Preamble

These notes are intended to supplement the lectures and make up for the lack of a textbook for the course this semester. As such, they will be updated continuously during the semester, in part to reflect the discussion in class, but also to preview material that will be presented and supply further explanation, examples, analysis, or discussion. The current state of the notes will be available in PDF format online at

http:\math.vanderbilt.edu\~tschantz\aanotes.pdf

While I will distribute portions of these notes in class, you will need to read ahead of class and get revised versions of previous notes by going online.

Your active participation is required in this class, in particular, I solicit any suggestions, corrections, or contributions to these notes that you may have. Using online, dynamically evolving, lecture notes instead of a static textbook is an experiment. I hope the result will be more concise, more relevant, and, in some ways, more useful than a textbook. On the other hand, these notes will be less polished, considerably less comprehensive, and not complete until the course has ended. Since I will be writing these notes as the course proceeds, I will be able to adjust the content, expanding on topics requiring more discussion and omitting topics we cannot get to. Writing these notes will take time, so please be patient. After reading these notes, you should have little difficulty finding more details in any textbook on the subject.
1 Introduction

This is a course in abstract algebra, emphasis on the abstract. This means we will be giving formal definitions and proofs of general theorems. You should already have had some exposure to proofs in linear algebra. In this course, you will learn to construct your own proofs.

1.1 The Mathematical Method

You have probably heard of the scientific method: observe a phenomena, construct a model of the phenomena, make predictions from the model, carry out experiments testing the predictions, refine and iterate. By carrying out experiments, the scientist assures himself, and others, that the understanding he develops of the world is firmly grounded in reality, and that the predictions derived from that understanding are reliable.

At first sight, what a mathematician does seems to be entirely different. A mathematician might be expected to write down assumptions or axioms for some system, state consequences of these assumptions, and supply proofs for these theorems. But this captures only part of what a mathematician does. Why does the mathematician go to the trouble of stating “obvious” facts and then supplying sometimes arcane proofs of theorems from such assumptions anyway?

In fact, mathematicians usually have one or more examples in mind when constructing a new theory. A theory (of first order logic) without any examples is inconsistent, meaning any theorem can be proven from the assumptions (this is essentially the completeness theorem of first order logic). The mathematician will identify the common important properties of his examples, giving an abstract definition of a class of similar examples. Theorems proved from these properties will then apply to any example satisfying the same assumptions (this is essentially the soundness theorem for first order logic). An applied mathematician may construct a theory based on observable properties of real world phenomena, expecting the theorems to hold for these phenomena to the extent that the phenomena satisfy the assumptions. A pure mathematician will usually be interested in examples from within mathematics which can be proved to satisfy the assumptions of his theory.

For example, applying this method to the notion of proof used by mathematicians, the theory of first order logic is a mathematical theory of logical statements and proofs. The soundness and completeness theorems for first
order logic are general statements about what can and cannot be proved from a set of assumptions in terms of what is or is not true in all examples satisfying that set of assumptions. Mathematical logic formalizes the procedures of proof commonly used in doing mathematics.

This is not a course in mathematical logic so we will accept a more usual informal style of proof. What is essential is that proofs be correct (in that they apply to all possible examples) and adequate (in that they produce the required conclusion from stated assumptions in clear accepted steps). That is, a proof must be an understandable explanation of why some consequence holds in all examples satisfying given assumptions. In particular, a proof will be written in English, expressing the logic of the argument, with additional symbols and formulas as needed. The reason for every step need not always be specified when obvious (to someone with experience), but it must be possible to supply reasons for each step if challenged. A real proof is a matter of give and take between author and reader, a communication of an important fact which may require questioning, explanation, challenge, and justification. If you don’t understand a proof in every detail, you should ask for clarification, and not accept the conclusion until you do understand the proof.

In this course we will consider numerous examples with interesting common properties. We will give abstract definitions of classes of such examples satisfying specific properties. We will use inductive reasoning\(^1\) to try and determine what conclusions follow in all of our examples and therefore what conclusions might follow in all examples satisfying the abstract definition, even those examples we have not imagined. Then we will attempt to find proofs of these conclusions from the stated definitions, or else to construct new examples satisfying the definitions but failing the expected conclusion. The examples we take are themselves often mathematically defined and so to work out whether a specific conclusion is true in a particular example may require more than simple computation but instead also require a proof. In the abstract, we may also develop algorithms for determining answers to certain questions that apply to any specific example, allowing us to utilize simple computations in the course of proving general facts.

Mathematics is the science of necessary consequence. By abstracting from

\(^1\)Inductive reasoning is the process of determining general properties by examining specific cases. Unless all the cases can be examined, inductive reasoning might lead to incorrect generalizations. Proof by induction is a valid method of deductive reasoning not to be confused with inductive reasoning.
the properties of specific examples to a general definition, the mathematician can use deductive reasoning to establish necessary properties of any example fitting the general definition. Making the right definitions and proving the best theorems from those definitions is the art of mathematics.

1.2 Basic Logic

There is nothing magic about the logic we will use in proving theorems. If I say of two statements $A$ and $B$ that “$A$ and $B$” holds I just mean $A$ holds and also $B$ holds. When I say “$A$ or $B$” holds, I mean that one or both of $A$ and $B$ holds (inclusive or), to say only one holds I must say “$A$ or $B$, but not both” (exclusive or).\footnote{In casual use, inclusive and exclusive or are often confused. Mathematically, we can afford casually saying “$A$ or $B$” when we mean exclusive or only when it is obvious that $A$ and $B$ cannot both hold.} If I say “if $A$, then $B$” holds I mean that in those examples where $A$ holds I also have that $B$ holds, i.e., in every case either $A$ fails (and the implication is said to be vacuously true) or $B$ holds. When $A$ and $B$ hold in exactly the same cases I often say “$A$ if and only if $B$”, abbreviated “$A$ iff $B$”, meaning if $A$, then $B$ and if $B$, then $A$, or put more simply either $A$ and $B$ both hold or neither $A$ nor $B$ holds. I sometimes say “it is not the case that $A$” or simply “not $A$” holds, meaning simply that $A$ fails to hold. Long combinations of statements can be hard to interpret unambiguously. In a course on mathematical logic we would have symbols for each logical connective and simply use parentheses to express the intended combination of statements. Instead, we will use various techniques of good exposition to clearly communicate the meaning of complex statements.

If basic statements are simple facts that can be checked, then working out the truth of compound sentences is a matter of calculation with truth values. Things become much more interesting when basic statements are true or false depending on a specific context, that is depending on what instance is taken for variable quantities. Variables are understood in ordinary English (e.g. “If a student scores 100 on the final, then they will get an A in the course.” most likely refers to the same student in both clauses), but we will not hesitate to use letters as explicit variables (as I have done in the last paragraph referring to statements $A$ and $B$). A statement such as $x + y = y + x$ might be asserted as true for all values of the variables (in some understood domain). On the other hand, the statement $2x + 1 = 6$ makes an assertion which is probably true or false depending on the value of $x$ (in the understood domain). If
we are to unambiguously understand the intent of a statement involving variables, then we need to know how the variables are to be interpreted. What we need are logical quantifiers.

Write $A(x)$ for a statement involving a variable $x$. There are two common interpretations of such a statement, usually implicitly understood. It may be that the assertion is that $A(x)$ holds no matter what value is assigned to $x$ (a universal or universally quantified statement). To make this explicit we say “for all $x, A(x)$” holds. Here we assume some domain of possible values of $x$ is understood. Alternatively, we might further restrict the specification of $x$ values for which the statement is asserted to hold by saying something like “for all integers $x, A(x)$” holds. Thus you probably interpreted $x+y = y+x$ as an identity, asserted to hold for all $x$ and $y$. This would be true if $x$ and $y$ range over numbers and $+$ means addition. It would be false if $x$ and $y$ range over strings and $+$ means concatenation (as in some programming languages) since even though $x+y = y+x$ is sometimes true, it fails when $x$ and $y$ are different single letters.

The second common interpretation is the assertion that $A(x)$ holds for at least one $x$. We say in this case “there exists an $x$ such that $A(x)$” holds or occasionally “for some $x, A(x)$” holds (an existential or existentially quantified statement). Thus when you consider $2x + 1 = 6$ you don’t expect the statement to be true for all $x$ and probably interpret the statement to mean there is an $x$ and you probably want to solve for that value of $x$. The statement “there exists an $x$ such that $2x+1 = 6$ is true if we care considering arithmetic of rational or real numbers since $x = 5/2$ makes $2x+1 = 6$ true. But the existential statement would be false if we were considering only integer $x$. To explicitly restrict the domain of a variable we might say something like “there exists an integer $x$ such that $A(x)$.

Mathematics gets especially interesting when dealing with statements that are not simply universal or existential. Universal statements are like identities, existential statements like equations to solve. But a combination of universal and existential requires a deeper understanding. The first example most students see is the definition of limit, “for every $\epsilon$, there exists a $\delta$ such that,...”. In such statements the order of the quantifiers is important ($\delta$ will usually depend on $\epsilon$ so the problem in proving a limit using the definition is in finding a formula or procedure for determining $\delta$ from $\epsilon$ and establishing that this $\delta$ satisfies the definition of limit). While it is informally reasonable to express quantifiers as “$A(x)$, for all $x$” or “$A(x)$, for some $x$”, it is preferable to keep quantifiers specified up front whenever there is any
chance for confusion (e.g., we might have trouble interpreting “there exists a \( \delta \) such that \( \ldots \) for any \( \epsilon \)”). The shorthand notation for “for all \( x \), \( A(x) \)” is \( \forall x \ A(x) \) and for “there exists \( x \) such that \( A(x) \)” is \( \exists x \ A(x) \), but one should never use the shorthand notation at the expense of clear exposition.

Finally a few words on definitions. A new terminology or notation is introduced by a definition. Such a definition does not add to the assumptions of the theory, but instead specifies how the new terminology is to be understood in terms of more basic concepts. A definition introducing the notion of widgets might start “An \( x \) is a widget if \( \ldots \)”. Because “widget” is a new term and this is its definition, the definition must be understood as an “iff” statement even though one ordinarily writes simply “if”. Moreover the statement is understood to be universally quantified, i.e., “for all \( x \), we say \( x \) is a widget iff \( \ldots \)”. To define a new function \( f \) we may give a formula for \( f \), writing \( f(x) = t \) and again meaning for all (possible) \( x \) if not otherwise specified. We can also define a new function by specifying the property the function has by saying “\( f(x) = y \) if \( A(x, y) \) holds” (meaning iff and for all \( x \) and \( y \) provided we can also show i) for every \( x \) there exists a \( y \) such that \( A(x, y) \) holds, and ii) for every \( x \) and \( y_1 \) and \( y_2 \), if both \( A(x, y_1) \) and \( A(x, y_2) \) hold, then \( y_1 = y_2 \). Part i) verifies the existence of a value for \( f(x) \) and part ii) shows that this value is uniquely defined. These conditions are sometimes combined by saying for every \( x \) there exists a unique \( y \) such that \( A(x, y) \). If for all \( x \) there is a unique \( y \) satisfying the condition then we say \( f \) is well-defined.

### 1.3 Proofs

Now that our logical language has been clarified, it is possible to state some simple logical rules and strategies for proving statements. We assume certain facts are given. A proof explains, in a stepwise fashion, how other statements follow from these assumptions, each statement following from the given information or from earlier conclusions. The following rules should be obvious, you may think of other variations. Informal proofs may omit the majority of obvious steps.

- From “\( A \) and \( B \)” conclude “\( A \)” (alternatively “\( B \)”).
- From “\( A \)” and “\( B \)” (separately) conclude “\( A \) and \( B \)”.
- From “\( A \)” (alternatively “\( B \)”)) conclude “\( A \) or \( B \)”.
• From “A or B” and “not B” conclude “A”.
• From “A” and “if A, then B” conclude “B”.
• From “not B” and “if A, then B” conclude “not A”.
• From “B or not A” conclude “if A, then B”.
• From “A iff B” conclude “if A, then B” (alternatively “if B, then A”).
• From “if A, then B” and “if B, then A” conclude “A iff B”.
• From “not not A” conclude A.
• From “A” conclude “not not A”.

A less obvious idea in creating a proof is the technique of making certain intermediate assumptions, for the sake of argument as it were. These intermediate assumptions are for the purpose of creating a subproof within the proof. But these intermediate assumptions are not assumptions of the main proof, only temporary hypotheticals that are discharged in making certain conclusions.

• From a proof assuming “A” and concluding B, conclude “if A, then B” (and where A is not then an assumption needed for this conclusion but is included in the statement).
• From a proof assuming “not B” and concluding “not A”, conclude “if A, then B”.
• (Proof by contradiction) From a proof assuming “not A” and concluding “B and not B” (a contradiction), conclude “A”.
• (Proof by cases) From a proof assuming “A” and concluding “B” and a proof assuming “not A” and concluding “B”, conclude “B” (in either case).

To deal with quantifiers we need a few more rules.

• From “for all x, A(x)” and any expression t, conclude “A(t)” (i.e., that “A(x)” holds with x replaced by the expression t and with the additional proviso that the variables in t are different from quantified variables in A(x)).
• From “there exists $x$, $A(x)$” conclude “$A(c)$” where $c$ is a new symbol (naming the value that is asserted to exist).

• From a proof of “$A(c)$” where $c$ is a variable not appearing in any of the assumptions, conclude “for all $x$, $A(x)$”.

• From a proof of “$A(t)$” for some expression $t$, conclude “there exists an $x$, $A(x)$” (with the additional proviso that variables in $t$ are different from quantified variables in $A(x)$).

Additional rules specify how to treat equality.

• From “$s = t$” conclude “$f(s) = f(t)$”.

• From “$s = t$” conclude “$A(s)$ iff $A(t)$”.

• For any expression $t$, conclude “$t = t$”.

• From “$s = t$” conclude “$t = s$”.

• From “$r = s$” and “$s = t$” conclude “$r = t$”.

Most of these rules can be turned around to suggest strategies for constructing proofs. Proofs are generally written in a one-way manner, proceeding from assumptions to conclusion in a step-by-step fashion. However, to discover a proof one usually explores what can be proved from the assumptions at the same time as working backward from the desired conclusion to see what would be needed to be proved to establish the conclusion. These then are some of the strategies for proving theorems you might use.

• To prove “$A$ and $B$”, prove “$A$” and “$B$”. To use “$A$ and $B$”, you may use “$A$” and “$B$” (write out all the basic facts you have available or might need).

• To prove “$A$ or $B$”, see if one or the other of “$A$” or “$B$” might hold and be provable (rarely that easy), or assume “not $A$” and prove “$B$” (or vice-versa), or else assume both “not $A$” and “not $B$” and prove any contradiction (“$C$ and not $C$”). To use “$A$ or $B$” to prove “$C$”, see if you can prove “$C$” assuming “$A$” and also prove “$C$” assuming “$B$” (proof by cases in another guise).
To prove “if \( A \), then \( B \)”, assume “\( A \)” and prove “\( B \)”, or assume “not \( A \)” and prove “not \( B \)”, or assume “\( A \)” and “not \( B \)” and prove a contradiction. To use “if \( A \), then \( B \)” see if you can prove “\( A \)” and conclude then also “\( B \)”, or see if you can prove “not \( B \)” and conclude then “not \( A \)”. 

To prove “\( A \iff B \)”, assume “\( A \)” and prove “\( B \)” and also assume “\( B \)” and prove “\( A \)” (or sometimes a shortcut of connecting other equivalent conditions is possible). To use “\( A \iff B \)”, see if \( A \) or \( B \) (or one of their negations) are known or needed and replace one with the other.

To prove any statement “\( A \)”, assume “not \( A \)” and prove a contradiction (not always any easier).

To prove “for all \( x \), \( A(x) \)”, take a new symbol \( c \), not appearing (unquantified) in any assumptions, and prove “\( A(c) \)”. To use “for all \( x \), \( A(x) \)”, consider which instances “\( A(t) \)” might be useful.

To prove “there exists \( x \) such that \( A(x) \)”, give a particular \( t \) and prove “\( A(t) \)”. To use “there exists \( x \) such that \( A(x) \)”, take a new symbol \( c \) naming an element and assume “\( A(c) \)”.

The rules and strategies outlined above are not to be construed too strictly. In informal proofs, simpler steps are often skipped or combined (but can always be detailed if requested), symbols may be reused in different contexts when no confusion arises in the interests of good or consistent notation, analogous arguments mentioned or omitted, and of course earlier theorems may also be applied. What matters is that the author of the proof communicates with the reader the essential steps that lead to the conclusion, understanding that the reader will be able to supply simpler steps of the argument if the intent and plan of the proof is clear.

1.4 Mathematical Foundations

Underlying nearly all of the mathematics is the theory of sets. A set is determined by its elements. Often, a set is given by specifying the condition satisfied by, and only by, its elements, subject to certain restrictions on how to sets can be formed. Thus, informally, \( \{ x : A(x) \} \) is a notation for a set \( S \) such that for all \( x \), \( x \in S \) iff \( A(x) \) holds.\(^3\) The precise rules for constructing

\(^3\)Not every definition of this form determines a set however!
sets will not (generally) concern us in this course. However it is assumed that
the student is familiar with basic operation on sets, union $A \cup B$, intersection
$A \cap B$, difference $A - B$, familiar with Cartesian product $A \times B$ and Cartesian
powers $A^n$, and familiar with the concept of subset $A \subseteq B$.

We write $f : A \rightarrow B$ to express that $f$ is a function from $A$ into $B$, i.e.,
that to every element $a \in A$ there identified an element $f(a) \in B$. The
domain of $f$ is $A$, the image of $f$ is the set $f(A) = \{f(a) : a \in A\}$ of
elements. If $f : A \rightarrow B$ and $g : B \rightarrow C$ are functions, then the composition
$g \circ f : A \rightarrow C$ is the function such that for all $a \in A$, $(g \circ f)(a) = g(f(a))$. For
reasons having to do with reading from left to right, we will occasionally like
our functions to act on the right instead of the left, that is we would like a
notation such as $(a)f$ or $af$ (parentheses optional) for the result of applying
$f$ to an element $a \in A$. Then the same composition of $f : A \rightarrow B$ followed
by $g : B \rightarrow C$ can be written as $fg$ (left-to-right composition denoted by
concatenation instead of $\circ$) where $a(fg) = (af)g$ or simply $afg$. In this
notation, composing a string of functions reads correctly from left-to-right,
and while there’s not often a good reason to use such non-standard notation,
it will occasionally make more sense (and is in fact essentially standard in
certain special cases).

A function $f : A \rightarrow B$ is injective (is an injection, 1-1, or one-to-one) if
for any $a_1, a_2 \in A$, if $f(a_1) = f(a_2)$, then $a_1 = a_2$, (equivalently, different
elements of $A$ always map to different elements of $B$). A function $f : A \rightarrow B$
is surjective (is a surjection, or onto) if for every $b \in B$ there is some $a \in A$
such that $f(a) = b$. A function is bijective (is a bijection, or is 1-1 and onto)
if it is injective and surjective.

For $n$ a positive integer, an $n$-ary operation on a set $A$ is a function
from $A^n$ back into $A$, i.e., an $n$-ary operation takes $n$ elements from $A$ and
evaluates to another element of $A$. The rank (or arity) of an operation is the
number $n$ of arguments the operation takes. Unary (1-ary) operations map
$A$ to $A$, binary (2-ary) operations map $A^2$ to $A$ and are often written with
special symbols (like $+$, $\cdot$, $\ast$, $\circ$) using infix notation (i.e., $x + y$ or $x \ast y$ etc.),
and ternary (3-ary) and higher rank operations rarely arise. An (identified
constant) element in the set can be thought of as a 0-ary operation on a set.

In algebra, we are interested in sets with various combinations of operations
(and constants). Such structures are should be familiar.

- $\mathbb{N} = \{0, 1, 2, \ldots\}$ the set of natural numbers, with $+$ addition, $\cdot$
multiplication, 0 and 1 distinguished constants.
• $\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2 \ldots\}$ the set of integers, with the above operations and $−$ (unary) negation (and also often binary subtraction).

• $\mathbb{Q}$ the set of rational numbers with the above operations and a unary (partial) operation denoted by $(\ )^{-1}$ for reciprocals of nonzero rationals only (or a binary division operation $/$ with nonzero denominator).

• $\mathbb{R}$ the set of real numbers with similar operations.

• $\mathbb{C}$ the set of complex numbers $x + yi$ for $x, y \in \mathbb{R}$ and $i = \sqrt{-1}$, with similar operations.

• $n \times n$ matrices with real entries, taking addition and matrix multiplication, negatives, and inverses of matrices with nonzero determinant.

Additional examples are modular arithmetic and polynomials.

1.5 Algebras

In (universal) algebra we study algebras of the following very general sort.

Definition 1.5.1 An algebra $\mathbf{A} = \langle A; f_1, f_2, \ldots, f_n \rangle$ is a nonempty set $A$ together with a sequence $f_i$ of (basic) operations on that set. The similarity type of an algebra is the sequence of ranks of the operations of the algebra. Algebras are similar if they have the same similarity type (i.e. if the ranks of operations of one algebra are the same as the ranks for the other algebra, in order).

Corresponding operations of similar algebras are often denoted by the same symbols so that formulas in those symbols can apply to any one of similar algebras with the symbols denoting the operations of that algebra. Thus we can use $+$ as addition of integers and of rational numbers (and this is especially convenient because the integers are rational numbers and the $+$ operation on the integers is the same as the $+$ operation on rationals restricted to integer arguments), but also we can use $+$ as an operation on $n \times n$ matrices. Each of these systems satisfies the identity $x + y = y + x$, so the choice of $+$ to represent matrix addition (and not matrix multiplication say) is good for another reason. In studying algebras we seek to discover such common properties of our examples, abstract these as defining a certain
class of algebras, and study this class of algebras, proving theorems for all algebras of this class.

We will concentrate our study on three classes of algebras, groups, rings, and fields. Algebras of these sorts have uses throughout mathematics, show up in a variety of ways in nature, and have a number of uses in science and technology. These applications are a secondary concern in this course however, where we will instead concentrate on developing the theory and practicing the mathematical techniques used in algebra and throughout mathematics.

1.6 Problems

Problem 1.6.1 Give the (precise, formal) definition of when \( \lim_{x \to a} f(x) = L \) from calculus. Explain how the quantifiers \((\varepsilon, \delta, x)\) are to be interpreted and in what order. Is the definition of limit always well-defined (does there exist a limit and, if so, is it unique)?

Problem 1.6.2 Give another definition from calculus or linear algebra of similar complexity to the definition of limit and perform the same analysis of quantifiers and well-definition.

Problem 1.6.3 To define a set \( S \), it is enough to specify exactly when an \( x \) belongs to \( S \). For sets \( A \) and \( B \), give definitions of \( A \cup B \), \( A \cap B \), \( A - B \), \( A \times B \), and \( A^n \), and when \( A \subseteq B \).

Problem 1.6.4 To show sets are equal, it is enough to show that they have exactly the same elements. Prove that for any sets \( A \), \( B \), and \( C \), \( A \cup B = B \cup A \), \( A \cap B = B \cap A \), \( A \cup (B \cup C) = (A \cup B) \cup C \) and \( A \cap (B \cap C) = (A \cap B) \cap C \).

Problem 1.6.5 Prove that for any sets \( A \), \( B \), and \( C \), \( A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \) and \( A \cup (B \cap C) = (A \cup B) \cap (A \cup C) \).

