove the model completeness of algebraically closed fields.

Afterword

It is traditional, in works like this, at the beginning to justify the writing of yet another exposition of a subject already richly supplied with expositions. Not wanting to burden the first year graduate students with such things before they got to the actual mathematics, I am putting my excuses here at the end.

Of the fine books appropriate for first year graduate students, most follow the lead of B. L. van der Waerden's 1931 *Moderne Algebra*, which extends to more than 500 pages. More recent expositions, like David Dummit and Richard Foote's *Abstract Algebra*, Nathan Jacobson's *Basic Algebra I, II*, Serge Lang's *Algebra*, and Joseph Rotman's *Advanced Modern Algebra* come to about a thousand pages each. Even the shorter books like Michael Artin's *Algebra*, Pierre Grillet's *Abstract Algebra*, Thomas Hungerford's *Algebra*, and Martin Isaacs *Algebra: a graduate course* extend to more than 500 pages each. Every mathematician should have at least one of these on their bookshelf, since you never know how much you should really know! But I have never been able to get through more than 150 pages of this material in any semester. Even Larry Grove's *Algebra*, terse at 300 pages, is pushing it some.

My purpose: present those parts of algebra most likely to appear on PhD exams in the United States and keep it short.

By good fortune, Israel Herstein had just published his *Topics in Algebra* when it was my turn to take an undergraduate abstract algebra course. By another piece of luck, Nathan Jacobson's *Basic Algebra I* had just become available when it was my turn to teach this material for the first time. As with seemingly all expositions of algebra, the whole conceptual framework owes much to Emmy Noether. After her, the influence of Emil Artin is profound. I hope that some part of the penetrating insight of Noether, the elegrance of mind of Artin, and the good sense of Jacobson and Herstein can be found here.

In some ways this exposition is idiosyncratic. I put ring theory first, where most expositors have followed van der Waerden's lead and presented group theory first. Students come to the course richly supplied with examples of rings—making ring theory less of a leap than group theory, where the examples they know are few and not very typical. Group theory hardly enters into the development of ring theory at the beginning. But ring theory supplies an understanding of divisibility that is useful early in the development of group theory in getting hold of the theorems of Lagrange and Sylow. Also, I framed the initial notions like homomorphism from the point of view of the general theory of algebraic systems, just so they would apply equally well to rings, modules, and groups. With regret, I omitted an excursion into lattice theory.

Finally, after getting through Galois theory and the Artin-Schreier proof of the Fundamental Theorem of Algebra, there was always time in the spring semester for another topic or two. At times, I have included, for example, the rudiments of the theory of finite fields or some exposition of Hilbert's Theorem 90 or the work of Steinitz on algebraically closed fields. But the ones I like the best are the three topics you find here: Gauss's work on the constructibility of regular polygons, the Lindemann-Weierstrass Theorem (which has the transcendence of π as a consequence—something every mathematician knows but few of us have seen a proof), and Hilbert's Nullstellensatz, which establishes the link between commutative algebra and algebraic geometry. Concerning the latter, I first learned about Abraham Robinson's proof in 1969, while I was a graduate student. But in 1971, Abraham Robinson explained his approach to me over a beer at the Cheshire Cat, a bar off Euclid Avenue north of the campus in Berkeley. For me, anyway, it is a proof, as Uncle Paul would say, for THE BOOK. So I have put it here at the end of my little book.

George F. McNulty
June 2016

Bibliography

Algebra Texts: Get at Least One

Artin, Michael

(1991) *Algebra*. Englewood Cliffs, NJ: Prentice Hall Inc., pp. xviii+618. ISBN: 0-13-004763-5.

Birkhoff, Garrett and Saunders Mac Lane

(1965) *A survey of modern algebra*. Third edition. New York: The Macmillan Co., pp. x+437.

Dummit, David S. and Richard M. Foote

(2004) *Abstract algebra*. Third. Hoboken, NJ: John Wiley & Sons Inc., pp. xii+932. ISBN: 0-471-43334-9.

Grillet, Pierre Antoine

(2007) *Abstract algebra*. Second. Vol. 242. Graduate Texts in Mathematics. New York: Springer, pp. xii+669. ISBN: 978-0-387-71567-4.

Grove, Larry C.

(2004) *Algebra*. Reprint of the 1983 original, with an errata list on pp. xv–xvi. Mineola, NY: Dover Publications Inc., pp. xviii+299. ISBN: 0-486-43947-X.