Problem 1.6.6 Prove that for any sets \( A \), \( B \), and \( C \), if \( A \subseteq B \) and \( B \subseteq C \), then \( A \subseteq C \).

Problem 1.6.7 Prove that for any sets \( A \) and \( B \), \( A \subseteq A \cup B \) and \( A \cap B \subseteq A \). Prove that for any sets \( A \), \( B \), and \( C \), \( A \cup B \subseteq C \) if \( A \subseteq C \) and \( B \subseteq C \). Prove that for any sets \( A \), \( B \), and \( C \), \( C \subseteq A \cap B \) if \( C \subseteq A \) and \( C \subseteq B \).
Problem 1.6.8 Suggest and prove properties relating set difference \((A - B)\) to unions and intersections and to subsets. For example, \(A - (B \cup C) = (A - B) - C = (A - B) \cap (A - C)\).

Problem 1.6.9 Suppose \(f\), \(g\), and \(h\) are unary operations on a set \(A\). Show that \(f \circ (g \circ h) = (f \circ g) \circ h\). Give an example showing that it need not be the case that \(f \circ g = g \circ f\).

Problem 1.6.10 Show that if \(f : A \rightarrow B\) and \(g : B \rightarrow C\) are injective functions then \(g \circ f\) is injective. Show that if \(f : A \rightarrow B\) and \(g : B \rightarrow C\) are surjective functions, then \(g \circ f\) is surjective.

Problem 1.6.11 Let \(A\) be the set of all subsets of a set \(S\). Write + for \(\cup\), \(\cdot\) for \(\cap\), 0 for the empty set, and 1 for the set \(S\). Consider the algebra \(A = \langle A; +, \cdot, 0, 1 \rangle\). What properties does \(A\) share with ordinary arithmetic on the integers \(\langle \mathbb{Z}; +, \cdot, 0, 1 \rangle\)? What properties distinguish \(A\) from arithmetic on integers?

Problem 1.6.12 Let \(A\) be the set of all bijective functions from \(X = \{1, 2, \ldots, n\}\) into \(X\). Let \(i_X\) denote the identity function on \(S\) defined for \(x \in X\) by \(i_X(x) = x\). Show that \(i_X\) is an element of \(A\) and \(\circ\) is an operation on \(A\). Let \(A = \langle A; \circ, i_X \rangle\). What properties does \(A\) share with multiplication on the positive rational numbers \(\langle \mathbb{Q}^+; \cdot, 1 \rangle\) if we identify \(\circ\) with multiplication and \(i_X\) with 1? What properties distinguish these algebras?

Problem 1.6.13 (From my research.) A special class of algebras I have studied are those having a single binary operation (denoted by \(\cdot\) or simply by concatenation say) satisfying the identities \(xx = x\) and \((xy)(yx) = y\) with no further assumptions (in particular, not necessarily even associative). As an example of such an algebra I took the integers modulo 5, \(A = \{0, 1, 2, 3, 4\}\) and defined \(xy\) by \(xy = 2x - y \pmod{5}\). Verify the two identities defining my special class of algebras for the algebra \(A = \langle A; \cdot \rangle\). Show that this operation on \(A\) is not associative.

I discovered that there always exists a ternary operation \(p(x, y, z)\) on algebras in this special class satisfying the identities \(p(x, y, y) = x\) and \(p(x, x, y) = y\). The computer program I used found the formula

\[
p(x, y, z) = (((y(zy))(yx)z))((yz)(x)(zy)z)(x(zy))(zy)(xy)(xy)z)
\]

for such an operation. Verify that the identities are satisfied by this operation.
1.7 Sample Proofs

Here are some sample proofs from the lectures. We construct such proofs by drawing pictures (when possible), testing examples, working forward from the assumptions and backwards from the conclusions at the same time, and applying definitions to analyze compound notions in terms of more elementary notions. When we understand reasons why something should be true, we can then work out a complete, clear, linear proof of why it is true, but one shouldn’t expect such a proof to be immediately apparent (at least not often).

Definition 1.7.1 For sets $A$ and $B$, we define $A - B$ so that for any $x$, $x \in A - B$ iff $x \in A$ and $x \notin B$. (This determines $A - B$ uniquely because sets are determined by their elements. Existence of such a set is a basic consequence of an axiom of set theory.) We define $A + B$ (the symmetric difference of $A$ and $B$) so that for any $x$, $x \in A + B$ iff either $x \in A$ and $x \notin B$, or else $x \in B$ and $x \notin A$. (Alternatively, $A + B = (A - B) \cup (B - A)$)

Theorem 1.7.1 For any sets $A$ and $B$, $A + B = B + A$.

Proof: Sets are equal if they have the same elements. An $x$ is in $A + B$ iff $x$ is in $A$ but not $B$ or else in $B$ but not $A$. But this is the same as saying $x$ is in $B$ but not $A$ or else in $A$ but not $B$, i.e., iff $x \in B + A$ reading the definition with the roles of $A$ and $B$ reversed. That is an $x$ belongs to $A + B$ iff it belongs to $B + A$ and so $A + B = B + A$. □

Theorem 1.7.2 For any sets $A$, $B$, and $C$, $A + (B + C) = (A + B) + C$.

Proof: For any $x$, $x \in A + (B + C)$ iff either $x \in A$ and $x \notin B + C$ or else $x \in B + C$ and $x \notin A$. In the first case, $x \notin B + C$ is true iff either $x$ is in both $B$ and $C$ or else $x$ is in neither $B$ nor $C$. So the first case would mean either $x \in A$, $B$, and $C$, or else $x \in A$, $x \notin B$ and $x \notin C$. In the second case, $x \notin A$ and either $x \in B$, $x \notin C$, or else $x \notin B$, $x \in C$, i.e., the possibilities are either $x \notin A$, $x \in B$, $x \notin C$, or else $x \notin A$, $x \notin B$ and $x \in C$.

Analyzing $x \in (A + B) + C$ similarly, we have $x \in A + B$, $x \notin C$ or else $x \notin A + B$, $x \in C$. In the first case, $x \in A$, $x \notin B$, $x \notin C$, or else $x \notin A$, $x \in B$, $x \notin C$. In the second case, $x \in A$, $x \in B$, $x \in C$, or else $x \notin A$, $x \notin B$, $x \in C$. But these amount to the same four possibilities as we ended
up with for \( x \in A + (B + C) \). That is \( x \in A + (B + C) \) iff \( x \in (A + B) + C \) so \( A + (B + C) = (A + B) + C \). \( \square \)

In searching for additional facts, you might have guessed that (also in analogy with arithmetic) \((A + B) - B = A\) (or maybe not). If you try to prove such a theorem either you’ll be frustrated or make a mistake. In trying to find a proof you might also discover instead a reason why this is not true. In fact the simplest way of showing a general statement is not true is to come up with an example where it fails. In this case, \( A = \{1, 2\} \) and \( B = \{2, 3\} \) gives \((A + B) - B = \{1, 3\} - \{2, 3\} = \{1\} \neq B\).

On the other hand, \( A + \emptyset = A \) and \( A + A = \emptyset \) are easily shown so we can calculate \((A + B) + B = A + (B + B) = A + \emptyset = A\). Thus + plays both the role of addition and subtraction in analogy with arithmetic, \( \emptyset \) plays the role of additive identity \( 0 \), and there is an entirely different identity in \( A + A = \emptyset \) implying each set is (with respect to symmetric difference) its own additive inverse.

A universal statement (such as the last about all sets \( A, B, \) and \( C \)) requires considering an arbitrary instance. A statement involving an existential quantifier requires finding or constructing or otherwise determining an object that satisfies a condition, perhaps dependent on an arbitrary instance corresponding to an enclosing universal quantifier.

**Theorem 1.7.3** For any functions \( f : A \to B \) and \( g : B \to C \), if \( f \) and \( g \) are surjective functions, then \( g \circ f \) is surjective.

**Proof:** Assume \( f : A \to B \) and \( g : B \to C \) are surjective functions. To show that \( g \circ f : A \to C \) is surjective, we need to show that for every \( c \in C \) there is an \( a \in A \) such that \((g \circ f)(a) = c\). Suppose \( c \in C \) is given. Then since \( g \) is surjective there exists a \( b \in B \) such that \( g(b) = c \). For this \( b \in B \), since \( f \) is surjective, there is an \( a \in A \) such that \( f(a) = b \). Now \((g \circ f)(a) = g(f(a)) = g(b) = c\), showing there is an \( a \in A \) mapping under \( g \circ f \) to each given \( c \in C \), so \( g \circ f \) is surjective. \( \square \)

**Theorem 1.7.4** For any functions \( f : A \to B \) and \( g : B \to C \), if \( f \) and \( g \) are injective functions, then \( g \circ f \) is injective.

**Proof:** Assume \( f : A \to B \) and \( g : B \to C \) are injective functions. To show \( g \circ f : A \to C \) is injective, we need to show that for any \( a_1, a_2 \in A \), if \((g \circ f)(a_1) = (g \circ f)(a_2)\) then \( a_1 = a_2 \). So assume \( a_1, a_2 \in A \) and \((g \circ f)(a_1) =
\((g \circ f)(a_2)\). Then \(g(f(a_1)) = g(f(a_2))\) and since \(g\) is injective we get that \(f(a_1) = f(a_2)\). But then since \(f\) is surjective we get that \(a_1 = a_2\). Hence \((g \circ f)(a_1) = (g \circ f)(a_2)\) implies \(a_1 = a_2\) so \(g \circ f\) is injective. \(\square\)

**Definition 1.7.2** For any set \(A\), the identity function on \(A\), denoted \(i_A\), is the function \(i_A : A \to A\) defined by, for any \(a \in A\), \(i_A(a) = a\). (Since functions are determined by their domains and values, such a function is unique, while existence is again a basic consequence of the axioms of set theory.)

**Theorem 1.7.5** For any bijective function \(f : A \to B\) there is a unique function \(g : B \to A\) such that \(g \circ f = i_A\) and \(f \circ g = i_B\). (The \(g\) determined by this theorem is usually denoted \(f^{-1}\).)

**Proof:** Suppose a bijective function \(f : A \to B\) is given. First we establish the existence of a function \(g\) satisfying the conditions. We define \(g : B \to A\) by determining a value for \(g(b)\) for each \(b \in B\). Given \(b \in B\), since \(f\) is surjective, there exists an \(a \in A\) such that \(f(a) = b\). If there were different \(a_1, a_2 \in A\) such that \(f(a_1) = f(a_2) = b\), then \(f\) would not be injective. Hence the \(a \in A\) with \(f(a) = b\) is unique and we may define \(g(b) = a\). (Alternatively, we may simply choose, for each \(b\) any particular \(a \in A\) with \(f(a) = b\), but this would technically require the axiom of choice to make a possibly infinite number of choices all at once to define \(g\).)

Now we claim that \(f \circ g = i_B\). For any \(b \in B\), \(g(b)\) is defined to be an element \(a \in A\) with \(f(a) = b\). Hence \((f \circ g)(b) = f(g(b)) = f(a) = b = i_B(b)\) as required.

Next we claim that \(g \circ f = i_A\). For any \(a' \in A\), \(f(a')\) is an element in \(B\), say \(f(a') = b\). From the definition of \(g\), \(g(b) = a\) for an element \(a \in A\) with \(f(a) = b\). But then \(f(a) = b = f(a')\), so since \(f\) is injective, \(a = a'\). Hence \((g \circ f)(a') = g(f(a')) = g(b) = a = a' = i_A(a')\) as required. Hence the defined \(g\) satisfies the two conditions of the conclusion of the theorem and the existence of such a \(g\) is established.

Finally we need to show that such a \(g\) is unique. Suppose \(g_1, g_2 : B \to A\) both satisfy the conditions, i.e., \(g_1 \circ f = i_A\), \(f \circ g_1 = i_B\), \(g_2 \circ f = i_A\), and \(f \circ g_2 = i_B\). To show \(g_1\) and \(g_2\) are in fact the same function, consider any \(b \in B\). Then \(f(g_1(b)) = b = f(g_2(b))\), and since \(f\) is injective, \(g_1(b) = g_2(b)\). Hence \(g_1 = g_2\) and uniqueness is established. \(\square\)
2 Groups

The notion of a group turns out to be a useful abstraction with numerous instances. Groups are defined by simple identities, but the resulting theory allows strong conclusions about the structure and properties of groups. So that we know what to check in our examples, we give the definition of group first. (Though this is the opposite of the model of discovering abstractions by considering examples, it saves us time by not having to rediscover the more useful abstractions.)

2.1 Definitions

We give two variants of the definition of a group. These definitions are seen to be equivalent except for what operations are taken as basic.

Definition 2.1.1 (Version 1) A group \( G \) is a set \( G \) together with a binary operation, (usually) denoted by \( \cdot \) or simply concatenation, such that:

1. (Associativity) for all \( x, y, z \in G \), \( x(yz) = (xy)z \);
2. (Identity) there is an element \( e \in G \) such that for any \( x \in G \), \( ex = xe = x \); and
3. (Inverses) for \( e \) as above, and for any \( x \in G \), there is a \( y \in G \) such that \( xy = yx = e \).

Such an element \( e \in G \) is called an identity element of \( G \). An element \( y \in G \) with \( xy = yx = e \) is called an inverse of \( x \).

Theorem 2.1.1 In any group \( G \), there is a unique identity element, and each element \( x \in G \) has a unique inverse.

Proof: Suppose \( e_1 \) and \( e_2 \) are both identities for \( G \), i.e., for all \( x \in G \), \( e_1x = xe_1 = x \) and \( e_2x = xe_2 = x \). Consider \( e_1e_2 \). Since \( e_1 \) is an identity, \( e_1e_2 = e_2 \). But also since \( e_2 \) is an identity, \( e_1e_2 = e_1 \). Hence \( e_1 = e_1e_2 = e_2 \) and there is only one identity in a group.

Take any \( x \in G \) and suppose \( y_1 \) and \( y_2 \) are inverses of \( x \), i.e., \( xy_1 = y_1x = e \) and \( xy_2 = y_2x = e \). Consider \( y_1(xy_2) = (y_1x)y_2 \). Since \( y_1 \) is an inverse of \( x \), \( y_1x = e \) so \( (y_1x)y_2 = ey_2 = y_2 \). But also since \( y_2 \) is an inverse of \( x \),
$y_1(xy_2) = y_1e = y_1$. Hence $y_1 = y_1e = y_1(xy_2) = (y_1x)y_2 = ey_2 = y_2$ and there is only one inverse of any given element $x \in G$. □

The (unique) identity element of a group with (multiplication) operation $\cdot$ is often denoted by $1$ (or $1_G$ if there’s any possibility of confusion), while the (unique) inverse of an element $x$ is denoted by $x^{-1}$. We thus may define the identity and inverse operation in any group. Alternatively we may take these as basic operations of the group structure.

**Definition 2.1.2 (Version 2)** A group $G$ is an algebra $\langle G; \cdot, (\cdot)^{-1}, 1 \rangle$ such that

1. for all $x, y, z \in G$, $x(yz) = (xy)z$;
2. for all $x \in G$, $1x = x1 = x$; and
3. for all $x \in G$, $x(x^{-1}) = (x^{-1})x = 1$.

We take $(\cdot)^{-1}$ to be of higher precedence than $\cdot$ so that $xy^{-1}$ means $x(y^{-1})$ instead of $(xy)^{-1}$. Because of associativity we can write $xyz$ to mean either of $x(yz) = (xy)z$. In general, any product of factors can be written without parentheses since however we put parentheses in will give the same result. This is a basic theorem of ordinary arithmetic, but since we are working in an abstract setting now it is worthy of stating and proving such a theorem.

**Theorem 2.1.2 (Generalized associativity)** For any group $G$ and any elements $x_1, x_2, \ldots, x_n \in G$, any way of parenthesizing the product $x_1x_2\ldots x_n$ evaluates to the same value in $G$.

**Proof:** We proceed by (strong) induction on $n$, the number of factors in the product. If $n = 1, 2$ there is only one way of putting in parentheses and evaluating so there is nothing to prove. If $n = 3$, then there are two possible ways of evaluating $x_1x_2x_3$, either as $x_1(x_2x_3)$ or as $(x_1x_2)x_3$. But these are equivalent by associativity in $G$ (basis cases). Assume $n > 3$ and assume the result is true for products of fewer than $n$ factors (the induction hypothesis). We show that any way of putting parentheses in $x_1x_2\ldots x_n$ evaluates to the same as $(x_1x_2\ldots x_{n-1})x_n$ where the product of the first $n-1$ factors is independent of the parentheses taken, by induction hypotheses. Fully parenthesizing the product of $n$ factors, we evaluate an expression where the last multiplication is of the product of the first $k$ factors and the product of the last $n-k$ factors, for some $k$, $1 \leq k < n$, i.e., we evaluate $(x_1x_2\ldots x_k)(x_{k+1}\ldots x_n)$ where
again the product of the first \(k\) and last \(n-k\) factors is independent of the parentheses taken, by induction hypothesis. If \(k=n-1\) this is the product we are comparing to. Otherwise \(k<n-1\) and we can write

\[
(x_1x_2\ldots x_k)(x_{k+1}\ldots x_n) = (x_1x_2\ldots x_k)((x_{k+1}\ldots x_{n-1})x_n) \quad \text{(ind. hyp.)}
\]

\[
= ((x_1x_2\ldots x_k)(x_{k+1}\ldots x_{n-1}))x_n \quad \text{(assoc.)}
\]

\[
= (x_1x_2\ldots x_{n-1})x_n \quad \text{(ind. hyp.)}
\]

which is the product we wanted. Hence the theorem is true for products of \(n\) factors if it is true for products of fewer than \(n\) factors (induction step). Hence by induction the theorem is true for any number of factors. \(\square\)

For an integer \(n > 0\), write \(x^n\) for the product of \(x\) with itself \(n\) times. Write \(x^{-n} = (x^{-1})^n\) and \(x^0 = 1\). Then clearly,

**Theorem 2.1.3** For any group \(G\), any \(x \in G\), and any integers \(m, n\),

\[x^{m+n} = x^m x^n.\]

If you’ve been trying to think of an example of a group, perhaps you’ve considered the real numbers under multiplication. Indeed, the multiplication of real numbers is associative and 1 is an identity element. However, 0 does not have an inverse since \(0y = 0 \neq 1\) for every real number \(y\). But 0 is the only exception and the set \(G = \mathbb{R} - \{0\}\) is such that ordinary multiplication of numbers from \(G\) gives a number in \(G\), multiplication is still associative, 1 \(\in G\) is still an identity, and every element \(x \in G\) has an inverse \(x^{-1} = 1/x\), its multiplicative inverse or reciprocal. Thus the nonzero real numbers form a group under multiplication. Our notation is taken to mimic that of ordinary arithmetic.

On the other hand, ordinary multiplication also satisfies the commutative law \(xy = yx\). We do not assume that a group must satisfy this property. We can, in fact, find examples of groups that do not satisfy this property, so commutativity does not follow from the conditions taken to define a group, and we must be careful not to use commutativity when reasoning about groups in general. The algebras satisfying this extra property form a special subclass of groups.

**Definition 2.1.3** A group \(G\) is Abelian if for all \(x, y \in G\), \(xy = yx\).

In an Abelian group, not only does it not matter what parentheses we take in a product, but the order of the factors does not matter. Hence,
for example, we can show \(((xy)y)(x(yx)) = x^3y^3\) for any \(x, y\) in an Abelian group, whereas this would not hold in the class of all groups (we can find a non-Abelian group where it fails).

Another arithmetic example of a group is the integers under ordinary addition. The sum of integers is an integer (so \(+\) is an operation on the set of integers), addition is associative, 0 is an identity element under addition (since \(0 + x = x + 0 = x\)), and each integer \(x\) has an (additive) inverse \(-x\) (since \(x + (-x) = (-x) + x = 0\)). Writing this group with multiplicative notation would be too confusing (since we already have a multiplication operation defined on the integers, but the integers under multiplication do not form a group, and we already have an element 1 which is not the additive identity element). Instead we may write a group using additive notation with binary operation \(+\), identity 0, and the inverse of \(x\) written \(-x\). In additive notation, the product \(x^n = xx\ldots x\) becomes a sum of \(n\) terms \(x + x + \ldots + x = nx\). Generally, we will reserve additive notation for Abelian groups, i.e., when we use \(+\) for a group operation it will (almost always) be safe to assume commutativity as well as associativity.

As further examples, consider addition and multiplication of \(n \times n\) real matrices. The sum of matrices is defined componentwise, addition is thus associative and commutative, the zero matrix is an identity, and negation can be taken componentwise to define an additive inverse. Thus \(n \times n\) real matrices under addition form an Abelian group. On the other hand, if we restrict to matrices with nonzero determinant, then matrix multiplication is associative (but not commutative for \(n > 1\)), the identity matrix is a multiplicative identity, and inverse matrices exist, giving a (non-Abelian) group of \(n \times n\) real matrices with nonzero determinant under matrix multiplication.

Instead of adding conditions such as commutativity to define subclasses of groups such as Abelian groups, we may also consider theories with fewer restrictions on the algebras. Thus an algebra with a single associative binary operation is called a semigroup. An algebra with an associative binary operation and identity element is called a monoid. Abelian groups are a special class of groups which are special cases of monoids which are, in turn, a subclass of semigroups. The generalized associativity theorem proved above only depends on the associative law and so also holds for semigroups. The rules for positive exponents work in semigroups, while for nonnegative exponents work in monoids.

More examples (and nonexamples) of groups are given in the next section and in the problems below. To summarize: a group is a set together with a
binary operation (a binary function defined on the set, and having values in the set), such that the operation is associative, there is an identity element for the operation, and each element has an inverse.

2.2 More examples

A square has a number of symmetries; rotational symmetry about its center in multiples of 90°, and mirror symmetries about lines through the center parallel to the sides or diagonally. To understand symmetries it proves useful to consider a group structure on symmetries. Imagine a physical square and consider the symmetries of the square as operations to be carried out on the square, picking up the square and putting it back down in the same space with perhaps different corners now in different positions. Operations on the square are the same if they would result in the same finally position. For a group product of symmetries $x$ and $y$ we perform the operation $x$ and then perform the operation $y$ (starting from the ending position of $x$), the product $xy$ is defined by the resulting position of the square, a result that we might recognize more simply as an equivalent single operation. Note that we are considering products of operations on the square as performed as read in left-to-right order (for the moment at least).

To make things easier, take the square with sides parallel to $x$-$y$ coordinate axes through the center of the square. Label the corners of the square in its initial position 1, 2, 3, and 4 according to which quadrant of the coordinate system the corner lies in (first being $(+, +)$, fourth being $(+, -)$ in counter-clockwise order). Then we can enumerate 8 possible symmetries of the square moving corners 1 to 4 into quadrants 1 to 4 according to table 2.2.

The multiplication table for these symmetries is given in table 2.2. The entry in the $x$ row and $y$ column is the product $xy$ of $x$ followed by $y$ acting on the square. We can see from the table that $i$ is the identity for this product by comparing the first row and first column to the row and column labels. Moreover, by locating the $i$ entries in the table, we easily see that $\ell^{-1} = r$, $r^{-1} = \ell$ and each other element is its own inverse. The only condition defining a group that is not so easily checked merely by inspection of the multiplication table is associativity. But we are simply composing functions, even if we are writing composition reading from left to right. Since composition of functions is associative, this product is also associative. If we wanted to write $xy$ as composition of functions (in the usual order) we would write $y \circ x$, the only change being to reflect the multiplication table on its
Table 1: Symmetries of a square as permutations of corners.

<table>
<thead>
<tr>
<th>symbol</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>ℓ</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>s</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>r</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>v</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>h</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>u</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>d</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Main diagonal swapping the row and columns. Note that this multiplication is not commutative (i.e., the group is not Abelian) since the multiplication table is not symmetric, e.g., \( uℓ = v \) but \( ℓu = h \).

Table 2: Multiplication of symmetries of square.

\[
\begin{array}{c|cccccccc}
 x \cdot y & i & ℓ & s & r & v & h & u & d \\
\hline
 i & i & ℓ & s & r & v & h & u & d \\
 ℓ & ℓ & s & r & i & u & d & h & v \\
s & s & r & i & ℓ & h & v & d & u \\
r & r & i & ℓ & s & d & u & v & h \\
v & v & d & h & u & i & s & r & ℓ \\
h & h & u & v & d & s & i & ℓ & r \\
u & u & v & d & h & ℓ & r & i & s \\
d & d & h & u & v & r & ℓ & s & i \\
\end{array}
\]

A permutation of a set \( X \) is a bijective function of \( X \) into \( X \). Fix a set \( X \) and consider the set \( S_X \) of all permutations of \( X \). Since a composition of bijective functions is bijective, the set \( S_X \) is closed under composition. The identity function on \( X \), \( i_X \), is bijective so is in \( S_X \). The inverse of a bijective function in \( S_X \) is defined and bijective so is in \( S_X \). Hence \( S_X \) is a group with multiplication the (ordinary) composition of functions, called the symmetric group on \( X \). If \( X = \{1, 2, \ldots, n\} \) we write \( S_n \) for \( S_X \).

Suppose \( n \) is a positive integer. The integers under addition modulo \( n \) form an additive group (that is, a group where we use additive notation.
with operations $+,-,$ and additive identity element 0). The elements are the congruence classes \{[0],[1],\ldots,[n-1]\} mod $n$ where \([i]\) stands for the set of numbers congruent to $i$ mod $n$, \([i]\) = \([i+kn]\) for any integer $k$, and \([i]+[j]=\[i+j]\) (one needs to check that this is well-defined). The identity element is \([0]\), and additive inverses are given by $-[i]=[-i]=[n-i]$ (again checking this is well-defined no matter what representative $i$ is taken for the equivalence class \([i]\)). The additive group of integers mod $n$ is often denoted $\mathbb{Z}_n$.

The integers mod $n$ under multiplication cannot form a group under multiplication since \([1]\) must be the identity and \([0]\) cannot have an inverse (except when $n=1$). In fact if $i$ and $n$ have a common factor $k$ then for any $j$, \([i][j]=[ij]\) will be a residue class all of whose elements are multiples of $k>1$ and so not congruent to 1 mod $n$. To get a multiplicative group of integers mod $n$ we restrict ourselves to the congruence classes \([i]\) of integers $i$ relatively prime to $n$. Multiplication mod $n$ can then be seen to be a well-defined operation on congruence classes relatively prime to $n$, which is associative, has identity and inverses. This group, denoted $U_n$, is Abelian.

2.3 Subgroups

If we want to understand a group, its properties and what its internal organization or structure is, we might start by analyzing subsets that also happen to be groups under the same operations.

**Definition 2.3.1** A nonempty subset $H$ is a subgroup of a group $G$ if any product (by the multiplication of $G$) of elements of $H$ is in $H$, and the inverse of any element of $H$ is in $H$.

Of course the point in calling a subset $H$ of $G$ a sub-group if it is closed under the multiplication and inverse of $G$ is that $H$ is a group.

**Theorem 2.3.1** If $H$ is a subgroup of a group $G$, then $H$ with the operation of multiplication in $G$ restricted to $H$, the inverse operation of $G$ restricted to $H$, and identity element the identity of $G$, is a group $H$.

A group $H$ is a subgroup of a group $G$ only if $H \subseteq G$ and multiplication and inverses in $H$ are exactly the same whether we work in $G$ or in $H$. Since this is usually the case we are interested in, we may use the same operation symbols for $G$ and $H$ even though technically we should restrict
the operations in $H$ to $H$. A subgroup $H \subseteq G$ is called proper if $H \neq G$. A group $G$ always has $G$ and $\{1\}$ as subgroups, the later is called the trivial subgroup.

**Theorem 2.3.2** Any subgroup of a subgroup $H$ of a group $G$ is a subgroup of $G$.

**Proof:** If $K$ is closed under the operation of $H$, operations which are inherited from the group $G$, then $K$ is closed under the operations of $G$ and so is a subgroup of $G$. $\square$

One particularly simple kind of subgroup of a group $G$ is that generated by a single element $x$. If $H$ is a subgroup containing $x$ then $x^2$, $x^3$, $\ldots$, $x^{-1}$, $x^{-2}$, $\ldots$ must all also belong to $H$. On the other hand, the set $H$ of all powers of $x$ in $G$ is closed under multiplication and inverses since multiplying amounts to adding exponents, and the inverse of $x^n$ is simply $x^{-n}$.

**Definition 2.3.2** A group $G$ is cyclic is there is an element $x \in G$ such that every element of $G$ is a power of $x$. If $x$ is an element of a group $G$, the cyclic subgroup of $G$ generated by $x$ is the subgroup $\langle x \rangle = \{x^n : n \in \mathbb{Z}\}$.

In general, the subgroup generated from a subset $X$ of a group $G$ is defined by taking all possible products of elements and inverses of elements of $X$. More precisely we have the following.

**Theorem 2.3.3** For $X$ a subset of a group $G$, let $X_0 = X \cup \{1\}$ and, for each $i > 0$, let $X_i$ be the union of $X_{i-1}$ and the set of all products and inverses of elements of $X_{i-1}$. Define $\langle X \rangle = \bigcup_{i \geq 0} X_i$. Then $\langle X \rangle$ is the smallest subgroup of $G$ containing the subset $X$, i.e., if $H$ is a subgroup of $G$ containing $X$ then $\langle X \rangle \subseteq H$.

**Proof:** (Outline) First check that $\langle X \rangle$ is a subgroup of $G$. If $x, y \in \langle X \rangle$ then $x \in X_i$ and $y \in X_j$ for some $i$ and $j$ and then $xy \in X_{1+\max(i,j)}$ so $xy \in \langle X \rangle$. Similarly for inverses. Then check that if $X \subseteq H$ then, by induction, each $X_i \subseteq H$ and so $\langle X \rangle \subseteq H$. $\square$

**Definition 2.3.3** The order of an element $x$ of a group $G$, denoted $o(x)$, is the least positive integer $n$ such that $x^n = 1$, if such exists, and otherwise is $\infty$. The order of a group $G$, denoted $o(G)$, is the number of elements in $G$, if $G$ is finite, and otherwise is $\infty$.

Hence, for any $x$ in a group $G$ the order of $x$ and the order of the subgroup $\langle x \rangle$ are the same.
2.4 Problems

Problem 2.4.1 Determine which of the following are groups.

1. Integers under multiplication.
2. Real numbers under addition.
3. Nonzero complex numbers under multiplication.
4. Rational numbers under addition.
5. Positive real numbers under multiplication.
6. Positive real numbers under addition.
7. Nonnegative real numbers under the operation $x \ast y = \sqrt{x^2 + y^2}$.
8. Dyadic rationals (rationals which, when written in lowest terms, have denominator a power of two) under addition.
10. Positive dyadic rationals under addition.
11. Polynomials with real coefficients in a variable $x$ under addition.
12. Polynomials as above under multiplication.
15. Nonzero rational functions (quotients of polynomials) under multiplication.
16. The set of all bijective continuous functions from $\mathbb{R}$ into $\mathbb{R}$ under composition.
17. The set of all bijective function $\mathbb{R} \to \mathbb{R}$ under pointwise addition.
18. The set of all differentiable functions $\mathbb{R} \to \mathbb{R}$ under pointwise addition.
19. The set of $2 \times 2$ integer matrices having determinant $\pm 1$ under matrix multiplication.
20. The set of $2 \times 2$ rational matrices having nonzero determinant under matrix multiplication.

**Problem 2.4.2** Classify each of the above algebras as semigroup, monoid, group, or Abelian group (or none of these).

**Problem 2.4.3** Prove that $G = \{ (a, b) : a, b \in \mathbb{R}, b \neq 0 \}$, with multiplication $(a, b)(c, d) = (a + bc, bd)$ is a group (find the identity element and give a formula for inverses, checking each of the identities defining a group).

**Problem 2.4.4** We need to be careful with certain simplifications in groups.

1. Prove that in any Abelian group $(xy)^{-1} = x^{-1}y^{-1}$.

2. Give an example of a group $G$ together with elements $x, y \in G$ such that $(xy)^{-1} \neq x^{-1}y^{-1}$.

3. Determine, with proof, a formula for $(xy)^{-1}$ in terms of $x^{-1}$ and $y^{-1}$ that works in any group.

4. Prove that, for any group $G$, if for all $x, y \in G$, $(xy)^{-1} = x^{-1}y^{-1}$, then $G$ is an Abelian group.

**Problem 2.4.5** Prove that the intersection of subgroups of a group $G$ is again a subgroup of $G$.

**Problem 2.4.6** Prove that if the union of subgroups $H$ and $K$ of a group $G$ is a subgroup of $G$, then either $H \subseteq K$ or $K \subseteq H$.

**Problem 2.4.7** List all of the cyclic subgroups in the group of symmetries of the square.

**Problem 2.4.8** Complete the enumeration of subgroups of the group of symmetries of the square by finding two noncyclic proper subgroups.

**Problem 2.4.9** Find all subgroups of $\mathbb{Z}_{12}$.

**Problem 2.4.10** Show that any subgroup of a cyclic group is cyclic.
2.5 Lagrange’s theorem

The subgroups of the eight element group of symmetries of the square have either 1 (the trivial subgroup), 2, 4, or 8 elements (the whole group). The subgroups of \( \mathbb{Z}_{12} \) have either 1, 2, 3, 4, 6, or 12 elements. The five element group \( \mathbb{Z}_5 \) has no proper nontrivial subgroups. Why are subgroups of a finite group restricted to only certain sizes? For example, why is the only subgroup containing more than half of the elements of a group the whole group itself? The key restriction on the sizes of subgroups is given by the following theorem.

**Theorem 2.5.1 (Lagrange’s theorem)** If \( G \) is a finite group and \( H \) is a subgroup of \( G \), then the order of \( H \) is a divisor of the order of \( G \).

The converse of this result does not hold, a group \( G \) need not have subgroups of every possible order dividing the order of \( G \). On the other hand, for every integer \( n > 0 \), the group \( \mathbb{Z}_n \) of integers mod \( n \) has subgroups of order \( k \) for every positive integer \( k \) dividing \( n \).

To understand this result, consider a proper subgroup \( H \) of a group \( G \). Multiplying the elements of \( H \) by each other gives only elements of \( H \), in fact in any row of the multiplication table corresponding to an element of \( H \), each element of \( G \) appears once and the subgroup elements appear in the columns corresponding to multiplication by a subgroup element. Multiplying an element of \( H \) by an element of \( G - H \) must therefore give an element of the group not in the subgroup. But the elements of a column of the multiplication table are also all different, so multiplying the elements of \( H \) by an element of \( G - H \) must give exactly as many different elements of \( G - H \) as there are in \( H \). For example, take \( G \) the group of symmetries of the square as given before and let \( H = \{ i, \ell, s, r \} \). Multiplying these elements by \( v \) gives \( \{ v, u, h, d \} \), as many distinct elements not in \( H \). Thus a proper subgroup can have at most half of the elements of the group.

If not all of the elements of \( G \) are exhausted by taking \( H \) and the result of multiplying \( H \) by an element of \( G - H \), then pick another element of \( G \) to multiply each element of \( H \) by. For example, take \( H = \{ i, s \} \). Then multiplying \( H \) by \( \ell \) we get \( \{ \ell, r \} \). Multiplying \( H \) by \( v \) gives \( \{ v, h \} \). These are again as many new and distinct elements as there are in \( H \). Again, multiplying \( H \) by \( u \) gives \( \{ u, d \} \) again new and distinct elements and exhausting the elements of the group. There are four disjoint subsets of two elements each, the order of \( H \), and these exhaust the \( 4 \times 2 = 8 \) elements in the group.
In general, the order of $H$ will be seen to divide the order of the group if we can show that this process always breaks the group into disjoint subsets of the same size as $H$.

To construct a proof of this result we make a few definitions.

**Definition 2.5.1** If $H$ is a subgroup of a group $G$ and $g \in G$, then we write $Hg = \{hg : h \in G\}$. We call such a set a right coset of $H$ in $G$. Similarly, we define left cosets by $gH = \{gh : h \in H\}$.

Anything we do with right cosets could be done as well with left cosets, but it need not be the case that left cosets are the same as right cosets (though for Abelian groups these notions clearly are the same). Note that $g \in Hg$ since the identity element is in $H$. Thus every element of the group is in some coset. In the last example we get right cosets of $H = \{i, s\}$ are $Hi = H = Hs$, $H\ell = \{\ell, r\} = Hr$, $ Hv = \{v, h\} = Hr$ and $Hu = \{u, d\} = Hd$, and it wouldn’t matter which new elements we had picked at each stage to enumerate these right cosets. Note that $Hg_1 = Hg_2$ when $g_1$ and $g_2$ are in the same coset of $H$, but that different cosets cannot overlap. The appropriate idea is captured by the following definition.

**Definition 2.5.2** A partition of a set $X$ is a set $P$ of subsets of $X$ such that:

1. each element of $X$ is an element of some $S \in P$; and
2. for any $S_1, S_2 \in P$ if $S_1 \cap S_2 \neq \emptyset$ then $S_1 = S_2$.

The elements of $P$ are called blocks of the partition.

Alternatively, we may just settle on keeping track of when two elements are in the same block of a partition. Indeed it may be that a partition of a set is most naturally defined in terms of when elements are equivalent in some sense, equivalent elements taken as a block of the partition.

**Definition 2.5.3** An equivalence relation on a set $X$ is a set of pairs $\theta$ of elements of $X$ such that:

1. (reflexive) for all $x \in X$, $(x, x) \in \theta$;
2. (symmetric) for all $x, y \in X$, if $(x, y) \in \theta$, then $(y, x) \in \theta$; and
3. (transitive) for all \(x, y, z \in X\), if \((x, y) \in \theta\) and \((y, z) \in \theta\). then \((x, z) \in \theta\).

Instead of writing \((x, y) \in \theta\), we write \(x \sim y\) or \(x \equiv y\), and say \(x\) is equivalent to \(y\) (if we have only one equivalence relation in mind), or we may write \(x \sim_{\theta} y\), \(x \equiv_{\theta} y\), or simply \(x \theta y\). Write \([x]\) (or \([x]_{\theta}\) or \(x/\theta\)) for the set \(\{y \in X : x \sim y\}\).

The subsets of \(X\) of the form \([x]\) are called equivalence classes or blocks of the equivalence relation.

Thus the reflexive, symmetric, and transitive properties become, for all \(x, y, z \in X\): \(x \sim x\); if \(x \sim y\) then \(y \sim x\); and if both \(x \sim y\) and \(y \sim z\) then \(x \sim z\).

**Theorem 2.5.2** If \(P\) is a partition of \(X\), then defining \(x \sim y\) iff \(x\) and \(y\) are in the same block of \(P\) gives an equivalence relation on \(X\). For any equivalence relation \(\sim\) on \(X\), the set \(P\) of equivalence classes is a partition of \(X\).

**Proof:** If \(P\) is a partition, each \(x\) is in a block of the partition and is certainly in the same block as itself and so \(\sim\) is reflexive. If \(x\) and \(y\) are in the same block then \(y\) and \(x\) are in the same block so \(\sim\) is symmetric. And if \(x\) and \(y\) are in the same block and \(y\) and \(z\) are in the same block then \(x\) and \(z\) are in the same block so \(\sim\) is transitive. That is, “in the same block” is naturally understood as an equivalence relation.

Suppose \(\sim\) is an equivalence relation on \(X\) and take \(P = \{[x] : x \in X\}\). Then each \(x \in [x]\) so every element of \(X\) is in some subset in \(P\). If for some elements \(x\) and \(z\) of \(X\), \([x]\) and \([z]\) are elements of \(P\) that overlap in some \(y\), then \(x \sim y\) and \(z \sim y\). But then \(y \sim z\) by symmetry so \(x \sim z\) by transitivity. Now for any \(w \in [x]\), \(x \sim w\) and since \(z \sim x\) also \(z \sim w\) and \(w \in [z]\). Likewise, for any \(w \in [z]\), \(z \sim w\) so \(x \sim w\) and \(w \in [x]\). That is, if \([x]\) and \([z]\) are elements of \(P\) that overlap then \([x] = [z]\). Hence \(P\) is a partition of \(X\).

□

Our goal now is to show that the right cosets of \(H\) in \(G\) partition \(G\) into blocks of the same size as \(H\). First we show a 1-1 correspondence between \(H\) and any right coset of \(H\), i.e., right cosets are the same size as \(H\).

**Lemma 2.5.3** Suppose \(H\) is a subgroup of a group \(G\) and \(g \in G\). Then the map \(f : H \rightarrow Hg\) defined by \(f(h) = hg\) is a bijection.
Proof: Since $Hg$ consists of exactly the elements of the form $hg$ for $h \in H$, $f$ is surjective. If $h_1g = h_2g$ then $(h_1g)g^{-1} = (h_2g)g^{-1}$ so $h_1 = h_11 = h_1(gg^{-1}) = (h_1g)g^{-1} = (h_2g)g^{-1} = h_21 = h_2$ (from here on out such tedious application of the group axioms will be abbreviated). Thus $f$ is also injective. □

As noted before, each $g \in G$ belongs to a right coset, namely to $Hg$. Thus it remains to see that overlapping right cosets are in fact equal. We may note here that $g_2 \in Hg$ iff $g_2 = hg$ for some $h \in H$, which is iff $g_2g^{-1} = h$, or in other words, iff $g_2g^{-1} \in H$. We might thus attack the problem of showing the right cosets of $H$ partition $G$ by showing instead that defining $g \sim g_2$ iff $g_2g^{-1} \in H$ gives an equivalence relation on $G$. We can also proceed by a direct approach.

Lemma 2.5.4 Suppose $H$ is a subgroup of a group $G$ and $g_1$ and $g_2$ are elements of $G$ such that $Hg_1 \cap Hg_2 \neq \emptyset$, then $Hg_1 = Hg_2$. Consequently, for any $g_1, g_2 \in G$, $g_2 \in Hg_1$ iff $Hg_1 = Hg_2$.

Proof: Suppose $g_3 \in Hg_1 \cap Hg_2$. Then for some $h_1$ and $h_2$ in $H$, $g_3 = h_1g_1 = h_2g_2$. Now suppose $g \in Hg_1$. Then for some $h \in H$, $g = h_1g_1 = hh_1^{-1}h_1g_1 = hh_1^{-1}g_3 = hh_1^{-1}h_2g_2$ and $hh_1^{-1}h_2 \in H$ so $g \in Hg_2$. Similarly, if $g \in Hg_2$ then $g \in Hg_1$. Hence $Hg_1 = Hg_2$. For the rest, if $g_2 \in Hg_1$, then $Hg_1$ and $Hg_2$ overlap and are equal, and conversely, if $Hg_1 = Hg_2$ then $g_2 \in Hg_2 = Hg_1$. □

Hence we have that the right cosets of $H$ in $G$ are a partition of $G$ into subsets of size equal to $H$. If $G$ is finite, the order of $H$ is a divisor of the order of $G$. The same analysis could be carried out with left cosets. We conclude that the number of left cosets of $H$ is the same as the number of right cosets of $H$ and is equal to the order of $G$ divided by the order of $H$.

Definition 2.5.4 The index of a subgroup $H$ in a group $G$, denoted by $[G : H]$, is the number of cosets of $H$ in $G$ (left or right) if this number is finite, and otherwise is taken as $\infty$.

Corollary 2.5.5 If $H$ is a subgroup of a finite group $G$, then $o(G) = o(H)[G : H]$. 30
2.6 Homomorphisms

We next consider functions between groups, considering not only what happens to individual elements under such a map, but also how the group operation in the domain group corresponds to the group operation in the range. We often have different group operations (and inverses and identities) involved in the same discussion. If we use the same multiplication symbol for two different groups $G$ and $H$ this could be confusing. To remove the confusion, we must understand a multiplication involving elements of a group $G$ should mean multiplication in $G$, while for elements of $H$ it should mean multiplication in $H$. This generally causes few problems except possibly when an element could belong to either $G$ or $H$, but this will generally happen only when $H$ is a subgroup of $G$ (or vice-versa) and then the multiplication in the subgroup is exactly that of the larger group anyway. Hence we will not make any special effort to use different symbols for the multiplications in different groups, nor for inverses in different groups, letting the context of an expression determine the operation intended. The identity element of a group may need to be distinguished (for clarity) and instead of writing $1 \in G$ or $1 \in H$ to say which group $1$ is the identity of, we may write $1_G$ or $1_H$.

Consider the function $\phi$ from $\mathbb{Z}_4$, the integers mod 4 under addition, to the group $G$ of symmetries of the square defined by $\phi([0]) = i$, $\phi([1]) = \ell$, $\phi([2]) = s$, and $\phi([3]) = r$. Then the way addition combines elements of $\mathbb{Z}_4$ corresponds to the way multiplication works in $G$. For example, $\phi([1] + [2]) = \phi([3]) = r = \ell s = \phi([1]) \phi([2])$. In fact, since this $\phi$ is injective, we might say $\mathbb{Z}_4$ is essentially the same as the image subgroup $\{i, \ell, s, r\}$. On the other hand, consider the function $\psi$ from $G$ to $\mathbb{Z}_4$ defined so $\psi$ maps $\{i, \ell, s, r\}$ all to $[0]$ and maps $\{v, h, u, d\}$ all to $[2]$. Then, although $\psi$ is far from injective, multiplication of elements from $G$ is still reflected in how their images add in $\mathbb{Z}_4$. For example, $\psi(\ell v) = \psi(u) = [2] = [0] + [2] = \psi(\ell) + \psi(v)$. This implies a certain structure in the multiplication of elements of $G$ that might not have been otherwise apparent.

In the above examples, only the multiplication of the groups was considered, but in fact the inverses are similarly respected and identity elements map to identity elements. Such maps between groups are especially nice because they give us a way of comparing groups, determining some of the structure of a group in terms of another group. We begin with the appropriate definition.

**Definition 2.6.1** A map $\phi : G \to H$ between groups $G$ and $H$ is a group
homomorphism if, for any \(g_1, g_2 \in G\), \(\phi(g_1g_2) = \phi(g_1)\phi(g_2)\) (where the first multiplication is interpreted in \(G\) and the second in \(H\)).