Herstein, I. N.

(1975) *Topics in algebra*. Second. Lexington, Mass.: Xerox College Publishing, pp. xi+388.

Hungerford, Thomas W.

(1980) *Algebra*. Vol. 73. Graduate Texts in Mathematics. Reprint of the 1974 original. New York: Springer-Verlag, pp. xxiii+502. ISBN: 0-387-90518-9.

Isaacs, I. Martin

(2009) *Algebra: a graduate course*. Vol. 100. Graduate Studies in Mathematics. Reprint of the 1994 original. Providence, RI: American Mathematical Society, pp. xii+516. ISBN: 978-0-8218-4799-2.

Jacobson, Nathan

(1975a) *Lectures in abstract algebra*. Volume I: Basic concepts, Reprint of the 1951 edition, Graduate Texts in Mathematics, No. 30. New York: Springer-Verlag, pp. xii+217.

(1975b) *Lectures in abstract algebra*. Volume II: Linear algebra, Reprint of the 1953 edition [Van Nostrand, Toronto, Ont.], Graduate Texts in Mathematics, No. 31. New York: Springer-Verlag, pp. xii+280.

(1975c) *Lectures in abstract algebra*. Volume III: Theory of fields and Galois theory, Second corrected printing, Graduate Texts in Mathematics, No. 32. New York: Springer-Verlag, pp. xi+323.

(1985) *Basic algebra. I*. Second. New York: W. H. Freeman and Company, pp. xviii+499. ISBN: 0-7167-1480-9.

Jacobson, Nathan

(1989) *Basic algebra. II.* Second. New York: W. H. Freeman and Company, pp. xviii+686. ISBN: 0-7167-1933-9.

Lang, Serge

(2002) *Algebra.* third. Vol. 211. Graduate Texts in Mathematics. New York: Springer-Verlag, pp. xvi+914. ISBN: 0-387-95385-X. DOI: 10.1007/978-1-4613-0041-0. URL: http://dx.doi.org/10.1007/978-1-4613-0041-0.

Mac Lane, Saunders and Garrett Birkhoff

(1979) *Algebra.* Second. New York: Macmillan Inc., pp. xv+586. ISBN: 0-02-374310-7.

Rotman, Joseph J.

(2010) *Advanced modern algebra.* Vol. 114. Graduate Studies in Mathematics. Second edition. Providence, RI: American Mathematical Society, pp. xvi+1008. ISBN: 978-0-8218-4741-1.

Waerden, B. L. van der

(1943) *Moderne Algebra. Parts I and II.* G. E. Stechert and Co., New York, pp. 272+224.

(1991a) *Algebra. Vol. I.* Based in part on lectures by E. Artin and E. Noether, Translated from the seventh German edition by Fred Blum and John R. Schulenberger. New York: Springer-Verlag, pp. xiv+265. ISBN: 0-387-97424-5. DOI: 10.1007/978-1-4612-4420-2. URL: http://dx.doi.org/10.1007/978-1-4612-4420-2.

(1991b) *Algebra. Vol. II.* Based in part on lectures by E. Artin and E. Noether, Translated from the fifth German edition by John R. Schulenberger. New York: Springer-Verlag, pp. xii+284. ISBN: 0-387-97425-3.

THE SOURCES: YOU CAN TAKE A LOOK

Artin, Emil

(1965) *The collected papers of Emil Artin.* Edited by Serge Lang and John T. Tate. Addison–Wesley Publishing Co. Reading, Mass., xvi+560 pp. (2 plates).

(1998) *Galois theory.* second. Edited and with a supplemental chapter by Arthur N. Milgram. Mineola, NY: Dover Publications Inc., pp. iv+82. ISBN: 0-486-62342-4.

(2007a) *Algebra with Galois theory.* Vol. 15. Courant Lecture Notes in Mathematics. Notes by Albert A. Blank, Reprint of the 1947 original [ıt Modern higher algebra. Galois theory, Courant Inst. Math. Sci., New York]. Courant Institute of Mathematical Sciences, New York, pp. viii+126. ISBN: 978-0-8218-4129-7.

(2007b) *Exposition by Emil Artin: a selection.* Vol. 30. History of Mathematics. Edited by Michael Rosen. American Mathematical Society, Providence RI; London Mathematical Society, London, pp. x+346. ISBN: 978-0-8218-4172-3; 0-8218-4172-6.