**Theorem 2.6.1** If \(\phi : G \to H\) is a group homomorphism, then \(\phi(1_G) = 1_H\) and, for all \(g \in G\), \(\phi(g^{-1}) = \phi(g)^{-1}\)

**Proof:** Since \(\phi\) is a group homomorphism, for a \(g \in G\), \(\phi(g) = \phi(1_Gg) = \phi(1_G)\phi(g)\), hence \(\phi(1_G) = \phi(g)\phi(g)^{-1} = \phi(1_G)\phi(g)\phi(g)^{-1} = \phi(1_G)\). Then \(\phi(g)\phi(g)^{-1} = \phi(gg^{-1}) = \phi(1_G) = 1_H = \phi(g)\phi(g)^{-1}\) so, multiplying on the left by \(\phi(g)^{-1}\), we have \(\phi(g^{-1}) = \phi(g)^{-1}\phi(g)\phi(g)^{-1} = \phi(g)^{-1}\phi(g)\phi(g)^{-1} = \phi(g)^{-1}\).

Thus to establish a map is a homomorphism we need check only the multiplications in the two groups, but if we have a homomorphism then it respects the multiplication and inverse operations and maps the identity in the domain group to the identity of the range. Next we observe that the image of a group homomorphism is always a subgroup of the range.

**Theorem 2.6.2** If \(\phi : G \to H\) is a group homomorphism, then \(\phi(G) = \{\phi(g) : g \in G\}\) is a subgroup of \(H\).

**Proof:** Suppose \(h_1, h_2 \in \phi(G)\). Then for some \(g_1, g_2 \in G\), \(\phi(g_1) = h_1\) and \(\phi(g_2) = h_2\). Hence \(h_1h_2 = \phi(g_1)\phi(g_2) = \phi(g_1g_2) \in \phi(G)\) and \(h_1^{-1} = \phi(g_1)^{-1} = \phi(g_1^{-1}) \in \phi(G)\). Since \(1_H = \phi(1_G) \in \phi(G)\) as well, \(\phi(G)\) is a subgroup of \(H\).

When are groups “essentially” the same? When there is a homomorphism between them which is a bijection.

**Definition 2.6.2** Groups \(G\) and \(H\) are isomorphic, and we write \(G \cong H\), if there is a homomorphism \(\phi : G \to H\) which is a bijection (the map \(\phi\) is then said to be an isomorphism).

**Theorem 2.6.3** Isomorphism is an equivalence relation on groups, i.e., for any groups \(G, H\) and \(K\): \(G \cong G\); if \(G \cong H\) then \(H \cong G\); and if both \(G \cong H\) and \(H \cong K\) then \(G \cong K\).

**Proof:** The identity map on \(G\) is a homomorphism from \(G\) to \(G\) which is a bijection, so isomorphism is reflexive. If \(\phi : G \to H\) is a homomorphism and a bijection, consider the bijection \(\phi^{-1} : H \to G\) (the inverse as a function, defined since \(\phi\) is a bijection). We check for \(h_1, h_2 \in H\), since
\( \phi \circ \phi^{-1} \) is the identity on \( H \), \( \phi^{-1}(h_1h_2) = \phi^{-1}(\phi(\phi^{-1}(h_1))\phi(\phi^{-1}(h_2))) = \phi^{-1}(\phi(\phi^{-1}(h_1))\phi^{-1}(h_2))) \), since \( \phi \) is a homomorphism, since \( \phi^{-1} \circ \phi \) is the
identity on \( G \) this in turn is \( \phi^{-1}(h_1)\phi^{-1}(h_2) \). Hence isomorphism is symmetric.
Finally suppose \( \phi : G \to H \) and \( \psi : H \to K \) are homomorphisms and bijections, the composition \( \psi \circ \phi : G \to K \) is a bijection so it remains only
to check that the composition of homomorphisms is a homomorphism. \( \square \)

**Theorem 2.6.4** If \( \phi : G \to H \) and \( \psi : H \to K \) are group homomorphisms,
then \( \psi \circ \phi : G \to K \) is a group homomorphism.

**Proof:** Suppose \( g_1, g_2 \in G \). Then we have \( \psi(\phi(g_1g_2)) = \psi(\phi(g_1)\phi(g_2)) = \psi(\phi(g_1))\psi(\phi(g_2)) \). \( \square \)

Isomorphic groups are, in some sense, the same group, though perhaps represented by differently as operations on sets. A homomorphism which is injective is an isomorphism of the domain with the image of the homomorphism. A homomorphism which is not injective maps some elements in the domain group to the same element in the range and thus some of the information about how multiplication works in the domain is not carried over to the image.

Suppose \( \phi : G \to H \) is a group homomorphism. Taking \( g_1 \sim g_2 \) if
\( \phi(g_1) = \phi(g_2) \) for elements \( g_1, g_2 \in G \), we can easily see \( \sim \) is an equivalence
relation on \( G \). We next consider the corresponding partition of \( G \).

**Definition 2.6.3** The kernel of a group homomorphism \( \phi : G \to H \), denoted
er\( ker(\phi) \), is the set \( \{ g \in G : \phi(g) = 1_H \} \).

**Theorem 2.6.5** Suppose \( \phi : G \to H \) is a group homomorphism and \( K = ker(\phi) \). Then \( K \) is a subgroup of \( G \).

**Proof:** Clearly \( 1_G \in K \) since \( \phi(1_G) = 1_H \). If \( g_1, g_2 \in K \), then \( \phi(g_1) = \phi(g_2) = 1_H \) so \( \phi(g_1g_2) = \phi(g_1)\phi(g_2) = 1_H1_H = 1_H \) and so \( g_1g_2 \in K \). If \( g \in K \)
then \( \phi(g) = 1_H \) so \( \phi(g^{-1}) = \phi(g)^{-1} = 1_H^{-1} = 1_H \). Hence \( K \) is a subgroup of
\( G \). \( \square \)

**Theorem 2.6.6** Suppose \( \phi : G \to H \) is a group homomorphism, \( K = ker(\phi) \), \( g \in G \) and \( h = \phi(g) \). Then \( \phi^{-1}(h) = \{ g_2 \in G : \phi(g_2) = h \} \) is
the right coset \( Kg \) of \( K \).
Proof: If $g_2 \in \phi^{-1}(h)$, then $\phi(g_2) = h = \phi(g)$. Hence we have $\phi(g_2g^{-1}) = \phi(g_2)\phi(g)^{-1} = 1_H$ so $g_2g^{-1} \in K$ and so $g_2 \in Kg$. If $g_2 \in Kg$, then $g_2 = kg$ for some $k \in K$ and $\phi(g_2) = \phi(kg) = \phi(k)\phi(g) = 1_H \phi(g) = \phi(g)$. Hence $\phi^{-1}(h) = Kg$. □

An entirely similar proof would show that the elements mapping to $\phi(g)$ are those in the left coset $gK$. Thus every right coset of the kernel of a homomorphism is also a left coset and vice-versa. Since not every subgroup in a group need have this property, kernels of homomorphisms are somewhat special.

Definition 2.6.4 A subgroup $H$ of a group $G$ is normal, and we write $H \triangleleft G$, if, for every $g \in G$, $gHg^{-1} = \{ghg^{-1} : h \in H\}$ is a subset of $H$.

Theorem 2.6.7 If $H$ is a normal subgroup of $G$, then for every $g \in G$, $gHg^{-1} = H$. A subgroup is normal iff all of its right cosets are also left cosets.

Proof: If $gHg^{-1} \subseteq H$, then $H \subseteq g^{-1}Hg$. If this is true for all $g \in G$, then for $g^{-1}$ this says $H \subseteq gHg^{-1}$. But then $gHg^{-1} \subseteq H \subseteq gHg^{-1}$ and $gHg^{-1} = H$. For a normal subgroup $H$, $gHg^{-1} = H$ implies $gH = Hg$ so the left and right cosets of $H$ containing $g$ are equal for any $g \in G$. If every right cosets of $H$ is also a left coset then for any $g \in G$ the right coset containing $g$ must be the left coset containing $g$ and so $gH = Hg$ and $gHg^{-1} = H$ making $H$ a normal subgroup. □

Thus to check a subgroup is normal it suffices to show $gHg^{-1} \subseteq H$, but then it follows also that the right cosets are the same as the left cosets. This does not necessarily mean that every element of a normal subgroup commutes with every element of the group. In an Abelian group, clearly every subgroup is normal. That kernels of homomorphisms are normal subgroups should be clear but we also offer a direct proof.

Theorem 2.6.8 The kernel of a group homomorphism $\phi : G \to H$ is a normal subgroup of $G$.

Proof: Let $K$ be the kernel of $\phi$ and take $g \in G$. For $k \in K$, $\phi(gkg^{-1}) = \phi(g)\phi(k)\phi(g)^{-1} = \phi(g)1_H\phi(g)^{-1} = 1_H$ so $gKg^{-1} \subseteq K$ and $K$ is a normal subgroup of $G$. □
2.7 Quotient groups and the homomorphism theorem

Given a group $G$ and a normal subgroup $K$ of $G$, is $K$ the kernel of a homomorphism? If so, the multiplication in the image of the homomorphism can be determined from the multiplication in $G$. Put differently, the kernel of a homomorphism determines a partition of the domain with particularly nice properties. If $K$ is the kernel of $\phi$, $Kg_1$ and $Kg_2$ are cosets of $K$, and we take $g_3 \in Kg_1$ and $g_4 \in Kg_2$, then $g_3g_4 \in K(g_1g_2)$ since $\phi(g_3g_4) = \phi(g_3)\phi(g_4) = \phi(g_1)\phi(g_2) = \phi(g_1g_2)$. That is, multiplication of the elements from two given cosets of the kernel will result in elements all in another coset of the kernel.

But this property holds for any normal subgroup $K$, and a similar result holds for inverses.

**Theorem 2.7.1** Suppose $K$ is a normal subgroup of $G$, and $g_1, g_2 \in G$. Then for all $g_3 \in Kg_1$ and $g_4 \in Kg_2$, $g_3g_4 \in K(g_1g_2)$. For all $g_3 \in Kg_1$, $g_3^{-1} \in K(g_1)^{-1}$.

**Proof:** Let $g_3 = k_1g_1$ and $g_4 = k_2g_2$ for $k_1, k_2 \in K$. Then $g_3g_4 = k_1g_1k_2g_2 = (k_1g_1k_2g_1^{-1})(g_1g_2)$. Now since $K$ is normal, $g_1k_2g_1^{-1} \in g_1Kg_1^{-1} = K$. Hence $k_1g_1k_2g_1^{-1} \in K$ and $g_3g_4 \in K(g_1g_2)$. Likewise we have $g_3^{-1} = g_1^{-1}k_1^{-1} = (g_1^{-1}k_1^{-1}g_1)(g_1^{-1})$. But $g_1^{-1}k_1^{-1}g_1 \in g_1^{-1}Kg_1 = K$ so $g_3^{-1} \in (Kg_1)^{-1}$. □

Now this observation allows us to define a group whose elements are the cosets of a normal subgroup.

**Definition 2.7.1** Given a normal subgroup $K$ of $G$ define the quotient group of $G$ by $K$, denoted $G/K$, to be the set of all right cosets of $K$ in $G$ with multiplication defined by $(Kg_1)(Kg_2) = K(g_1g_2)$, inverse defined by $(Kg_1)^{-1} = K(g_1^{-1})$ and $1_{G/K} = K1_G = K$. Multiplication and inverses are well defined by the last theorem since if $Kg_3 = Kg_1$ and $Kg_4 = Kg_2$ then $K(g_3g_4) = K(g_1g_2)$, and $K(g_3^{-1}) = K(g_1^{-1})$.

**Theorem 2.7.2** The quotient group $G/K$ is a group. The map $\psi : G \rightarrow G/K$ defined by $\psi(g) = Kg$ for $g \in G$ is a surjective group homomorphism with kernel $K$. The homomorphism $\psi$ is called the quotient map of $G$ onto $G/K$.

**Proof:** Suppose $Kg_1, Kg_2, Kg_3$ are any cosets of $K$ in $G$. Then we check

$$(Kg_1)((Kg_2)(Kg_3)) = (Kg_1)(K(g_2g_3)) = K(g_1(g_2g_3)) = K((g_1g_2)g_3) = (K(g_1)(Kg_2))(Kg_3) = ((Kg_1)(Kg_2))(Kg_3)$$

so the multiplication is associative.
Next \((Kg_1)(K1_G) = Kg_1 = K(1_Gg_1) = (K1_G)(Kg_1)\) so \(1_{G/K}\) is an identity element. Finally, we note that \((Kg_1)(Kg_1)^{-1} = K(g_1g_1^{-1}) = K1_G = K(g_1^{-1}g_1) = (Kg_1)^{-1}(Kg_1)\) and we have inverses. Hence \(G/K\) is a group. For \(g_1, g_2 \in G\), \(\psi(g_1g_2) = K(g_1g_2) = (Kg_1)(Kg_2) = \psi(g_1)\psi(g_2)\) so \(\psi\) is a homomorphism. The kernel of \(\psi\) is the set of \(g \in G\) with \(K = 1_{G/K} = \psi(g) = Kg\), i.e., all \(g \in K\). □

Hence every normal subgroup is the kernel of some homomorphism. Now we can connect the image of a homomorphism with the quotient group derived from the kernel of the homomorphism in the domain group.

**Theorem 2.7.3** (First homomorphism theorem) Let \(\phi : G \rightarrow H\) be a group homomorphism, \(K\) the kernel of \(\phi\), and \(\psi : G \rightarrow G/K\) be the quotient map of \(G\) onto \(G/K\). Then the subgroup \(\phi(G)\) of \(H\) is isomorphic to \(G/K\). Moreover, there is an isomorphism \(\alpha : G/K \rightarrow \phi(G)\) such that \(\phi = \alpha \circ \psi\).

**Proof:** Define a map \(\alpha : G/K \rightarrow \phi(G)\) by taking, for each coset \(Kg\), \(\alpha(Kg) = \phi(g)\). We must check that this is well-defined, i.e., if we had taken a different representative of the same coset, say \(Kg_2 = Kg\), then \(\alpha(Kg_2) = \alpha(Kg)\). But if \(Kg = Kg_2\) then \(g_2 = kg\) for some \(k\) in the kernel of \(\phi\) so \(\phi(g_2) = \phi(k)\phi(g) = 1_H\phi(g) = \phi(g)\), i.e. \(\alpha(Kg_2) = \alpha(Kg)\) as required. Clearly, \((\alpha \circ \psi)(g) = \alpha(Kg) = \phi(g)\) so \(\alpha\) has been defined just so as to make \(\phi = \alpha \circ \psi\).

Now \(\alpha\) is a surjection since for \(h \in \phi(G)\) there is a \(g \in G\) with \(\phi(g) = h\) and hence \(\alpha(Kg) = \phi(g) = h\). To check \(\alpha\) is injective, suppose \(\alpha(Kg_1) = \alpha(Kg_2)\). Then \(\phi(g_1) = \phi(g_2)\) and \(\phi(g_2g_1^{-1}) = \phi(g_2)\phi(g_1)^{-1} = 1_H\phi(g_1)^{-1} = 1_H\) so \(g_2g_1^{-1} \in K\) and \(Kg_2 = Kg_1\) (even though it may be \(g_1 \neq g_2\)). Finally to see \(\alpha\) is a homomorphism we check \(\alpha((Kg_1)(Kg_2)) = \alpha(K(g_1g_2)) = \phi(g_1g_2) = \phi(g_1)\phi(g_2) = \alpha(Kg_1)\alpha(Kg_2)\). Hence \(\alpha\) is an isomorphism. □

### 2.8 Problems

**Problem 2.8.1** List the right cosets of the subgroup \(H\) of \(\mathbb{Z}_{12}\) generated by \([3]\). What is the order of \(H\)? What is the index of \(H\) in \(G\)?

**Problem 2.8.2** Determine the left and right cosets of the subgroup \(H = \{i, h\}\) in the group of symmetries of the square.

**Problem 2.8.3** Determine the normal subgroups of the group of symmetries of the square.
Problem 2.8.4 Prove that a subgroup of index 2 in a group $G$ is a normal subgroup of $G$.

Problem 2.8.5 Prove that if $H$ and $K$ are normal subgroups of a group $G$, then $H \cap K$ is a normal subgroup of $G$.

Problem 2.8.6 Prove that the homomorphic image of a cyclic group is cyclic.

Problem 2.8.7 Suppose $G$ is a cyclic subgroup of order $n$ generated by $x \in G$ and suppose $H$ is a group and $y \in H$. Show that there is a homomorphism $\phi : G \to H$ with $\phi(x) = y$ iff the order of $y$ is a divisor of $n$. Show that if a homomorphism $\phi$ with $\phi(x) = y$ exists, then it is unique.

Problem 2.8.8 Prove that if $G$ is a group of order $p$, where $p$ is prime, then $G$ is a cyclic group and therefore $G$ is isomorphic to $\mathbb{Z}_p$.

Problem 2.8.9 Let $H$ be the subgroup of the group of integers $\mathbb{Z}$ under addition generated by an integer $n > 0$. Since $\mathbb{Z}$ is Abelian, $H$ is a normal subgroup of $\mathbb{Z}$. What is the quotient group $\mathbb{Z}/H$?

Problem 2.8.10 Prove that if $\phi : G \to H$ is a surjective group homomorphism, then the subgroups of $H$ are in 1-1 correspondence with the subgroups of $G$ containing $\ker(\phi)$, a subgroup $G_2$ of $G$ containing $\ker(\phi)$ corresponding to the subgroup $H_2 = \phi(G_2)$ in $H$.

Problem 2.8.11 Prove that if $\phi : G \to H$ and $\psi : H \to K$ are group homomorphisms, then $\ker(\phi)$ is a subgroup of $\ker(\psi \circ \phi)$.

Problem 2.8.12 Prove that in any group $G$, the trivial group and the whole group $G$ are always normal subgroups of $G$. What are the quotients of $G$ by these normal subgroups?
2.9 Refinements of the homomorphism theorem

We further refine our understanding of homomorphisms by considering how subgroups and normal subgroups of the domain and image groups correspond. We may as well assume now that we start with a surjective group homomorphism.

**Theorem 2.9.1** (Second homomorphism theorem) Suppose \( \phi : G \to H \) is a surjective group homomorphism with kernel \( K \) (so \( G/K \) is isomorphic to \( H \)). Then there is a 1-1 correspondence between the subgroups of \( G \) containing \( K \) and the subgroups of \( H \), a subgroup \( M \) of \( G \) containing \( K \) corresponding to \( M' = \phi(M) \) and a subgroup \( M' \) of \( H \) corresponding to \( M = \phi^{-1}(M') = \{ g \in G : \phi(g) \in M' \} \). Moreover for corresponding subgroups \( M \subseteq G \) and \( M' \subseteq H \), \( M/K \cong M' \) (by the restriction of the isomorphism of \( G/K \) and \( H \) to the subgroup \( M/K \)). The normal subgroups of \( G \) containing \( K \) correspond to the normal subgroups of \( H \) by this correspondence.

**Proof:** First, we note that if \( M \) is a subgroup of \( G \) then \( \phi(M) \) is a subgroup of \( H \) since \( \phi \) restricted to \( M \) is a homomorphism. Likewise, if \( M' \) is a subgroup of \( H \), then \( \phi^{-1}(M') \) is a subgroup of \( G \) containing \( K \), since if \( x, y \in \phi^{-1}(M') \) then \( \phi(x), \phi(y) \in M' \) so \( xy \) and \( x^{-1} \) must be in \( \phi^{-1}(M') \) since \( \phi(xy) = \phi(x)\phi(y) \in M' \) and \( \phi(x^{-1}) = (\phi(x))^{-1} \in M' \), but \( K = \phi^{-1}(0) \subseteq \phi^{-1}(M') \) since \( 0 \in M' \).

To show the 1-1 correspondence between subgroups \( M \subseteq G \) containing \( K \) and subgroups \( M' \subseteq H \) we show that the maps sending an \( M \) to \( M' = \phi(M) \) and an \( M' \) to \( M = \phi^{-1}(M') \) are inverses, i.e., for a subgroup \( M \) with \( K \subseteq M \subseteq G \), \( \phi^{-1}(\phi(M)) = M \) and for a subgroup \( M' \subseteq H \), \( \phi(\phi^{-1}(M')) = M' \).

In the first part, if \( x \in M \) then \( \phi(x) \in \phi(M) \) means that \( x \in \phi^{-1}(\phi(M)) \) so we have \( M \subseteq \phi^{-1}(\phi(M)) \). On the other hand, if \( x \in \phi^{-1}(\phi(M)) \) then \( \phi(x) = \phi(m) \) for some element \( m \in M \) so \( xm^{-1} \in \ker(\phi) = K \). But \( K \subseteq M \) so \( x \in mK \subseteq M \) and so we have \( \phi^{-1}(\phi(M)) \subseteq M \) and is in fact equal. For the second part, if \( y \in M' \) then since \( \phi \) is surjective, \( y = \phi(x) \) for some \( x \in G \), so \( x \in \phi^{-1}(M') \) and \( y = \phi(x) \in \phi(\phi^{-1}(M')) \). Conversely, for \( y \in \phi(\phi^{-1}(M')) \), \( y = \phi(x) \) for some \( x \in \phi^{-1}(M') \) so \( y = \phi(x) \in M' \). This \( M' = \phi(\phi^{-1}(M')) \) and the 1-1 correspondence for subgroups is established.

Now if \( M \subseteq G \) is a subgroup containing \( K \) then \( \phi \) is also a normal subgroup of \( M \) and the homomorphism \( \phi \) restricts to a surjective homomorphism of \( M \) onto \( M' = \phi(M) \) with kernel \( K \). Hence by the first homomorphism theorem, \( M' \cong M/K \).
Finally, for a normal subgroup \( M \) of \( G \) containing \( K \), \( M' = \phi(M) \) is a normal subgroup of \( H \) since for \( h \in H \), taking \( g \in G \) with \( \phi(g) = h \), we have \( hM'h^{-1} = \phi(gMg^{-1}) = \phi(M) = M' \). Conversely, if \( M' \) is a normal subgroup of \( H \), then \( M = \phi^{-1}(M') \) is a normal subgroup of \( G \) since for \( g \in M \), \( \phi(gMg^{-1}) = \phi(g)M'\phi(g)^{-1} = M' \) so \( gMg^{-1} \subseteq \phi^{-1}(M') = M \). \( \square \)

There is also a nice correspondence between quotients of the domain and image groups of a homomorphism.

**Theorem 2.9.2** (Third homomorphism theorem) Suppose \( \phi : G \to H \) is a surjective group homomorphism with kernel \( K \). Suppose that \( N \) is a normal subgroup of \( G \) containing \( K \) and \( N' = \phi(N) \) is the corresponding normal subgroup of \( H \). Then \( G/N \) is isomorphic to \( H/N' \). In terms of quotient groups, we have \( G/N \cong (G/K)/(N/K) \).

**Proof:** We start by noting that \( G/K \cong H \) by an isomorphism \( \alpha : G/K \to H \) and \( N/K \cong N' \) by the restriction of \( \alpha \) to \( N/K \) by the first and second homomorphism theorems. Since for \( g \in G/K \) and \( x \in N/K \), \( \alpha(gxg^{-1}) = \alpha(g)\alpha(x)\alpha(g)^{-1} = \alpha(g)N'\alpha(g)^{-1} = N' \), we get \( gxg^{-1} \in N/K \) and so \( N/K \) is a normal subgroup of \( G/K \). The cosets of \( N/K \) in \( G/K \) correspond to the cosets of \( N' \) in \( H \) by \( \alpha((N/K)g) = N'\alpha(g) \) and so \( (G/K)/(N/K) \cong H/N' \). Hence the first assertion reduces to the second, that \( G/N \cong (G/K)/(N/K) \).

To see the second assertion, define a homomorphism \( f : G/K \to G/N \) by \( f(Kg) = Ng \), checking that this is well-defined since if \( Kg' = Kg \) then \( g'g^{-1} \in K \subseteq N \) so \( Ng' = Ng \), and that this is a homomorphism since \( \phi((Kg)(Kg')) = \phi(Kgg') = Ngg' = NgNg' = \phi(Kg)\phi(Kg') \). Note that \( f \) is clearly surjective. Now a coset \( Kg \) is in the kernel of this homomorphism iff \( Ng = N \), meaning \( g \in N \) so \( Kg \subseteq N \), i.e., iff \( Kg \in (N/K) \). Thus the kernel of \( f \) is \( N/K \) and hence by the first homomorphism theorem applied to \( f \), \( (G/N) \cong (G/K)/(N/K) \). \( \square \)

### 2.10 Symmetric groups and Cayley’s theorem

Recall that the symmetric group \( S_X \) (sometimes denoted \( \text{Sym}(X) \)) consists of all bijective functions from \( X \) to \( X \) (permutations of \( X \)) under composition. We consider these examples because they are easily understood and computed, at least for finite sets \( X \), and most applications of groups are to symmetries of some sort of structure on a set \( X \) which is thus explicitly a subgroup of the group of all permutations of \( X \). In fact, we will see that every group is a subgroup of a symmetric group.
If \( X = \{x_1, x_2, \ldots, x_n\} \) is a finite set, then we can specify a permutation \( f \in S_X \) by listing the \( x \in X \) and corresponding \( f(x) \) perhaps in a table such as

\[
\begin{pmatrix}
  x_1 & x_2 & \cdots & x_n \\
  f(x_1) & f(x_2) & \cdots & f(x_n)
\end{pmatrix}
\]

This notation takes quite a bit of room. If there is a natural order to the elements of \( X \) it might make sense to list only the second row of this table. A different alternative is to take advantage of the fact that each \( x_i \) also appears in the second row of the table just once as \( f(x_j) \) since \( f \) is a permutation. This alternative also reveals something of the nature of \( f \).

**Definition 2.10.1** Suppose \( f \in S_X \) for \( X \) finite. A cycle of \( f \) is a sequence \( a_1, a_2, \ldots, a_k \) of distinct elements of \( X \) with \( a_{i+1} = f(a_i) \) for each \( i < k \) and \( a_1 = f(a_k) \). If \( a_1, a_2, \ldots, a_k \) are distinct elements of \( X \), we define \( g = (a_1 a_2 \ldots a_k) \) to be the permutation of \( X \) such that \( g(x) = a_{i+1} \) if \( x = a_i \) for \( i < k \), \( g(a_k) = a_1 \) and \( g(x) = x \) if \( x \) is not among the \( a_i \). Then \( g \) has \( a_1, a_2, \ldots, a_k \) as a cycle, in fact as its only cycle of length greater than one when \( k > 1 \), and we say \( g \) is a cycle. A cycle of \( f \) of just one element is a fixed point of \( f \), i.e., an \( x \) with \( f(x) = x \). The permutation denoted by \( (x_i) \) is taken simply to be the identity map on \( X \). The cycle \( (a_1 a_2 \ldots a_k) \) is equivalent to \( (a_2 \ldots a_k a_1) \) or any other cyclic permutation of the \( a_i \) sequence. A cycle of two elements is called a transposition.

The inverse of the cycle \( (a_1, a_2, \ldots, a_k) \) is the cycle \( (a_k, \ldots, a_2, a_1) \). For example, on \( X = \{1, 2, \ldots, 6\} \), the product \( f = (1 2 3)(4 5) \) is the function with \( f(1) = 2 \), \( f(2) = 3 \), \( f(3) = 1 \), \( f(4) = 5 \), \( f(5) = 4 \) and \( f(6) = 6 \), having as cycles the three cycle \( 1, 2, 3 \), a transposition \( 4, 5 \) and a cycle consisting of only the fixed point 6. It is not difficult to take a permutation and write down a product of cycles equivalent to the permutation.

**Theorem 2.10.1** Every permutation \( f \in S_X \) for a finite \( X \) can be written as the product of its different cycles.

**Proof:** Informally, we describe an algorithm for listing the cycles of \( f \). Let \( a_1 = x_1, a_2 = f(a_1), a_3 = f(a_2), \) etc., until hitting an \( a_k \) with \( f(a_k) = a_1 \). Write down the cycle \( (a_1 a_2 \ldots a_k) \). If all of the elements of \( X \) are listed in this cycle then this is \( f \). Otherwise, pick an \( x_j \in X \) which is not one of the \( a_i \). Repeat the process letting \( b_1 = x_j, b_2 = f(b_1), \) etc., until finding
a second complete cycle of $f$. Writing this cycle next to the previous cycle we can continue to find a third cycle, fourth cycle, etc., until exhausting all of the elements of $X$. Now we claim that the original $f$ is the product of these cycles. For $x \in X$, $x$ is an element of just one of these factor cycles, say $(c_1 c_2 \ldots c_m)$ with $x = c_i$ and $f(x) = c_{i+1}$ (or $c_1$ when $i = m$). Neither $x$ nor $f(x)$ belongs to any other cycle. Computing the product applied to $x$, when we compute a different cycle $(d_1 d_2 \ldots)$ applied to $x$ or $f(x)$ we get back $x$ or $f(x)$. Thus the product of cycles applied to $x$ evaluates from left to right giving $x$ at each stage until the cycle $(c_1 c_2 \ldots c_m)$ is reached and $f(x)$ at each stage thereafter. But then the product maps $x$ to $f(x)$, and since this applies to any $x \in X$, the product is equal to $f$. A more technical (but no more revealing) proof would involve perhaps an inductive proof on the number of elements in $X$. □

We ordinarily omit cycles of length 1 since these are the identity. The cycle structure of a permutation is immediately revealed by writing it as a product of disjoint cycles. This also allows us to compute products of cycles reasonably easily. For example, if $f = (123)(345)(136)$, we compute $f(1) = ((123)(345))(3) = (123)(4) = 4$, then $f(4) = 5$, and $f(5) = 1$ for a cycle, $f(2) = 3$, $f(3) = 6$, and $f(6) = 2$ for a second cycle so $f = (145)(263)$ is the product of disjoint cycles. The product $gh$ of two disjoint cycles is equal to the reverse product $hg$ since what $g$ fixes what $h$ moves and vice-versa. On the other hand, $(13)(12) = (123)$ but $(12)(13) = (132)$ is its inverse. In fact, every permutation of a finite set is a product of transpositions, this since we can compute that $(a_1 a_2 \ldots a_n) = (a_n a_1) \ldots (a_2 a_1)$ and more is true.

**Theorem 2.10.2** Any $f \in S_X$ for a finite $X$ is a product of transpositions. If $f$ is some product of an even number of transpositions, then every product of transpositions giving $f$ involves an even number of transpositions. The set of permutations representable as a product of an even number of transpositions is a subgroup $A_X$ (called the alternating group on $X$), of index 2 in $S_X$ if $X$ has at least 2 elements. (Such permutations are called even permutations, all products of an odd number of transpositions being considered odd permutations.)

**Proof:** Every $f$ is a product of cycles by the previous theorem so the observation that cycles are products of transpositions will complete the proof of the first assertion. If $f$ and $g$ are each representable as a product of an even number of transpositions then $fg$ and $f^{-1}$ are also, by multiplying even
products for \( f \) and \( g \) together for \( fg \) and by reversing the order of the even product for \( f \) to get \( f^{-1} \) (the any transposition is its own inverse). Hence the even permutations form a subgroup \( A_X \). Suppose there is any permutation \( f \) of \( X \) which is not representable as a product of an even number of transpositions. Then every element of \( A_X f \) is a product of an odd number of transpositions but not of an even number of transpositions since \( A_X f \) is disjoint from \( A_X \). Then if \( g \) is a product of an odd number of transpositions in some way, then \( gf \) is a product of an even number of transpositions so belongs to \( A_X \). But then \( g \) cannot also belong to \( A_X \) else \( gf \in A_X f \) which is disjoint from \( A_X \). That is, every \( g \) which is representable as a product of an odd number of transpositions can only be represented as a product of transpositions with an odd number of factors. The same then applies to permutations representable as a product of an even number of transpositions, they can only be represented as products of an even number of transpositions. Every element of \( S_X \) is a product of transposition and so belongs to \( A_X \) or \( A_X f \), i.e., \( A_X \) has index 2 in \( S_X \). In particular, transpositions cannot also be products of an even number of transpositions.

It remains to be seen why not every element of \( S_X \) can be represented as a product of an even number of transpositions, supposing now that \( X \) has at least two elements \( a \) and \( b \), proceeding by induction on the size of \( X \). The transposition \( (ab) \) is a product of an odd number (1) of transpositions. If \( X \) consists of only \( a \) and \( b \), then the only transposition is \( (ab) \) and a product of an even number of transpositions is an even power of this transposition is the identity and not this transposition (establishing the basis case). Suppose instead then that \( X \) has more than two elements, \( (ab) \) is also a product of an even number of transpositions say \( (ab) = (a_1 b_1) \ldots (a_{2k} b_{2k}) \). If some element \( x \in X \) is not among the \( a_i \) or \( b_i \), then this product would also be an even product in \( S_{X-\{x\}} \) representing the \( (ab) \) transposition in \( X-\{x\} \) a set of at least 2 elements where by induction hypothesis not every permutation is even and so where, by the argument of the last paragraph, transpositions are not products of an even number of transpositions, a contradiction. Hence we must have every \( x \in X \) used in any even product giving \( (ab) \).

Fix \( a \in X - \{a,b\} \) and take an even product \( (ab) = (a_1 b_1) \ldots (a_{2k} b_{2k}) \) involving \( c \) a minimal number of times. This \( c \) cannot appear just once among the \( a_i \) and \( b_i \) otherwise at some point in this product \( c \) maps to some other element \( d \) of \( X \) according to the one transposition \( (cd) \) in the product and this \( d \) then possibly maps to another \( e \) by the remainder of the product but \( e \neq c \) since \( c \) is not involve in the remainder and so \( c \) would not be fixed by
the product. Instead $c$ must appear in at least two different transpositions in the product.

Take a product $(ab) = (a_1 b_1) \ldots (a_{2k} b_{2k})$ where $c$ occurs a minimum number of times and two occurrences of $c$ appear in transpositions as close together as possible in those products where $c$ appears a minimum number of times, say $(cd)(xy) \ldots (ce)$ is a subsequence in this product with $(cd)$ as close to $(ce)$ as in any product of transpositions evaluating to $(ab)$, and with $(x,y)$ the first transposition after $(c,d)$ if there are any intervening transpositions between these occurrences of $c$. If $x$ and $y$ are both different from $c$ and $d$ then we have $(cd)(xy) = (xy)(cd)$ and we could have the $(cd)$ closer to $(ce)$ in a product of the same (even) length representing $(ab)$. If $x$ or $y$ is $d$, say $x = d$, but $y \neq c$, then $(cd)(dy) = (dy)(cy)$ and again we could have the occurrences of $c$ closer together in an even product for $(ab)$. Otherwise, $x$ or $y$ is $c$, say $x = c$ and we have successive transpositions involving $c$. Then if $y \neq d$, $(cd)(cy) = (dy)(cd)$ and we could have taken a product involving fewer occurrences of $c$. Finally, if $x = c$ and $y = d$, then $(cd)(cd) = 1$ and we could take two fewer occurrences of $c$ in the product. But in every case we have contradicted the minimality assumption on an even product representing $(ab)$. Hence any representation of $(ab)$ as a product of transpositions must be a product of an odd number of transpositions, completing the proof of the theorem. \[\square\]

The even permutations in $S_X$ for $X$ finite form an important subgroup $A_X$, called the alternating group. A simple group is a group with no proper nontrivial normal subgroups, i.e., such that every homomorphic image is trivial or isomorphic to the group. The finite simple Abelian groups are the cyclic groups $\mathbb{Z}_p$ for $p$ a prime. If $X$ has $n$ elements for $n \geq 5$ then $A_X$ (or simply $A_n$) is a non-Abelian simple group of $n!/2$ elements. The complete classification of finite simple groups was a major accomplishment of finite group finished only a few decades ago. One reason for the importance of symmetric groups is the following result.

**Theorem 2.10.3 (Cayley’s theorem)** Every group $G$ is isomorphic to a subgroup of some symmetric group $S_X$ (for some $X$ depending on $G$). If $G$ is finite, then $X$ can be taken to be finite.

**Proof:** Take $X = G$ and define a map $\phi : G \to S_X$ by taking $\phi(g)$ to be the function on $X$ defined by $(\phi(g))(x) = gx$. Then $\phi(g)$ is in $S_X$ because $\phi(g^{-1})$...
is seen to be the inverse of $\phi(g)$. The map $\phi$ is a homomorphism since
\[ \phi(gh)(x) = (gh)(x) = g(hx) = \phi(g)(\phi(h)(x)) = (\phi(g) \circ \phi(h))(x). \]
Now $\phi$ is injective since if $\phi(g) = \phi(h)$ then $g = \phi(g)(1) = \phi(h)(1) = h$. Hence $G$ is isomorphic to $\phi(G) \subseteq S_X$ and $X$ is finite if $G$ is finite. □

2.11 Direct products and finite Abelian groups

We have analyzed groups in terms of subgroups and homomorphisms or quotients. Another tool in understanding groups is illustrated by an example. Consider the additive group integers mod 12, $G = \mathbb{Z}_{12}$. This has as subgroups $H = \langle 4 \rangle$ consisting of the 3 different equivalence classes of multiples of 4, and $K = \langle 3 \rangle$ consisting of the 4 different equivalence classes of multiples of 3. Since $G$ is Abelian, $H$ and $K$ are normal subgroups, and $G/H \cong \mathbb{Z}_4$ and $G/K \cong \mathbb{Z}_3$ the isomorphisms realized by taking the equivalence classes of representative elements modulo 4 and 3 respectively. Now the equivalence class of a number modulo 12 is determined if we know the equivalence classes of the number modulo 4 and 3. In terms of $G$, we have that $H \cap K = \{0\}$, each $H$ coset has a unique representative from $K$ and each $K$ coset has a unique representative of $H$, and each element of $G$ can be uniquely represented as a sum of an element of $H$ and an element of $K$. Switching to multiplicative notation, we make the following definition.

**Definition 2.11.1** If $G$ is a group with normal subgroups $H$ and $K$ such that, for each element $g \in G$, there exists a unique element $h \in H$ and a unique element $k \in K$ with $g = hk$, then we say $G$ is an internal direct product of normal subgroups $H$ and $K$.

Then we have the following result.

**Theorem 2.11.1** Suppose $G$ is an internal direct product of normal subgroups $H$ and $K$. Then $H \cap K = \{1\}$, and for each $h \in H$ and $k \in K$, $hk = kh$. Moreover, $K \cong G/H$ and $H \cong G/K$.

**Proof:** If $g \in H \cap K$ then $g = hk$ with $h = g$ and $k = 1$ or else with $h = 1$ and $k = g$. Since elements of $G$ factor uniquely as products of elements of $H$ and $K$, $h = g = k = 1$ and $G \cap H = \{1\}$. Suppose $h \in H$ and $k \in K$. Since $H$ and $K$ are normal subgroups $kHk^{-1} = H$ and $h^{-1}Kk = K$. Now
$$kh = (khk^{-1})k$$ is a product of an element of $H$ and an element of $K$ but also $kh = h(h^{-1}kh)$ is such a product. Thus we conclude $(khk^{-1}) = h$ and $k = h^{-1}kh$, that is $kh = hk$. Define a map $f$ of $K$ to $G/H$ by $f(k) = Hk$, the restriction of the quotient map of $G$ onto $G/H$ and hence a homomorphism. Then if $g = hk$, then $Hg = H(hk) = Hk = f(k)$ so $f$ is surjective. On the other hand if $f(k_1) = f(k_2)$, then $Hk_1 = Hk_2$ or $k_1k_2^{-1} \in H$. But $k_1k_2^{-1} \in K$ and $H \cap K = \{1\}$ so $k_1 = k_2$, and $f$ is injective. Thus $K \cong G/H$ and then similarly $H \cong G/K$. $$
abla$$

Starting from groups $H$ and $K$, is there a group $G$ which is an internal direct product of groups isomorphic to $H$ and $K$? We construct such a group as follows.

**Definition 2.11.2** Suppose $H$ and $K$ are groups. Then the (external) direct product of $H$ and $K$ is the group $G = H \times K$ whose elements are those of the Cartesian product with operations defined componentwise by $(h_1,k_1)(h_2,k_2) = (h_1h_2,k_1k_2)$ and $(h,k)^{-1} = (h^{-1},k^{-1})$ and taking identity element $1_G = (1_H,1_K)$.

**Theorem 2.11.2** The direct product $G = H \times K$ of groups is a group. The subgroups $\bar{H} = \{(h,1) : h \in H\}$, and $\bar{K} = \{(1,k) : k \in K\}$ are normal subgroups of $G$ isomorphic to $H$ and $K$ respectively and $G$ is an internal direct product of $\bar{H}$ and $\bar{K}$. Conversely, if a group $G$ is an internal direct product of normal subgroups $H$ and $K$, then $G$ is isomorphic to the direct product $H \times K$.

**Proof:** Clearly $H \times K$ is closed under the multiplication, and inverse operations defined, and contains the identity element chosen. The only thing to check to see that $H \times K$ is a group is that the identities defining groups hold in $H \times K$. But associativity holds because it holds in each component separately. For identity we have $(h,k)(1,1) = (h1,k1) = (h,k) = (1h,1k) = (1,1)(h,k)$ so we have $(1,1)$ an identity. Similarly, inverses work because they work in each component separately.

The subset $\bar{H}$ is a subgroup of $G$ since products and inverses of elements with second component 1 will also have second component 1 and thus stay in $\bar{H}$. Of course, the definition of multiplication is exactly as needed to make the map of $H$ into $\bar{H}$ taking $h$ to $(h,1)$ an isomorphism of $H$ and $\bar{H}$. The map sending $(h,k)$ to $k$ is likewise a homomorphism of $H \times K$ onto $K$ with kernel $\bar{H}$ so $\bar{H}$ is a normal subgroup of $H \times K$. Similarly for $K$ and
Clearly, every element \((h, k) = (h, 1)(1, k)\) of \(H \times K\) is a product of elements from \(\bar{H}\) and \(\bar{K}\), but this representation must also be unique since \((h', k') = (h', 1)(1, k') = (h, 1)(1, k) = (h, k)\) means \(h' = h\) and \(k' = k\) so \(H \times K\) is an internal direct product of subgroups \(\bar{H}\) and \(\bar{K}\).

On the other hand, given a group \(G\) which is an internal direct product of subgroups \(H\) and \(K\), we define a map \(\phi : H \times K \rightarrow G\) by \(\phi(h, k) = hk\). By the uniqueness of representation of elements of \(G\) as products of elements from \(H\) and \(K\), \(\phi\) is a bijection. But \(\phi\) is also a homomorphism since for \(h, h' \in H\) and \(k, k' \in K\),

\[
\phi(((h, k)(h', k'))) = \phi(hh', kk') = hh'kk' = hh'k' = \phi(h, k)\phi(h', k')
\]
since \(kh' = h'k\) by the preceding theorem. □

So internal and external direct products work out to essentially the same thing (up to isomorphism). We usually identify \(\bar{H}\) with \(H\) and \(\bar{K}\) with \(K\) and take the internal direct product of \(H\) and \(K\) as a (somewhat simpler external) direct product of \(H\) and \(K\) and suppress the distinction between internal and external direct products. Note that \((H \times K) \times L\) and \(H \times (K \times L)\) are isomorphic groups and might be more conveniently be represented by componentwise operations on the set of triples from \(H, K,\) and \(L\) we usually work with in the Cartesian product \(H \times K \times L\). Also \(H \times K\) is isomorphic to \(K \times H\).

For finite Abelian groups, this construction is extremely useful. We state the next result without proof as an example of what can be proved.

**Theorem 2.11.3** Suppose \(G\) is a Abelian group of order \(n = p_1^{k_1}p_2^{k_2} \cdots p_n^{k_n}\) with the \(p_i\) the distinct prime factors of \(n\). Then \(G \cong H_1 \times H_2 \times \cdots \times H_n\) where each \(H_i\) is an Abelian group of order \(p_i^{k_i}\). Moreover, each \(H_i\) is isomorphic to a direct product of cyclic groups \(\mathbb{Z}_{p_i^{m_{ij}}}\) where the exponents \(m_{ij}\) of \(p_i\) in the orders of these groups sum to the exponent \(k_i\) (and possibly some \(m_{ij}\) are equal so \(H_i\) may have repeated factor groups).

Thus the 24 element Abelian groups are, up to isomorphism,

\[
\begin{align*}
\mathbb{Z}_2 \times \mathbb{Z}_3 & \cong \mathbb{Z}_{24} \\
\mathbb{Z}_2 \times \mathbb{Z}_6 \times \mathbb{Z}_4 & \cong \mathbb{Z}_6 \times \mathbb{Z}_4 \cong \mathbb{Z}_2 \times \mathbb{Z}_{12} \\
\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_3 & \cong \mathbb{Z}_6 \times \mathbb{Z}_2 \times \mathbb{Z}_2
\end{align*}
\]
2.12 Problems

Problem 2.12.1 Complete the proof that every permutation is a product of transpositions by showing that a cycle can be written \((a_1 a_2 \ldots a_k) = (a_k a_1) \ldots (a_2 a_1)\).

Problem 2.12.2 Check that if \(c, d, x, y\) are distinct elements of \(X\) then in \(S_X\),

1. \((cd)(xy) = (xy)(cd)\),
2. \((cd)(dy) = (cdy) = (dy)(cy)\),
3. \((cd)(cy) = (cyd) = (cy)(dy) = (dy)(cd)\), and
4. \((cd)(cd) = 1\).

Problem 2.12.3 Show that the direct product \(H \times K\) of Abelian groups \(H\) and \(K\) is an Abelian group.

Problem 2.12.4 Show that \(U_{21}\), the multiplicative group of equivalence classes mod 21 having multiplicative inverses, is a direct product of cyclic subgroups of orders 2 and 6 so is isomorphic to \(\mathbb{Z}_2 \times \mathbb{Z}_6\).

Problem 2.12.5 Determine the structure of \(U_{35}\) as a direct product of cyclic subgroups.
3 Rings and fields

We’ve been working with algebras having a single binary operation satisfying conditions reminiscent of arithmetic operations, associativity, identities, and inverses. But ordinary arithmetic has both an addition and a multiplication operation, so we next consider algebras with two binary operations satisfying familiar conditions. We have a theory of groups encompassing more examples than just arithmetic, we can prove general theorems about such algebras, and we can establish general constructions of groups from other groups allowing greater understanding of the structure of examples. With this experience we repeat the same sort of analysis for the new algebras, using examples as a guide to what might be the interesting and useful assumptions, and proving general theorems.

3.1 Definitions and examples

We consider algebras with two binary operations, denoted by $+$ and $\cdot$, and consider conditions on these operations similar to those satisfied by the arithmetic on the integers.