Artin, Emil and Otto Schreier

(1927a) "Algebraische Konstruktion reeller Körper". In: *Abh. Math. Sem. Univ. Hamburg* 5.1, pp. 85–

99. ISSN: 0025-5858. DOI: 10.1007/BF02952512. URL: http://dx.doi.org/10.1007/BF02952512.

Artin, Emil and Otto Schreier

(1927b) “Eine Kennzeichnung der reell abgeschlossenen Körper”. In: *Abh. Math. Sem. Univ. Hamburg* 5.1, pp. 225–231. ISSN: 0025-5858. DOI: 10.1007/BF02952522. URL: http://dx.doi.org/10.1007/BF02952522.

Cayley, Arthur

(1854) “On the theory of groups as depending on the symbolic equation $\theta^n = 1$”. In: *Philosophical Magazine* 7, pp. 40–47.

(1858) “A Memoir of the Theory of Matrices”. In: *Phil. Trans.* 148.

Eisenstein, Gotthold

(1975a) *Mathematische Werke. Band I.* Chelsea Publishing Co., New York, pp. xiii+502.

(1975b) *Mathematische Werke. Band II.* Chelsea Publishing Co., New York, xiii+pp. 503–929.

Euclid of Alexandria

(2002) *Euclid’s Elements.* All thirteen books complete in one volume, The Thomas L. Heath translation, Edited by Dana Densmore. Green Lion Press, Santa Fe, NM, pp. xxx+499. ISBN: 1-888009-18-7; 1-888009-19-5.

Gauss, Carl Friedrich

(1986) *Disquisitiones arithmeticae.* Translated and with a preface by Arthur A. Clarke, Revised by William C. Waterhouse, Cornelius Greither and A. W. Grootendorst and with a preface by Waterhouse. Springer-Verlag, New York, pp. xx+472. ISBN: 0-387-96254-9.

Hamilton, William Rowan

(1967) *The mathematical papers of Sir William Rowan Hamilton. Vol. III: Algebra.* Edited by H. Halberstam and R. E. Ingram. Cunningham Memoir No. XV. Cambridge University Press, London-New York, xxiv+672 pp. (1 plate).

(1969) *Elements of quaternions. Vols. I, II.* Edited by Charles Jasper Joly. Chelsea Publishing Co., New York, Vol. I: xxxiii+594 pp., Vol. II: liv+502.

Hausdorff, Felix

(1914) *Grundzüge der Mengenlehre.* English translation 2nd ed., Chelsea, New York (1962). W. de Gruyter & Co., Leipzig.

Hermite, Charles

(1874) “Sur la fonction exponentielle”. In: *C. R. Acad. Sci. Paris* 77, pp. 18–24, 74–79, 226–233.

Hilbert, David

(1893) “Ueber die vollen Invariantensysteme”. In: *Math. Ann.* 42.3, pp. 313–373. ISSN: 0025-5831. DOI: 10.1007/BF01444162. URL: http://dx.doi.org/10.1007/BF01444162.

Hilbert, David

(1970) *Gesammelte Abhandlungen. Band II: Algebra, Invariantentheorie, Geometrie*. Zweite Auflage. Springer-Verlag, Berlin-New York, pp. viii+453.

(1978) *Hilbert's invariant theory papers*. Lie Groups: History, Frontiers and Applications, VIII. Translated from the German by Michael Ackerman, With comments by Robert Hermann. Math Sci Press, Brookline, Mass., pp. ix+336. ISBN: 0-915692-26-0.

(1993) *Theory of algebraic invariants*. Translated from the German and with a preface by Reinhard C. Laubenbacher, Edited and with an introduction by Bernd Sturmfels. Cambridge University Press, Cambridge, pp. xiv+191. ISBN: 0-521-44457-8; 0-521-44903-0.

König, Denes

(1927) "Über eine Schlussweise aus dem Endlichen ins Unendliche". In: *Acta Sci. Math. (Szeged)* 3, pp. 121–130.

Kronecker, Leopold

(1887) "Ein Fundamentalsatz der allgemeinen Arithmetik". In: *J. Reine Angew. Math.* 100, 490–510, also pp, 209–240 in *Kronecher's Werke III.*

(1968) *Leopold Kronecker's Werke. Bände I–V*. Herausgegeben auf Veranlassung der Königlich Preussischen Akademie der Wissenschaften von K. Hensel. Chelsea Publishing Co., New York, Vol. I: ix+483 pp. (1 plate), Vol. II: viii+540 pp., Vol. III, Part I: vii+473 pp., Part II: iii+215 pp. (Parts I and II bound as one), Vol. IV: x+508 pp., Vol. V: x+527.