**Definition 3.1.1** A ring is an algebra $\mathbb{R} = \langle R; +, \cdot, -, 0 \rangle$ such that

1. Ignoring multiplication, $\langle R; +, -, 0 \rangle$ is an Abelian group, i.e., $+$ is associative and commutative, $0$ is an identity for $+$ (the additive identity), and $-$ is a unary operation giving inverses of elements for $+$ (the additive inverses of elements).

2. Multiplication is an associative operation, i.e., for all $x, y, z \in R$, $(xy)z = x(yz)$.

3. Multiplication distributes over addition on both the left and right, i.e., for all $x, y, z \in R$, $x(y + z) = (xy) + (xz)$ and $(y + z)x = (yx) + (zx)$.

Note that $-$ and 0 are determined uniquely from assuming $R$ has an identity and inverses for $+$, but we include these as extra operations of the algebra for convenience and these conditions then become identities. We sometimes write $x - y$ for $x + (-y)$. We will take $\cdot$ to have precedence over $+$ so in the distributive law we could have written $xy + xz$ instead of the more explicit $(xy) + (xz)$. 

48
Of course, arithmetic on the integers satisfies these conditions and we may take as a ring $\mathbb{Z} = \langle \mathbb{Z}; +, -, 0 \rangle$ including the multiplication operation along with the (additive) group considered before. Likewise, we can take multiplication modulo $n$ to make $\mathbb{Z}_n$ a ring. The set of all even integers is a different example of a ring, though the additive group of even integers is actually isomorphic to the group of integers, as rings these are different because the multiplication acts differently. The rational numbers, real numbers, and complex numbers are examples of rings (in fact, these are fields, to be defined below). The set of all even integers is a different example of a ring, though the additive group of even integers is actually isomorphic to the group of integers, as rings these are different because the multiplication acts differently. The rational numbers, real numbers, and complex numbers are examples of rings (in fact, these are fields, to be defined below). The set of continuous functions forms a ring, the continuous functions are a subring of this, and the differentiable functions are an even smaller subring. The set of $n \times n$ matrices with integer coefficients is a ring if we take the sum of matrices defined by adding their corresponding entries and take matrix multiplication (or more generally matrices with entries from a given ring). The trivial ring is the ring with one element 0.

Now, most of these examples also have an identity element for multiplication, (even the trivial ring). We always assume an additive identity exists, but we might not want to assume a multiplicative identity exists. Most of these examples have a multiplication operation which is commutative. We always assume addition is commutative, but we might not want to assume a commutative multiplication. If we need these properties then we make explicit the assumptions by referring to restricted classes of rings.

**Definition 3.1.2** A ring with unit is an algebra $\langle R; +, -, 0, 1 \rangle$ which is a ring having an additional constant 1 which is a (two-sided) multiplicative identity (determined uniquely). A commutative ring is a ring in which multiplication is commutative. (And a commutative ring with unit is a ring with unit in which multiplication is commutative, clearly.)

The identity of a ring is the additive identity 0. The multiplicative identity 1, if one exists, is a unit of the ring. (Actually, the units of a ring are those elements having multiplicative inverses, the assumption of a unit in the ring implying the existence of a multiplicative identity. The multiplicative identity might better be called the unity element of the ring.)

We establish some additional basic facts implied by the identities defining rings.

**Theorem 3.1.1** Suppose $R$ is a ring, and $x, y, z \in R$. Then
1. \( x0 = 0x = 0 \).

2. \((-x)y = x(-y) = -(xy)\).

3. \( x(y - z) = xy - xz \) and \((y - z)x = yx - zx\).

4. If \( R \) is a ring with unit, \(-x = (-1)x\).

**Proof:** We check that

\[
x0 = (x0 + x0) + -(x0) = x(0 + 0) + -(x0) = x0 + -(x0) = 0
\]

and a similar proof shows \(0x = 0\). Next we compute

\[
(-x)y = ((-x)y + xy) + -(xy) = (-x + x)y + -(xy) = 0y + -(xy) = -(xy)
\]

and similarly \( x(-y) = -xy \). Of course then

\[
x(y - z) = x(y + -z) = xy + x(-z) = xy + -xz = xy - xz
\]

and the other distributive law for subtraction is similar. Finally if \(1 \in R\) is a multiplicative identity, then \(-x = -(1x) = (-1)x\). □

In an additive group, we write \(nx = x + x + \ldots + x\) a sum of \(n\) \(x\)'s if \(n\) is a positive integer, \(nx = -x + -x + \ldots + -x\) a sum of \(-n\) \(-x\)'s if \(n\) is negative, and \(0x = 0\). In a ring, we have both this shorthand for multiple additions and also a matrix multiplication. Since \(0 \cdot x = 0\) in any ring, and \(1 \cdot x = x\) in any ring with unit, this notation should not usually cause confusion. We may thus write \(2x\) for \(x + x\) in a ring without necessarily implying the ring has an element 2 with \(2 \cdot x = x + x\). However, in a ring with unit, we might as well define 2 = 1+1 since then by distributivity \(2 \cdot x = (1+1) \cdot x = 1 \cdot x + 1 \cdot x = x + x\) (and is also equal to \(x \cdot 2\). We can then define 3 to be \(2 + 1\) in the ring, 4 to be \(3 + 1\), etc., and then \(-1, -2\), and so on to be the additive inverses of these. The only problem in using integers to label corresponding elements of a ring with unit is that these might not all be different, some positive \(m\) might be the same as 0 in \(R\). But this is just what we have in \(\mathbb{Z}_m\). What we have established is summarized in the following result (except that we haven’t defined subring and ring isomorphism yet) and is incorporated into our notation.
Theorem 3.1.2 If $R$ is a ring with unit, then the map taking the integer 1 to the multiplicative identity of $R$ extends to an additive group homomorphism of $\mathbb{Z}$ into $R$. Let $\bar{n}$ be the image of $n$ under this homomorphism. Then the image of this homomorphism is a subring of $R$ isomorphic to $\mathbb{Z}$ or $\mathbb{Z}_m$ for some positive integer $m$. For each integer $n$, taking $nx$ for an $x \in R$ as in an additive group agrees with the multiplication in $R$ of the element $\bar{n}$ and $x$ (in either order), i.e., $nx = \bar{n} \cdot x = x \cdot \bar{n}$ (and in practice we will write $\bar{n}$ as simply $n \in R$ instead of $\bar{n}$, just as $\bar{0} = 0$ and $\bar{1} = 1$ are the usual symbols denoting additive and multiplicative identities in a ring while still allowing different rings to have different additive and multiplicative identities).

Suppose $R$ is a commutative ring with unit. The set of polynomials in a variable (symbol) $x$ with coefficients from $R$, denoted $R[x]$, is the set all expressions of the form $a_0 + a_1x + a_2x^2 + \ldots + a_nx^n = \sum_{i=0}^{n} a_ix^i$ for a nonnegative integer $n$ and $a_i \in R$ where we consider polynomials equal if they differ only by terms $a_ix^i$ with $a_i = 0$. If $R$ is an ordinary arithmetic ring $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, or even $\mathbb{C}$, this agrees with our usual understanding of polynomials. Now addition of polynomials is understood by adding coefficients of like powers of $x$, essentially using the distributive law $a_ix^i + b_ix^i = (a_i + b_i)x^i$ as if $x$ were an element of $R$ (which it is not). The multiplication of polynomials is usually explained by multiplying out the product term by term using the distributive law and then collecting like powers of $x$. A general formula might be given as

\[
\left( \sum_{i=0}^{m} a_ix^i \right) \left( \sum_{j=0}^{n} b_jx^j \right) = \sum_{k=0}^{m+n} \left( \sum_{i+j=k} a_ib_j \right) x^k
\]

with the inner sum limited to $i$ and $j$ for which $a_i$ and $b_j$ are defined coefficients of the polynomials on the left. Without too much difficulty, one can verify that $R[x]$ is again a commutative ring with unit. The notation $R[x]$ is sometimes read $R$ adjoined $x$, since a new element $x$ is added to the ring $R$ and then closed under additions and multiplications to make a new ring. The $x$ is called an indeterminant, it has no specific value in $R$, and we are considering polynomials as distinct elements even if they might be equal as functions of $R$ into $R$ if, for $x \in R$, they always evaluated in $R$ to the same value.

A similar notation is used when we actually have a definite value (in some larger ring) in mind for the $x$. As a specific example, if $\alpha \in \mathbb{C}$ satisfies a polynomial equation over a ring with unit $R$ which is a subring of the
complex numbers, then we write $R[\alpha]$ for the ring with unit of all values in $\mathbb{C}$ of polynomials in $\alpha$ with coefficients from $R$. Thus $\mathbb{Q}[\sqrt{2}]$ is the set of all numbers of the form $a + b\sqrt{2}$ with $a$ and $b$ rational (and since $\sqrt{2}^2 = 2$ we do not need powers of $\sqrt{2}$ greater than 1, but $\sqrt{2}$ is not rational so we need at least the first power of $\sqrt{2}$ to extend the rationals). Then $\mathbb{Q}[\sqrt{2}]$ is a set of real numbers closed under addition, multiplication, and negation and containing 0 and 1 (and in fact this ring is also a field).

### 3.2 Integral domains and fields

If our inspiration in studying rings is to understand generalizations of the arithmetic of the integers or the rational, real, or complex numbers, then an important property we need to consider is the following.

**Definition 3.2.1** An integral domain is a commutative ring with unit $R$ such that, for any $a, b \in R$, if $ab = 0$ then $a = 0$ or $b = 0$. Nonzero elements $a$ and $b$ of a ring such that $ab = 0$ are called zero divisors, so another way of stating this definition is to say that an integral domain is a commutative ring with unit having no zero divisors.

The integers are an integral domain, but so is the ring $\mathbb{Q}[x]$, with the important consequence that if a polynomial $p$ factors as $p(x) = q(x)r(x)$ then the roots of $p$ are just the roots of $q$ and those of $r$. The integers modulo 6 is not an integral domain since $2 \cdot 3 = 6 \equiv 0 \mod 6$. More generally, for the integers modulo $n$ to be an integral domain we must have that if $ab$ is a multiple of $n$, then $a$ or $b$ is a multiple of $n$. But this is true exactly when $n$ is a prime number. In an integral domain, we can cancel a nonzero factor from both sides of an equation, in fact this condition is equivalent to no zero divisors.

**Theorem 3.2.1** Suppose $R$ is commutative ring with unit. Then $R$ is an integral domain iff, for any $a \neq 0$ in $R$ and any $x, y \in R$, $ax = ay$ implies $x = y$.

**Proof:** Since $ax = ay$ iff $a(x - y) = ax - ay = 0$, if $R$ has no nonzero divisors, then $ax = ay$ implies $a = 0$ or $x - y = 0$ which if $a \neq 0$ must mean $x = y$. Conversely, if $ax = ay$ implies $x = y$ for any $a \neq 0$, then when $ab = 0$ either $a = 0$ or $ab = 0 = a0$ making $b = 0$. □
In the rational, real, or complex numbers, we have the additional property that nonzero elements have multiplicative inverses. We distinguish the rings having this property.

**Definition 3.2.2** A nontrivial commutative ring with unit $R$ is a field if, for any nonzero $a \in R$, there exists an $x \in R$ with $ax = 1$.

In a field, $0 \neq 1$. For $a \neq 0$, an $x$ with $ax = 1$ is nonzero, and if $ab = 0$ then $b = xab = x0 = 0$, so a field is an integral domain. Thus the nonzero elements of a field form a multiplicative group, and the multiplicative inverse of a nonzero element is unique and denoted by $a^{-1}$. The rational numbers are our basic example of a field and we can solve linear equations in a field the same way we do over the rationals. To solve a polynomial equation we may need a field extending the rationals, for example, the polynomial equation $x^2 - 2 = 0$ has no solution in the rationals but will in the larger field $\mathbb{Q}[\sqrt{2}]$. We also have finite examples of fields, in fact we have the following result.

**Theorem 3.2.2** A finite commutative ring with unit is a field iff it is an integral domain.

**Proof:** Every field is an integral domain, so assume $R$ is a finite integral domain and suppose $a \neq 0$. If $ax = ax'$ then $x = x'$. Hence the map of $R$ to $R$ mapping $x$ to $ax$ is an injective map of a finite set to itself and hence a permutation of $R$. So there exists an $x$ with $ax = 1$, and $x$ is the multiplicative inverse of $a$. Each $a \neq 0$ has a multiplicative inverse so $R$ is a field. □

The map of $\mathbb{Z}$ sending $x$ to $2x$ is injective but not surjective which is possible only because $\mathbb{Z}$ is infinite.

### 3.3 Homomorphism theorems

Generalizing the notion of group homomorphism we have the following.

**Definition 3.3.1** For rings $R$ and $S$, a function $f : R \to S$ is a ring homomorphism if, for all $x, y \in R$, $f(x + y) = f(x) + f(y)$ and $f(xy) = f(x)f(y)$. (Note then that also, as for groups, $f(-x) = -f(x)$ and $f(0) = 0$.) The kernel of a ring homomorphism $f : R \to S$ is $\ker(f) = \{ r \in R : f(r) = 0 \}$. 

53
As with groups, an isomorphism of rings is a bijective homomorphism of rings. A ring homomorphism is an additive group homomorphism of the additive group structures of the rings. To understand ring homomorphisms we proceed in analogy with the group case. We would next like to analyze the kernels of ring homomorphisms. The kernel of a homomorphism is of course a subgroup of the additive group structure, but we should expect more. Every subgroup of the Abelian additive group of a ring is a normal subgroup and so the kernel of a group homomorphism, but these will not all be kernels of ring homomorphisms.

**Definition 3.3.2** A subring of a ring $R$ is a nonempty subset $S \subseteq R$ such that, for all $x, y \in S$, $x + y \in S$, $-x \in S$, and $xy \in S$ (and we get $0 \in S$ but do not require $1 \in S$ if $R$ is a ring with unit). An ideal of a ring $R$ is a nonempty subset $I \subseteq R$ which is a subring of $R$ and is such that, for every $x \in I$ and every $r \in R$, $rx \in I$ and $xr \in I$ (sometimes called a two-sided ideal since we can multiply by an arbitrary $r \in R$ on either left or right to get $rI \subseteq I$ and $Ir \subseteq I$).

Now kernels are not just subrings but ideals.

**Theorem 3.3.1** Suppose $f : R \rightarrow S$ is a ring homomorphism. The image of $f$ is a subring $f(R) \subseteq S$. The kernel of $f$ is an ideal $I = \ker(f)$ and elements $x, y \in R$ map to the same element of $S$ iff they belong to the same (additive group) coset of $I$, i.e., iff $I + x = I + y$.

**Proof:** For $x, y \in I = \ker f$ and $r \in R$ we have $f(x) = f(y) = 0$ so $f(x + y) = f(x) + f(y) = 0$, $f(-x) = -f(x) = 0$, and $f(xy) = f(x)f(y) = 0$ but more generally $f(rx) = f(r)f(x) = f(r)0 = 0$ and similarly $f(xr) = 0$. Hence $x + y$, $-x$, $xy$ and more generally $rx$ and $xr$ all belong to the kernel of $f$. Since $0 \in \ker f$, the kernel of $f$ is a subring but also an ideal of $R$. If $s_1, s_2 \in f(R) \subseteq S$ then there are $r_1, r_2 \in R$ with $f(r_1) = s_1$ and $f(r_2) = s_2$ so $s_1 + s_2 = f(r_1 + r_2)$, $-s_1 = f(-r_1)$, $s_1s_2 = f(r_1r_2)$, and $0 = f(0)$ all in $f(R)$ a subring of $S$. Now $f(x) = f(y)$ iff $f(x + -y) = f(x) + -f(y) = 0$, that is iff $x + -y \in I$, which is to say iff $I + x = I + y$. □

Now it will be that any ideal of a ring is the kernel of some homomorphism. If a ring homomorphism maps each coset of its kernel to a different coset, then to find a homomorphism with a given ideal we should consider again the definition of operations on the set of cosets generalizing the definition of a quotient group.
**Definition 3.3.3** Suppose $R$ is a ring and $I$ is an ideal of $R$. Define $R/I = \{I + r : r \in R\}$ to be the set of cosets of $I$ in $R$, with operations defined using representative elements of the cosets by

$$(I + a) + (I + b) = I + (a + b)$$

$$-(I + a) = I + (-a)$$

$$(I + a)(I + b) = I + ab$$

and taking $0$ in $R/I$ to be the coset $I + 0$. The map $f : R \to R/I$ defined by $f(r) = I + r$ is called the quotient map of $R$ onto $R/I$.

That these operations are well-defined is included in the following theorem.

**Theorem 3.3.2** Suppose $R$ is a ring and $I$ is an ideal of $R$. Then the operations on $R/I$ defined above by taking representative elements are well-defined. The quotient $R/I$ is a ring. The quotient map $f : R \to R/I$ is a ring homomorphism with kernel $I$.

**Proof:** There are a number of things to check. Suppose $I + a = I + a'$ and $I + b = I + b'$ so $a$ and $a'$ represent the same coset and $b$ and $b'$ represent the same coset. To check the operations defined only depend on the coset we check that using $a'$ and $b'$ in place of $a$ and $b$ in these definitions would give the same result. Note that $a - a'$ and $b - b' \in I$. Then $I + (a + b) = I + (a' + b')$ since $(a + b) - (a' + b') = (a - a') + (b - b') \in I$. Also $I + (-a) = I + (-a')$ since $(-a) - (-a') = -(a - a') \in I$. The trickier check is that $I + (ab) = I + (a'b')$ since $ab - a'b' = a(b - b') + (a - a')b' \in I$ as both of the last terms are $I$ as $I$ is an ideal (and not just a subring). Thus $R/I$ is a well-defined algebra. Does it satisfy the identities defining a ring? Typical elements of $R/I$ are $I + x$, $I + y$, and $I + z$ for $x, y, z \in R$. But then, for example

$$(I + x)((I + y) + (I + z)) = I + (x(y + z)) = I + (xy + xz)$$

$$= (I + xy) + (I + xz)$$

$$= (I + x)(I + y) + (I + x)(I + z)$$

showing that the distributive law works in $R/I$ since it holds in $R$ and operations are defined on representative elements. The other identities check similarly. The kernel of the quotient map is the set of $x \in R$ with $I + x = I + 0$ the $0$ element of $R/I$. But $I + x = I$ iff $x \in I$, so the kernel of the quotient map is just $I$. □

This leads us to the first homomorphism theorem for rings.
**Theorem 3.3.3** (First homomorphism theorem for rings) Let \( \phi : R \to S \) be a ring homomorphism, \( I \) the kernel of \( \phi \), and \( \psi : R \to R/I \) be the quotient map of \( R \) onto \( R/I \). Then the subring \( \phi(R) \) of \( S \) is isomorphic to \( R/I \). Moreover, there is an isomorphism \( \alpha : R/I \to \phi(R) \) such that \( \phi = \alpha \circ \psi \).

**Proof:** Define \( \alpha : R/I \to \phi(R) \) by taking \( \alpha(I + r) = \phi(r) \). Then \( \alpha \) is well-defined since if \( I + r = I + r' \), \( r - r' \in I \) so \( \phi(r) = \phi(r') \) and \( \alpha(I + r) = \alpha(I + r') \).

To check \( \alpha \) is injective reverse each of the implications in the last sentence. Clearly, \( \alpha \) is surjective. We need \( \alpha \) is a ring homomorphism. For \( r, r' \in R \),

\[
\alpha((I + r)(I + r')) = \alpha(I + (rr')) = \phi(rr') = \phi(r)\phi(r') = \alpha(I + r)\alpha(I + r')
\]

and similarly for addition and negation. Finally \( (\alpha \circ \psi)(r) = \alpha(I + r) = \phi(r) \) so \( \phi = \alpha \circ \psi \) for \( \psi \) the quotient map taking \( r \) to \( I + r \). \( \square \)

As for groups there is a second and a third homomorphism theorem giving the correspondence between subrings and ideals in the domain and image rings of a homomorphism and the relation between quotients by comparable ideals. We will be satisfied with simply stating these results.

**Theorem 3.3.4** (Second homomorphism theorem for rings) Suppose \( \phi : R \to S \) is a surjective ring homomorphism with kernel \( I \) (so \( R/I \) is isomorphic to \( S \)). Then there is a 1-1 correspondence between the subrings of \( R \) containing \( I \) and the subrings of \( S \), a subring \( T \) of \( R \) containing \( I \) corresponding to \( T' = \phi(T) \) and a subring \( T' \) of \( S \) corresponding to \( T = \phi^{-1}(T') = \{r \in R : \phi(r) \in T'\} \). Moreover for corresponding subrings \( T \subseteq R \) and \( T' \subseteq S \), \( T/I \cong T' \) (by the restriction of the isomorphism of \( R/I \) and \( S \) to the subring \( T/I \)). The ideals of \( R \) containing \( I \) correspond to the ideals of \( S \) by this correspondence.

**Theorem 3.3.5** (Third homomorphism theorem for rings) Suppose \( \phi : R \to S \) is a surjective ring homomorphism with kernel \( I \). Suppose that \( J \) is an ideal of \( R \) containing \( I \) and \( J' = \phi(J) \) is the corresponding ideal of \( S \). Then \( R/J \) is isomorphic to \( S/J' \). In terms of quotient rings, we have \( R/J \cong (R/I)/(J/I) \).

### 3.4 Problems

**Problem 3.4.1** Let

\[
R = \{f : \mathbb{R} \to \mathbb{R} : f(x + y) = f(x) + f(y) \text{ for all } x, y \in \mathbb{R}\}
\]
Take \( f + g \) defined by \( (f + g)(x) = f(x) + g(x) \) (pointwise addition). Show that if \( f, g \in R \) then \( f + g \in R \). Show that the composition of two functions from \( R \) is again in \( R \). Show that \( R \) is a ring under pointwise addition (as \( + \)) and composition (as \( \cdot \)) (with appropriate \( - \) and \( 0 \)).

**Problem 3.4.2** Suppose \( R \) is a ring with unit. Show that \( R \) is the trivial ring iff \( 0 = 1 \) in \( R \).

**Problem 3.4.3** The hardest identity to verify in showing that \( R[x] \) is a ring with unit is probably the distributive law. Suppose \( a = \sum_{i=0}^{m} a_i x^i \), \( b = \sum_{j=0}^{n} b_j x^j \) and \( c = \sum_{j=0}^{n} c_j x^j \) (here extending \( b \) and \( c \) by 0 terms so they end at the same \( x^n \)). Prove \( a(b + c) = ab + ac \).

**Problem 3.4.4** Show that \( \mathbb{Q}[\sqrt{2}] \) is a field.

**Problem 3.4.5** Show that, in a commutative ring with unit, if the product of nonzero elements \( a \) and \( b \) is zero, then there is no \( x \) in the ring with \( ax = 1 \).