Krull, Wolfgang

(1929) "Idealtheorie in Ringen ohne Endlichkeitsbedingung". In: *Math. Ann.* 101.1, pp. 729–744. ISSN: 0025-5831. DOI: 10.1007/BF01454872. URL: http://dx.doi.org/10.1007/BF01454872.

(1999) *Gesammelte Abhandlungen/Collected papers. Vol. 1, 2*. With biographical contributions by H. Schöneborn, H.-J. Nastold, J. Neukirch and Paulo Ribenboim, Edited and with a preface by Ribenboim. Walter de Gruyter & Co., Berlin, Vol. 1: xiv+822 pp, Vol. 2: pp. i–viii and 823–1730. ISBN: 3-11-012771-7.

Kuratowski, Casimir

(1922) "Une méthode d'élimination des nombres transfinis des raisonnements mathématiques". In: *Fundamenta Mathematicae* 3, pp. 76–108.

Lagrange, J.-L.

(1770/1771) "Réflexions sur le résolution des algébrique des équations". In: *Nouv. Mém. Acad. Berlin.*

Lindemann, F.

(1882) "Ueber die Zahl π". In: *Math. Ann.* 20.2, pp. 213–225. ISSN: 0025-5831. DOI: 10.1007/BF01446522. URL: http://dx.doi.org/10.1007/BF01446522.

Noether, Emmy

(1983) *Gesammelte Abhandlungen*. Edited and with an introduction by Nathan Jacobson, With an introductory address by P. S. Alexandrov [P. S. Aleksandrov]. Berlin: Springer-Verlag, viii+777 pp. (1 plate). ISBN: 3-540-11504-8.

Rabinowitsch, J. L.

(1930) "Zum Hilbertschen Nullstellensatz". In: *Math. Ann.* 102.1, p. 520. ISSN: 0025-5831. DOI: 10.1007/BF01782361. URL: http://dx.doi.org/10.1007/BF01782361.

Robinson, Abraham

(1951) *On the metamathematics of algebra*. Studies in Logic and the Foundations of Mathematics. North-Holland Publishing Co., Amsterdam, pp. ix+195.

(1954) "On predicates in algebraically closed fields". In: *J. Symbolic Logic* 19, pp. 103–114. ISSN: 0022-4812.

(1956) *Complete theories*. North-Holland Publishing Co., Amsterdam, pp. vii+129.

Steinitz, Ernst

(1910) "Algebraische Theorie der Körper". In: *J. Reine Angew. Math.* 137, pp. 167–309.

(1950) *Algebraische Theorie der Körper*. Chelsea Publishing Co., New York, N. Y., 176 pp. (1 plate).

Sylow, M. L.

(1872) "Théorèmes sur les groupes de substitutions". In: *Math. Ann.* 5.4, pp. 584–594. ISSN: 0025-5831. DOI: 10.1007/BF01442913. URL: http://dx.doi.org/10.1007/BF01442913.

Tarski, Alfred and Robert L. Vaught

(1958) "Arithmetical extensions of relational systems". In: *Compositio Math* 13, pp. 81–102. ISSN: 0010-437X.

Teichmüller, Oswald

(1939) "Braucht der Algebraiker das Auswahlaxiom?" In: *Deutsche Math.* 4, pp. 567–577.

Tukey, John W.

(1940) *Convergence and Uniformity in Topology*. Annals of Mathematics Studies, no. 2. Princeton University Press, Princeton, N. J., pp. ix+90.

Zariski, Oscar

(1947) "A new proof of Hilbert's Nullstellensatz". In: *Bull. Amer. Math. Soc.* 53, pp. 362–368. ISSN: 0002-9904.

Zassenhaus, Hans

(1934) "Zum Satz von Jordan-Hölder-Schreier". In: *Abh. Math. Sem. Univ. Hamburg* 10.1, pp. 106–108. ISSN: 0025-5858. DOI: 10.1007/BF02940667. URL: http://dx.doi.org/10.1007/BF02940667.

Zorn, Max

(1935) "A remark on method in transfinite algebra". In: *Bull. Amer. Math. Soc.* 41.10, pp. 667–670. ISSN: 0002-9904. DOI: 10.1090/S0002-9904-1935-06166-X. URL: http://dx.doi.org/10.1090/S0002-9904-1935-06166-X.

Index