**Problem 3.4.6** Nothing is really a theorem until it has a proof. One rarely understands a theorem until they have worked through the proof for themselves. As a project, write out a proof of the second homomorphism theorem for rings. Good writing counts here, explain what needs to be proved as well as calculate why these things are true. You especially need to check that my statement of the theorem is accurate, no assumptions are left out and no conclusions are incomplete or wrong (my confidence is \(< 100\% \) that there isn’t a careless error that you’ll need to correct).
3.5 More on fields

While we have considered rings in general, fields might be considered more interesting (e.g., because they have multiplicative inverses of nonzero elements, more kinds of equations can be solved). Supposing we start with a ring, might we find a field containing that ring? Clearly, the ring must be commutative. We will assume that there is already a unit element in the ring. The ring cannot have zero divisors since a zero divisor cannot also have a multiplicative inverse (an exercise above). So we must start with an integral domain $R$. For any nonzero $b$ in $R$ we will need some $b^{-1}$ in our field, but then we will also need to be closed under addition, negation, and products as well as multiplicative inverses. For example, if we start with the integers, we must add an element for $2^{-1} = 1/2$, and then also $-1/2$, and $3 \cdot 2^{-1} = 3/2$ etc. On the other hand, we already know that $1/2 + 1/2 = 2 \cdot 2^{-1} = 1$ so we must define some combinations of these new elements to be equivalent. Of course, this idea applied to the integers leads to the rational numbers. The basic result is that addition, negation, and multiplication of fractions, and multiplicative inverse of nonzero fractions, all give rise to other fractions provided we remember to identify fractions that have the same reduced form. We generalize this idea to an arbitrary integral domain.

**Definition 3.5.1** Suppose $R$ is an integral domain (but $R \neq \{0\}$). Define an equivalence relation on the set of ordered pairs $\{(a, b) : a, b \in R, b \neq 0\}$ by $(a, b) \sim (c, d)$ if $ad = bc$ in $R$, and write $a/b$ for the equivalence class of $(a, b)$. The field of quotients (or fractions) of $R$ is the set $F = \{a/b : a, b \in R, b \neq 0\}$ with operations and constants defined by

\[
\begin{align*}
(a/b) + (c/d) &= (ad + bc)/(bd) \\
-(a/b) &= (-a)/b \\
0 &= 0/1 \\
(a/b)(c/d) &= (ac)/(bd) \\
1 &= 1/1 \quad \text{and additionally we write}
\end{align*}
\]

\[
(a/b)^{-1} = b/a \quad \text{(for} \ a \neq 0 \text{)}
\]

Here we are thinking of ordered pairs $(a, b)$ with $b \neq 0$ as representing fractions including all improper (unreduced) fractions. Rather than determine unique reduced fraction representatives, we need only know the equivalence classes of fractions that must be taken equal, writing $a/b$ for the set of all
equivalent pairs to \((a, b)\), so we get \(a/b = c/d\) when \(ad = bc\). This definition requires we check that \(\sim\) is an equivalence relation and that the operations defined for representatives of equivalence classes are in fact well-defined operations on the equivalence classes (i.e., where we write \(a/b\) we must not assume that \(a\) and \(b\) are determined uniquely, only that \((a, b)\) is one element of \(a/b\)). This is included in the following result.

**Theorem 3.5.1** Suppose \(R\) is an integral domain, \(R \neq \{0\}\). The relation defined on ordered pairs from \(R\) by \((a, b) \sim (c, d)\) if \(ad = bc\) in \(R\) is an equivalence relation. The operations defined above for representative elements of equivalence classes define operations on equivalence classes of this equivalence relation, that is, the field of quotients of \(R\) defined above, \(F\), is a well-defined algebra. Then \(F\) is a field. The map \(\phi : R \rightarrow F\) defined by \(\phi(r) = r/1\) is an injective ring homomorphism of \(R\) into \(F\). Thus \(R \cong \phi(R) \subseteq F\) and by identifying \(R\) with \(\phi(R)\) we may take \(R\) to be a subring of the field of quotients of \(R\).

**Proof:** We check \(\sim\) is an equivalence relation. Since \(ab = ba\), \((a, b) \sim (a, b)\). Since \(ad = bc\) implies \(cb = da\), \((a, b) \sim (c, d)\) implies \((c, d) \sim (a, b)\). Now assume \((a, b) \sim (c, d)\) and \((c, d) \sim (e, f)\). Then \(ad = bc\) and \(cf = de\) so \(adf = bcf = bde\), and since \(daf = d-be\) in an integral domain implies \(af = be\), \((a, b) \sim (e, f)\).

Next we check that the operations defined above on representatives are well-defined operations on equivalence classes. Suppose \((a, b) \sim (a', b')\) and \((c, d) \sim (c', d')\) so \(ab' = a'b\) and \(cd' = c'd\). Then \((ad + bc)b'd' = adbd' + bcbd' = a'd'bd + b'cd'b = (a'd' + b'c')bd\) so \((ad + bc, bd) \sim (a'd' + b'c', b'd')\) and addition is well-defined. Similarly \((-a)b' = -ab' = -a'b = (-a')b\) so \((-a, b) \sim (-a', b')\) and negation is well-defined. Next \(acbd' = a'c'bd\) so \((ac, bd) \sim (a'c', b'd')\) and multiplication is well-defined. Multiplicative inverse is well-defined since if \(a \neq 0\) then \(ab' \neq 0\) so \(a'b \neq 0\) and \(a' \neq 0\) and \(b' = ba'\) means \((b', a') \sim (b, a)\).

We may now write \(a/b\) for the \(\sim\) equivalence class of \((a, b)\).

Next we check the set of equivalence classes with these operations is a field. We check associativity, commutativity, identity, and inverses for addition, zero, and negation, associativity, commutativity, identity, and inverses for multiplication, one, and reciprocal, and distributivity of multiplication over addition.
\[
\left(\frac{a}{b} + \frac{c}{d}\right) + \frac{e}{f} = \frac{ad + bc}{bd} + \frac{e}{f} = \frac{(ad + bc)f + (bd)e}{bdf} = \frac{adf + bcf + bde}{bdf} = \frac{a(df) + b(cf + de)}{bdf} = \frac{a}{b} + \frac{e}{f} + \frac{c}{d} + \frac{e}{f}
\]

\[
\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} = \frac{cb + da}{db} = \frac{c + a}{d - b}
\]

\[
\frac{a}{b} + \frac{0}{1} = \frac{a + b0}{b1} = \frac{a}{b}
\]

\[
\frac{a}{b} + \frac{-a}{b} = \frac{a + (-a)}{b} = \frac{0}{bb} = \frac{0}{1} (\text{since } 0 \cdot 1 = 0 = 0 \cdot bb)
\]

\[
\left(\frac{a}{b}, \frac{c}{d}\right) \frac{e}{f} = \frac{ac}{bd} \frac{e}{f} = \frac{ace}{bdf} = \frac{a}{b} \left(\frac{c}{d} \frac{e}{f}\right)
\]

\[
\frac{ac}{bd} = \frac{ca}{db} = \frac{c}{d} \frac{a}{b}
\]

\[
\frac{a1}{b1} = \frac{a1 + b0}{b1} = \frac{a}{b}
\]

\[
\frac{a}{b} \left(\frac{a}{b}\right)^{-1} = \frac{ab}{ba} = \frac{ab}{ba} = \frac{1}{1} (\text{for } a \neq 0, \text{ since } ab \cdot 1 = ab = ba \cdot 1)
\]

\[
\left(\frac{a}{b} + \frac{c}{d}\right) \frac{e}{f} = \frac{ad + bc}{bd} \frac{e}{f} = \frac{(ad + bc)e}{bdf} = \frac{(ad + bc)e}{bdf} (\text{since } (ad + bc)e(bdf) = (ad + bc)e(f(bdf)))
\]

\[
= \frac{ade + bce}{bdf} = \frac{ae + c}{bf + df} = \frac{a}{b} + \frac{c}{d} + \frac{e}{f}
\]
Next we check $\phi : R \to F$ defined by $\phi(r) = r/1$ is an injective ring homomorphism. We have $\phi(r+s) = (r+s)/1 = (r1+s1)/(11) = r/1 + s/1 = \phi(r) + \phi(s)$, $\phi(-r) = (-r)/1 = -(r/1) = -\phi(r)$, and $\phi(rs) = (rs)/1 = (r/1)(s/1) = \phi(r)\phi(s)$. If $\phi(r) = \phi(s)$ then $r/1 = s/1$ so $r = r1 = s1 = s$. Thus $R \cong \phi(R) \subseteq F$ and we may identify $R$ with $\phi(R)$. \[\square\]

Instead of finding a field containing a given ring, might we find a field that is a quotient of the ring? This is how we get the integers modulo $n$, as a quotient of the integers by the ideal $n\mathbb{Z}$. We have that $\mathbb{Z}_n$ is a field iff $n$ is a prime. What conditions on the ideal correspond to the quotient ring being a field? Again we should start with a nontrivial commutative ring with unit. If we have zero divisors, then they will end up identified with 0 in the quotient ring if that quotient is a field, so we needn’t assume now that we have an integral domain.

**Definition 3.5.2** An ideal $I$ of a commutative ring with unit $R$ is called maximal if $I \neq R$ and, for any ideal $J$ containing $I$, $J = I$ or $J = R$, i.e., $I$ is maximal in the set of proper ideals of $R$ ordered by inclusion (and not simply of maximal size).

**Theorem 3.5.2** Suppose $R$ is a commutative ring with unit, and $I$ is an ideal of $R$. Then $R/I$ is a field iff $I$ is a maximal ideal of $R$. Hence also $R$ is a field iff the only ideals of $R$ are $\{0\}$ and $R$.

**Proof:** Suppose $I$ is a maximal ideal. For $I + a \neq I + 0$, so $a \notin I$, let $J = Ra + I = \{ra + s : r \in R, s \in I\}$. Then $J$ is an ideal of $R$ since $(ra + s) + (r'a + s') = (r + r')a + (s + s')$, $-(ra + s) = (-r)a + -s$, and $r'(ra + s) = (r'r)a + r's$ and for $r, r' \in R$, $s, s' \in I$, $s + s'$, $-s$, and $r's$ in $I$. But then $I \subseteq J$ but $a \in J$ so $J \neq I$. Since $I$ is maximal, $J = R$ and $1 \in J$. Thus for some $r \in R$ and $s \in I$, $1 = ra + s$ so $(I + r)(I + a) = I + 1$ and $I + r$ is the multiplicative inverse of $I + a$. Thus $R/I$ is a field.

Conversely, assume $R/I$ is a field. Suppose $J$ is an ideal of $R$, $I \subseteq J$, but $I \neq J$. Then there is an $a \in J - I$ so $I + a \neq I$ and $I + a$ has a multiplicative inverse in $R/I$, say $(I + r)(I + a) = I + 1$. Then $ra + s = 1$ for some $s \in I$, and $1 \in Ra + I \subseteq J$. But then $R = R1 \subseteq J$ so $J = R$ and $I$ was a maximal ideal. \[\square\]

For example, in the integers, the maximal ideals are exactly the ideals $\mathbb{Z}p$ for $p$ a prime number. The quotient by $\mathbb{Z}p$ is the ring $\mathbb{Z}/\mathbb{Z}p = \mathbb{Z}_p$ of integers mod $p$. To show $\mathbb{Z}_p$ is a field using the theorem, we would need to
show $I = \mathbb{Z}p$ is a maximal ideal. To see this, suppose $J$ is an ideal with $I \subseteq J \subseteq \mathbb{Z}$. Suppose $I \neq J$ so there is some number in $J - I$, but not only negative numbers in $J - I$ since both $I$ and $J$ are closed under negation, and not 0, so take the smallest positive $n \in J - I$. If $n > p$ then dividing $p$ into $n$ we get $n = qp + r$ for some quotient $q$ and remainder $r$ with $0 \leq r < p$. But then $r = n - qp \in J$, but not in $I$ else $n \in I$, contradicting the minimality of $n$. So $n < p$ and dividing $n$ into $p$, $p = qn + r$ for some $q$ and $r$ with $0 \leq r < n$. Again $r = p - qn \in J$ but if $r \neq 0$ then $r \notin J$ contradicting the minimality of $n$. Hence $r = 0$ and $n < p$ divides evenly into $p$. But if $p$ is prime then $n = 1$, and $J$ contains $\mathbb{Z}n = \mathbb{Z}$, that is $J = \mathbb{Z}$ and $I$ was a maximal ideal.

As another example, consider the ring of Gaussian integers $\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$. The ideal $I = (\mathbb{Z}[i]) \cdot 5$ is not maximal since $5 = (1+2i)(1-2i)$ and so $J = (\mathbb{Z}[i]) \cdot (1 + 2i)$ contains $I$ but is not $\mathbb{Z}[i]$ as $1 \in J$ since $1/(1 + 2i) = (1 - 2i)/5 \notin \mathbb{Z}[i]$. In the Gaussian integers, 5 is not a prime but factors into primes $(1 + 2i)$ and $(1 - 2i)$. Now $J$ is a maximal ideal in $\mathbb{Z}[i]$ and $\mathbb{Z}[i]/J$ is a field. We might enumerate the cosets of $J$ in $\mathbb{Z}/J$ checking each nonzero element of the quotient has a multiplicative inverse. Thus $J$ consists of integer combinations of $(1 + 2i)$ and $i(1 + 2i) = (-2 + i)$ a square grid of integer lattice points in the complex plane. The different cosets then are $J$, $J + 1$, $J + i$, $J + -1$, and $J + -i$, and nothing else since these translations of the $J$ grid cover each integer lattice point of the complex plane. Since $i \cdot -i = -1 \cdot -1 = 1$ these give a field.

The last two examples above have the property that every ideal in the ring is generated by a single element. The usefulness of an algorithm for dividing one element into another leaving a (somehow) smaller remainder in finding a generator for an ideal is illustrated above. Finally, the idea of factoring elements into irreducible elements and the correspondence between irreducible elements and maximal ideals was exploited. How these properties are related will be considered next.

### 3.6 Euclidean domains and principal ideal domains

Given polynomials $a(x)$ and $b(x)$ in $\mathbb{Q}[x]$, we can divide $a(x)$ by $b(x)$, repeatedly subtracting a rational times a power of $x$ multiple of $b(x)$ away from $a(x)$ until getting a remainder $r(x)$ of 0 or a polynomial of degree less than that of $b(x)$ (since otherwise subtracting a multiple of $b(x)$ from $r(x)$ would have given a smaller degree $r(x)$). That is, $a(x) = q(x)b(x) + r(x)$ for a
polynomial quotient $q(x)$ and a remainder polynomial $r(x)$. This generalizes division of integers except that the degree of the remainder polynomial is the measure of the size of the remainder instead of the absolute value of the remainder polynomial. In general, we hypothesize the existence of an appropriate remainder measuring function.

**Definition 3.6.1** An integral domain $R$ is a Euclidean domain of there exists a function $d : R \to \mathbb{N}$ such that for all $a, b \in R$ with $b \neq 0$, $d(a) \leq d(ab)$ and there exists $q, r \in R$ such that $a = qb + r$ and $d(r) < d(b)$.

For example, the integers are a Euclidean domain with $d(a) = |a|$. For polynomials, the degree of a polynomial almost works. We could take the more common definition of a Euclidean domain with a function $d$ defined for nonzero elements of the ring and the property that the remainder $r$ is either zero or has $d(r) < d(b)$ so that degree of polynomials would exactly. Instead we simply extend $d$ so $d(0)$ is smaller than $d(r)$ for nonzero $r$ perhaps by taking $d(0) = 0$ and adding 1 to all other $d(r)$. Thus defining $d(a(x))$ to be 0 of $a(x) = 0$ and $d(a(x)) = 1 + \deg(a(x))$ otherwise shows that polynomial rings over fields are all Euclidean domains.

In considering ideals in the integers, we noted that any ideal in $\mathbb{Z}$ is of the form $\mathbb{Z}n$, all multiples of $n$, $\pm n$ generating the same ideal. The ideal $\mathbb{Z}m$ is contained in $\mathbb{Z}n$ iff $n$ divides $m$. In general, for elements $a$ and $b$ of an integral domain $R$, ideals $Ra$ and $Rb$ are the same iff $a = ub$ for an element $u$ having a multiplicative inverse in $R$, that is $u$ a unit of $R$. The smallest ideal containing elements $a$ and $b$ is $Ra + Rb$. If this ideal is $Rc$, then $c$ is a divisor of both $a$ and $b$, but also of the form $c = ra + sb$ so that if $x$ is any divisor of $a$ and $b$ then $x$ also divides $c$. That is $c$ is (in some sense) a maximal common divisor of $a$ and $b$, with $c$ determined only up to a unit if any such $c$ exists. These ideas are captured in the following.

**Definition 3.6.2** Suppose $R$ is an integral domain and $a$ and $b$ are nonzero elements of $R$. We say $a$ is a unit of $R$ if there exists an $x \in R$ with $ax = 1$. We say $b$ divides $a$ and write $b|a$ if for some $x \in R$, $a = bx$. We say $a$ and $b$ are associates if $a = ub$ for a unit $u \in R$. We say $a$ is irreducible if $a$ is not zero and not a unit such that for any $x, y \in R$, if $xy = a$ then $x$ or $y$ is a unit in $R$. We say $a$ is prime if whenever $a$ divides $xy$, either $a$ divides $x$ or $a$ divides $y$. 63
Theorem 3.6.1 (Euclidean algorithm) Suppose $R$ is a Euclidean domain and $a$ and $b$ are nonzero elements of $R$. Let $a_0 = a$ and $a_1 = b$. If $a_{i-1}$ and $a_i$ are defined let $a_{i-1} = q_i a_i + r_i$ with $d(r_i) < d(a_i)$. If $r_i \neq 0$ let $a_{i+1} = r_i$ otherwise let $c = a_i$ and stop. Then the process eventual stops and defines $c$, such that $c$ is a divisor of both $a$ and $b$, $c = xa + yb$ for some $x, y \in R$, and for any $z$ dividing both $a$ and $b$, $z$ divides $c$.

Proof: If $r_i \neq 0$ then $a_{i+1} = r_i$ so $d(a_{i+1}) = d(r_i) < d(a_i)$. Thus $d(a_1) > d(a_2) > d(a_3) > \ldots$ until some $r_i = 0$. But there can be no infinite decreasing sequence of natural numbers, so some $r_i = 0$ say with $i = n$ and then $c = a_n$.

Now $c$ divides $a_n = c$ and $a_{n-1} = q_n a_n$. If $0 < i < n$ and $c$ divides $a_i$ and $a_{i+1}$ then $c$ divides $q_i a_i + a_{i+1} = q_i a_i + r_i = a_{i-1}$. Continuing in this way (by induction) $c$ divides each $a_i$ so $c$ divides $a_0 = a$ and $a_1 = b$.

Now we see $a_0 = 1a + 0b$, $a_1 = 0a + 1b$, and for all $i < n$, if $a_{i-1} = x_{i-1}a + y_{i-1}b$ and $a_i = x_i a + y_i b$ then

$$a_{i+1} = a_{i-1} - q_i a_i = (x_{i-1} - q_i x_i)a + (y_{i-1} - q_i y_i)b = x_{i+1}a + y_{i+1}b$$

for $x_{i+1} = x_{i-1} - q_i x_i$ and $y_{i+1} = y_{i-1} - q_i y_i$. Hence (by induction) $c = a_n = x_n a + y_n b$. Finally if $z$ divides $a$ and $b$, then $z$ divides any $xa + yb$ and so divides $c$. $\Box$

An ideal in a ring need not simply be all the multiples of some single element, i.e. not all ideals in rings are generated by a single element. The ideal generated by elements $a$ and $b$ in a Euclidean domain is that generated by the greatest common divisor $c$ determined by this algorithm. In general then we have the following.

Definition 3.6.3 Suppose $R$ is an integral domain. An ideal $I$ of $R$ is a principal ideal of $R$ if, for some $a \in R$, $I = Ra$. The ring $R$ is a principal ideal domain if every ideal of $R$ is a principal ideal.

Theorem 3.6.2 If $R$ is a Euclidean domain, then $R$ is a principal ideal domain.

Proof: Suppose $I$ is an ideal of $R$. If $I = \{0\}$ then $I = R0$ is principal. Otherwise take an element $b \in I$ having minimal value $d(b)$ (this since every nonempty set of natural numbers has a smallest element, though $b$ need not be unique). Now $Rb \subseteq I$. If $I \neq Rb$ take $a \in I - Rb$. Write $a = qb + r$ with $d(r) < d(b)$. Then $r = a - qb \in I$ contradicting the minimality of $b$. Hence instead $I = Rb$ is a principal ideal and $R$ is a principal ideal domain. $\Box$

In principal ideal domains, the maximal ideals are easily determined.
**Theorem 3.6.3** If $R$ is a principal ideal domain and $a$ is irreducible, then $Ra$ is a maximal ideal.

**Proof:** Suppose $I = Ra$ with $a$ irreducible. Take an ideal $J$ with $I \subseteq J$. Then $J = Ib$ for some $b$ and $b$ divides $a$, say $a = bc$. Since $a$ is irreducible, either $b$ or $c$ is a unit. If $b$ is a unit then $J = R$ and if $c$ is a unit then $a \in Rb = J$ so $I = J$. Hence $I$ is a maximal ideal. □

A prime element of an integral domain is also irreducible. In a principal ideal domain, irreducible elements are also prime (these are exercises). In the integers, nonzero elements factor uniquely into irreducible elements, up to signs and permutation. A principal ideal domain satisfies the appropriate generalization which we state without proof for comparison.

**Definition 3.6.4** An integral domain is a unique factorization domain if, for every $a \in R$ which is not 0 and not a unit, there exist irreducible elements $p_1, p_2, \ldots, p_n \in R$ with $a = p_1p_2 \ldots p_n$, and if $a = q_1q_2 \ldots q_m$ for irreducibles $q_1, q_2, \ldots, q_m \in R$, then $n = m$ and for some permutation $\sigma$ of $1, \ldots, n$, each $p_i$ is an associate of $q_{\sigma(i)}$.

**Theorem 3.6.4** Every principal ideal domain is a unique factorization domain.

There exist unique factorization domains that are not principal ideal domains and principal ideal domains that are not Euclidean domains.

### 3.7 Problems

**Problem 3.7.1** Show that if $R$ is an integral domain, then $R[x]$ is an integral domain.

**Problem 3.7.2** Let $p(x) = x^4 - 4$ and $q(x) = x^3 + x^2 - 2x - 2$. Determine polynomials $a(x), r(x),$ and $s(x)$ in $\mathbb{Q}[x]$ such that $a(x) = r(x)p(x) + s(x)q(x)$ and $a(x)$ divides $p(x)$ and $q(x)$ (so that the ideal generated by $p(x)$ and $q(x)$

**Problem 3.7.3** Suppose $R$ is an integral domain and $I = Ra$. Then $I = R$ iff $a$ is a unit in $R$.

**Problem 3.7.4** Show that in any integral domain, a prime element $a$ is also irreducible.

**Problem 3.7.5** Show that in a principal ideal domain, an irreducible element $a$ is also prime.
3.8 Polynomial rings

We will need a few additional facts about polynomials over a field. Recall that the degree of a nonzero polynomial \( p(x) \) is the largest power of \( x \) having a nonzero coefficient in \( p(x) \). For nonzero \( a(x), b(x) \in F[x] \), \( \text{deg}(a(x)b(x)) = \text{deg}(a(x)) + \text{deg}(b(x)) \). We call a polynomial monic if this coefficient of the largest power of \( x \) is 1. For a field \( F \) the elements of \( F[x] \) which are units are the nonzero elements of \( F \), hence every polynomial is an associate of a monic polynomial. We summarize the application of some of the earlier ideas to this case.

**Theorem 3.8.1** Suppose \( F \) is a field. Then \( F[x] \) is a Euclidean domain with \( d(p) = \text{deg}(p) + 1 \) and \( d(0) = 0 \) since for any \( a(x), b(x) \in F[x] \) with \( b(x) \neq 0 \), there exists polynomials \( q(x), r(x) \in F[x] \) such that \( a(x) = q(x)b(x) + r(x) \) with \( \text{deg}(r) < \text{deg}(b) \) or \( r(x) = 0 \) when \( b(x) \) is a constant (the division algorithm), and the polynomial of greatest degree dividing nonzero polynomials \( p(x) \) and \( q(x) \) can be computed using the Euclidean algorithm. Hence \( F[x] \) is a principal ideal domain (and a unique factorization domain). Every polynomial \( p(x) \in F[x] \) factors as a product of an element of \( F \) and monic irreducible polynomials in \( F[x] \) (actually uniquely up to the order of factors). Every ideal \( I \) in \( F[x] \) is \( I = (p(x)) \) for a monic polynomial, and the quotient \( F[x]/I \) is a field iff \( p(x) \) is an irreducible polynomial.

These are all easy consequences of our previous results. We need a few additional miscellaneous facts about polynomials.

**Theorem 3.8.2** Suppose \( p(x) \in F[x] \) is a nonzero polynomial over a field. Then \( a \in F \) is a root of \( p(x) \) (i.e., \( p(a) = 0 \) in \( F \)) iff \( p(x) = q(x)(x-a) \) for some polynomial \( q(x) \in F[x] \). If \( \text{deg}(p(x)) = n \) then \( p(x) \) has at most \( n \) different roots.

**Proof:** Suppose \( p(a) = 0 \) and write \( p(x) = q(x)(x-a) + r(x) \) for an \( r(x) \) of degree less than that of \( (x-a) \), i.e., \( r(x) \) a constant. Then \( 0 = p(a) = q(a)0 + r(0) \) so \( r(0) = 0 \) and \( x-a \) is a factor of \( p(x) \). Conversely, if \( p(x) = q(x)(x-a) \) then \( p(a) = q(a)0 = 0 \). To show that a polynomial of degree \( n \) has at most \( n \) roots proceed by induction on \( n \). If \( p(x) \) is a nonzero polynomial of degree 0 then it is a nonzero constant and has no roots. Suppose \( p(x) \) is of degree \( n > 0 \). If \( p(x) \) has no roots in \( F \) then it has at most \( n \) roots. Else if \( a \) is a root of \( p(x) \) then \( p(x) = q(x)(x-a) \) with \( q(x) \) of degree \( n-1 \) and so having
at most \( n - 1 \) roots. Since \( F \) is an integral domain \( p(b) = q(b)(b - a) \) is zero only if \( b = a \) or \( q(b) = 0 \), i.e., \( b \) is a root of \( q(x) \). Hence \( p(x) \) has at most the \( n - 1 \) roots of \( q(x) \) together with \( a \) (if \( a \) wasn’t already a root of \( q(x) \)) so \( p(x) \) has at most \( n \) roots. \( \square \)

If \( p(x) = \sum_{i=0}^{n} a_{i} x^{i} \in F[x] \) then we formally define the derivative of \( p(x) \) to be the polynomial \( p'(x) = \sum_{i=1}^{n} i a_{i} x^{i-1} \) (no limits required). Then the usual rules for derivatives apply \( (p(x) + q(x))' = p'(x) + q'(x) \) and \( (p(x)q(x))' = p'(x)q(x) + p(x)q'(x) \).

**Theorem 3.8.3** If the only polynomials that divide both \( p(x) \) and \( p'(x) \) are constants, then \( p(x) \) does not factor as \( (a(x))^{2}b(x) \) for a nonconstant polynomial \( a(x) \).

**Proof:** Suppose \( p(x) = (a(x))^{2}b(x) \), then
\[
p'(x) = 2a(x)a'(x)b(x) + (a(x))^{2}b'(x)
\]
so \( a(x) \) is a factor of both \( p(x) \) and \( p'(x) \). \( \square \)

### 3.9 Subfields and extensions

We now continue the analysis of fields. Recall that for a ring with unit, we can map the integers into the ring by taking \( n > 0 \) to \( 1 + 1 + \ldots + 1 \) (\( n \) times) and taking \( -n < 0 \) to the negative of this. The image of this map will either be \( \mathbb{Z} \) or isomorphic to a quotient of \( \mathbb{Z} \), so isomorphic to \( \mathbb{Z}_{n} \) for some \( n \).

**Definition 3.9.1** The characteristic of a field \( F \) is the least positive integer \( n \) such that \( 1 + 1 + \ldots + 1 \) (\( n \) times) is equal to 0 in \( F \), if there is any such positive integer, or is 0 if there is no such positive integer.

We say \( F \subseteq E \) is a subfield of a field \( E \), or \( E \) is an extension of \( F \), if \( F \) is a field and a subring of \( E \).

**Theorem 3.9.1** The characteristic of a field \( F \) is either 0 or prime \( p \). If the characteristic is 0, then \( F \) contains a subfield isomorphic to \( \mathbb{Q} \). If the characteristic is a prime \( p \), then \( F \) contains a subfield isomorphic to \( \mathbb{Z}_{p} \).
In either case, this subfield is called the prime subfield of \( F \), and the prime subfield is contained in every subfield of \( F \).
Proof: Suppose the characteristic of $F$ is 0. Then the subring of $F$ generated by 1 is isomorphic to $\mathbb{Z}$. The set of $ab^{-1}$ for $a$ and $b$ is this subring is then isomorphic to $\mathbb{Q}$. Now suppose the characteristic of $F$ is $n > 0$. Then the subring of $F$ generated by 1 is isomorphic to $\mathbb{Z}_n$. Now if $n$ is not prime, then $n = ab$ for $a, b < n$ and $a$ and $b$ correspond to nonzero elements of $F$ which multiply to $n$ which equals 0 in $F$. But then $F$ would have zero divisors and not be a field. Instead $n = p$ is a prime and $F$ has a subfield isomorphic to $\mathbb{Z}_p$. In either case, the subfield constructed was generated by 1 and hence will also be a subfield of any subfield of $F$. □

To understand fields we will use a small amount of linear algebra. It is not hard to generalize the usual real vector space definitions and results to vector spaces over an arbitrary field.

Definition 3.9.2 An algebra $V = (V; +, -, 0, r \cdot (\ ) (r \in F)$ is a vector space over a field $F$ if it is an Abelian group with operations $+$, and $-$, and constant 0, with additional operations defining scalar multiplication by elements $r \in F$ such that

- for all $r \in F$ and $x, y \in V$, $r \cdot (x + y) = r \cdot x + r \cdot y$;
- for all $r, s \in F$ and $x \in V$, $(r + s) \cdot x = r \cdot x + s \cdot y$;
- for all $r, s \in F$ and $x \in V$, $r \cdot (s \cdot x) = (rs) \cdot x$; and
- for all $x \in V$, $1x = x$.

Theorem 3.9.2 If $F$ is a subfield of $E$ then $E$ is a vector space over $F$ with scalar multiplication by elements of $F$ defined by multiplication in $E$.

Definition 3.9.3 Suppose $V$ is a vector space over a field $F$ and $S$ is a subset of $V$. Then a linear combination of elements of $S$ is an element of $V$ of the form $r_1 x_1 + r_2 x_2 + \ldots + r_n x_n$ for some $n$, some coefficients $r_i \in F$, and elements $x_i \in S$ where we may as well take the $x_i$ distinct and the empty sum is taken to be 0. The set of all linear combinations of $S$ is the span of $S$ in $V$. A subset $S$ of spans $V$ if every element of $V$ is a linear combination of elements of $S$. A subset $S$ is linearly independent if the only expression $r_1 x_1 + r_2 x_2 + \ldots + r_n x_n$ giving 0 as a linear combination of elements of $S$ is with all the coefficients $r_i = 0$. A basis of $V$ is a linearly independent set spanning $V$. The dimension of $V$ is the number of elements in any basis of $V$, if $V$ has a finite basis, or infinity if there is no finite basis (and this is a well-defined integer if there is a finite basis).
Thus every extension of $F$ has a dimension as a vector space over $F$.

**Definition 3.9.4** Suppose $F$ is a subfield of $E$ and $\alpha \in E$. Then $\alpha$ is an algebraic element of $E$ over $F$ if $\alpha$ is a root of some polynomial in $F[x]$. A nonzero polynomial of least degree having $\alpha$ as a root is called a minimal polynomial of $\alpha$. An extension $E$ of $F$ is an algebraic extension of $F$ if every element of $E$ is algebraic over $F$. If $\alpha \in E$ is algebraic over $F \subseteq E$, write $F(\alpha)$ for the smallest subfield of $E$ containing $F$ and $\alpha$, and write $F[\alpha]$ for the set of all linear combinations of powers of $\alpha$, i.e., all $p(\alpha) \in E$ for $p \in F[x]$.

**Theorem 3.9.3** Suppose $E$ is an extension of $F$. If $E$ has finite dimension as a vector space over $F$, then $E$ is an algebraic extension of $F$. If $\alpha \in E$ is algebraic over $F$ and $p(x) \in F[x]$ is a minimal polynomial for $\alpha$, then $p(x)$ is unique up to multiplication by a nonzero element of $F$, $p(x)$ is an irreducible polynomial in $F[x]$, the smallest subfield of $E$ containing $F$ and $\alpha$ is $F[\alpha]$, and $F[\alpha] \cong F[x]/(p(x))$ by an isomorphism mapping $\alpha$ to $(p(x)) + x$.

**Proof:** Suppose $E$ has dimension $n$ over $F$ and $\alpha \in E$. Then any $n + 1$ elements of $E$ will be linearly dependent. For the set $\{1, \alpha, \alpha^2, \ldots, \alpha^n\}$, there will be coefficients $r_i \in F$, $0 \leq i \leq n$, not all zero, such that $r_0 + r_1 \alpha + r_2 \alpha^2 + \ldots + r_n \alpha^n = 0$, i.e., $\alpha$ is a root of a polynomial and hence $\alpha$ is algebraic over $F$.

Suppose $\alpha \in E$ is algebraic over $F$ and $p(x)$ is a minimal polynomial for $\alpha$. If also $p_2(x)$ is minimal polynomial then $p_2(x) = q(x)p(x) + r(x)$ for $r(x)$ of lower degree than $p(x)$ and $0 = p_2(\alpha) = q(\alpha)p(\alpha) + r(\alpha) = q(\alpha)0 + r(\alpha) = r(\alpha)$ contradicting the minimality of the degree of $p(\alpha)$ unless $r(x)$ is nonzero. Hence $p(x)$ divides $p_2(x)$ and $p_2(x)$ divides $p(x)$ similarly, i.e., $p(x)$ and $p_2(x)$ differ by a unit of $F[x]$, by a factor of a nonzero element of $F$. If $p(x) = q(x)r(x)$ then $0 = p(\alpha) = q(\alpha)r(\alpha)$ so $q(alpha)$ or $r(\alpha)$ is zero and of no greater degree than $p(x)$, but then by the minimality of the degree of $p(x)$ the other factor is a unit and so $p(x)$ was irreducible.

Define a homomorphism of $F[x]$ onto $F[\alpha]$ by mapping a polynomial $a(x)$ to $a(\alpha)$ (easy to check a homomorphism). Then the kernel of this homomorphism includes $p(x)$ but if it includes a polynomial $a(x)$ not a multiple of $p(x)$ and we take $a(x) = q(x)p(x) + r(x)$ then $0 = a(\alpha) = q(\alpha)p(\alpha) + r(\alpha)$ and $r(x)$ is a polynomial of lower degree than $p(x)$ having $\alpha$ as a root contradicting the minimality of the degree of $p(x)$. Hence the kernel of this homomorphism is simply the ideal generated by $p(x)$ and $F[\alpha] \cong F[x]/(p(x))$,
a field since \(p(x)\) is irreducible. The smallest subfield of \(E\) containing \(F\) and \(\alpha\) contains \(F[\alpha]\) but this set is also a field so it equals the smallest subfield containing \(F\) and \(\alpha\). □

For example, we have considered the field \(\mathbb{Q}[\sqrt{2}]\) by viewing the elements as real numbers. But we could just as well formally take \(\sqrt{2}\) simply as a formal symbol about which the only thing we assume is that it is a root of \(x^2 - 2\). This polynomial is the minimal polynomial for \(\sqrt{2}\) over the rationals. As a polynomial over \(\mathbb{Q}[\sqrt{2}]\), \(x^2 - 2\) is not irreducible since it factors as \((x - \sqrt{2})(x + \sqrt{2})\). The field \(\mathbb{Q}[\sqrt{2}]\) is isomorphic to \(\mathbb{Q}[x]/(x^2 - 2)\), that is, we can consider polynomials in \(x\) modulo the ideal generated by \(x^2 - 2\), or taking \(x\) to be replaced by the symbol \(\sqrt{2}\) in the quotient ring, we have the \(\sqrt{2}\) is a root of \(x^2 - 2\). Now \(-\sqrt{2}\) also is a root of this same polynomial.

**Theorem 3.9.4** Suppose \(E\) and \(E'\) are extensions of a field \(F\), and suppose \(\alpha \in E\) and \(\beta \in E'\) have the same minimal polynomial, and take \(F[\alpha] \subseteq E\) and \(F[\beta] \subseteq E'\). Then \(F[\alpha] \cong F[\beta]\) by an isomorphism which is the identity on \(F\) and maps \(\alpha\) to \(\beta\).

**Proof:** Take \(p(x)\) the minimal polynomial of \(\alpha\) and \(\beta\). Then \(F[\alpha] \cong F[x]/(p(x)) \cong F[\beta]\). The first isomorphism corresponds \(\alpha\) and \((p(x)) + x\) and the second corresponds this with \(\beta\). □

**Definition 3.9.5** A splitting field for a nonzero polynomial \(p(x) \in F[x]\) of degree at least 1 is an extension \(E\) of \(F\) such that \(p(x)\) factors into linear factors in \(E[x]\) and \(p(x)\) does not factor into linear factors in any proper subfield of \(E\) (i.e., \(E\) is generated by the roots of \(p(x)\) and \(F\)).

**Theorem 3.9.5** For any nonzero polynomial \(p(x) \in F[x]\) of degree at least 1, there exists a splitting field for \(p(x)\). If \(E\) and \(E'\) are splitting fields for \(p(x)\) then \(E \cong E'\).

**Proof:** Proceed by induction on the degree of \(p(x)\). If \(p(x)\) is degree 1, then it is already factored into a linear factors in \(F[x]\), the root of \(p(x)\) is an element of \(F\), and any splitting field of \(p(x)\) is just \(F\).

Suppose \(p(x)\) has degree greater than 1. If \(p(x)\) is not irreducible, factor \(p(x)\) into irreducibles and let \(p_1(x)\) be an irreducible factor, and otherwise simply take \(p_1(x) = p(x)\). Then \(E_1 = F[x]/(p_1(x))\) is an extension of \(F\) in which \(p_1(x)\) has a root \(\alpha\). Factor \(p(x)\) over \(E_1\), say \(p(x) = q(x)(x - \alpha)\). Then \(q(x) \in E_1[x]\) is of lower degree than \(p(x)\) so by induction hypothesis there is
a splitting field $E_2$ for $q(x)$ over $E_1$. But then $p(x)$ factors into linear factors over $E_2$, and the field generated by $F$ and the roots of $p(x)$ in $E_2$ is a splitting field of $p(x)$.

Now if $E$ and $E'$ are splitting fields of $p(x)$, then $p_1(x)$ has roots $\alpha \in E$ and $\beta \in E'$, and $F[\alpha] \subseteq E$ and $F[\beta] \subseteq E'$ are isomorphic subfields of $E$ and $E'$ say by an isomorphism $\phi$ which is the identity on $F$ and takes $\alpha$ to $\beta$. Let $p(x) = q(x)(x - \alpha)$ in $F[\alpha][x]$ and $p(x) = r(x)(x - \beta)$ in $F[\beta][x]$ and the isomorphism of $\phi : F[\alpha] \to F[\beta]$ extends to an isomorphism $F[\alpha][x] \to F[\beta][x]$ taking $q(x)$ to $r(x)$. Now $E$ is a splitting field of $q(x)$ and $E'$ is a splitting field of $r(x)$, otherwise these would not have been splitting fields for $p(x)$. Then $E$ and $E'$ have isomorphic subfields, we might as well have taken them to be the same field, and since $E$ and $E'$ are both splitting fields of corresponding polynomials in the two subfield, or the same polynomial if we identify these subfields, and since $q(x)$ is of degree less than $p(x)$, $E \cong E'$ by induction hypothesis. □

### 3.10 Finite fields

We apply all of the tools developed in this course to understanding finite sets with addition, negation, additive identity 0, multiplication, multiplicative identity 1, and multiplicative inverses of nonzero elements, that satisfy the usual rules of arithmetic (that is, the field axioms). Finite fields are elegant combinatorial, number theoretical, algebraic objects.

**Theorem 3.10.1** The number of elements of a finite field is $p^n$ for $p$ the characteristic of the field and some positive integer $n$.

**Proof:** Suppose $F$ is a finite field of characteristic $p$, a prime since if $F$ is characteristic 0, then its prime subfield is the rational numbers which is infinite. Then $F$ is a vector space over $F$, of finite dimension $n$ since $F$ is finite, and then $F$ has $p^n$ elements, since in terms of linear combinations of $n$ basis elements, there are $p$ possible coefficients in $\mathbb{Z}_p$ for each basis element giving $p^n$ combinations, all distinct elements of $F$. □

For example, take the integers mod 5 and consider the polynomial $p(x) = x^2 - 2 \in \mathbb{Z}_5[x]$. Then $p(x)$ is irreducible so the ideal $(p(x))$ is a maximal ideal and $F = \mathbb{Z}_5[x]/(p(x))$ is a field. In this field, $(p(x)) + x$ is an element whose square is 2, and dropping the explicit reference to the ideal $(p(x))$ we might view $x \in F$ the same as $\mathbb{Z}_5 \subseteq F$ with the relation $x^2 = 2$ in $F$ understood.
If we think of this $x$ as $\sqrt{2}$ we could even write $F = \mathbb{Z}_5[\sqrt{2}]$. Every element of $F$ is a formal $\mathbb{Z}_5$ linear combination of 1 and $\sqrt{2}$, so there are 25 elements in $F$ with operations defined by

$$(a_1 + b_1\sqrt{2}) + (a_2 + b_2\sqrt{2}) = (a_1 + a_2) + (b_1 + b_2)\sqrt{2}$$

and

$$(a_1 + b_1\sqrt{2}) + (a_2 + b_2\sqrt{2}) = (a_1a_2 + 2b_1b_2) + (a_1b_2 + a_2b_1)\sqrt{2}$$

the coefficients taken mod 5. Now it would seem that there are a few different similar ways we could try to construct a field of 25 elements. However all fields of 25 elements are isomorphic.

**Theorem 3.10.2** For $p$ a prime and $n$ any positive integer, there exists, up to isomorphism, a unique field having $p^n$ elements. This field is denoted $\text{GF}(p^n)$ and is called the Galois field of $p^n$ elements.

**Proof:** First we construct a field of $p^n$ elements. Starting with $\mathbb{Z}_p$ we should take a dimension $n$ extension field. The trick that works is to take $f(x) = x^{p^n} - x$ and take $F$ to be the splitting field of $f(x)$. Then $f'(x) = p^n x^{p^n-1} - 1 = -1$, in $\mathbb{Z}_p$, so $f'(x)$ and $f(x)$ have no common factors so $f(x)$ has no repeated roots in $F$. Thus $f(x)$ factors over $F$ into $p^n$ distinct linear factors and $f(x)$ has exactly $p^n$ roots in $F$. Now the set of roots of $f(x)$ is closed under addition, negation, multiplication, and multiplicative inverses of nonzero elements (an interesting exercise to check). Thus the set of roots of $f(x)$ is already a field, that is the splitting field of $f(x)$ consists of just the roots of $f(x)$, that is, $F$ has $p^n$ elements.

Suppose we have a field $F'$ of $p^n$ elements. Then the characteristic of $F'$ is $p$, $F'$ has prime subfield $\mathbb{Z}_p$. The group of units of $F'$, the set of all nonzero elements of $F'$ under multiplication, is an Abelian group of $p^n - 1$ elements. By Lagrange’s theorem, every nonzero $x \in F'$ satisfies $x^{p^n-1} = 1$. But then $f(x) = x(x^{p^n-1} - 1) = 0$ for every $x \in F'$. That is $f(x)$ factors completely into $p^n$ factors over $F'$, $f(x)$ has as $p^n$ distinct roots the elements of $F'$, and $F'$ is the splitting field of $f(x)$. But then $F' \cong F$ since splitting fields are unique up to isomorphism. □

Addition in a finite field $F$ can be understood by taking a basis for the vector space $F$ over its prime subfield and taking addition componentwise. Multiplication is a little trickier to see, and is not easily connected to the vector space picture of the field, nevertheless multiplication in a finite field turns out to be quite simple.
Theorem 3.10.3  The set of nonzero elements of GF($p^n$) under multiplication is a cyclic group of order $p^n - 1$.

Proof: The nonzero elements of $F = GF(p^n)$ form an Abelian group under multiplication. If no element of this group has order $p^n - 1$ then for some $k < p^n - 1$ we would have every element $x$ satisfies $x^k = 1$ (by considering the structure of finite Abelian groups essentially). But then $g(x) = x(x^k - 1)$ is a polynomial having at most $k + 1$ roots but satisfied by every element of $F$ so $k + 1$ must be at least $p^n$ a contradiction. Hence instead some nonzero element of $F$ has order $p^n - 1$ and so generates the cyclic group of nonzero elements in $F$. □

Thus, for example, $u = 2 + \sqrt{2}$ is a generator of the group of units of $\mathbb{Z}[\sqrt{2}]$ and we have

$$
\begin{align*}
0 + 0\sqrt{2} &= 0 & 1 + 0\sqrt{2} &= u^0 & 2 + 0\sqrt{2} &= u^6 & 3 + 0\sqrt{2} &= u^{18} & 4 + 0\sqrt{2} &= u^{12} \\
0 + 1\sqrt{2} &= u^{15} & 1 + 1\sqrt{2} &= u^{10} & 2 + 1\sqrt{2} &= u^1 & 3 + 1\sqrt{2} &= u^{17} & 4 + 1\sqrt{2} &= u^{14} \\
0 + 2\sqrt{2} &= u^{21} & 1 + 2\sqrt{2} &= u^{23} & 2 + 2\sqrt{2} &= u^3 & 3 + 2\sqrt{2} &= u^{16} & 4 + 2\sqrt{2} &= u^7 \\
0 + 3\sqrt{2} &= u^9 & 1 + 3\sqrt{2} &= u^{19} & 2 + 3\sqrt{2} &= u^8 & 3 + 3\sqrt{2} &= u^4 & 4 + 3\sqrt{2} &= u^{11} \\
0 + 4\sqrt{2} &= u^3 & 1 + 4\sqrt{2} &= u^2 & 2 + 4\sqrt{2} &= u^5 & 3 + 4\sqrt{2} &= u^{13} & 4 + 4\sqrt{2} &= u^{22}
\end{align*}
$$

and

$$
\begin{align*}
u^0 &= 1 + 0\sqrt{2} & u^1 &= 2 + 1\sqrt{2} & u^2 &= 1 + 4\sqrt{2} & u^3 &= 0 + 4\sqrt{2} \\
u^4 &= 3 + 3\sqrt{2} & u^5 &= 2 + 4\sqrt{2} & u^6 &= 2 + 0\sqrt{2} & u^7 &= 4 + 2\sqrt{2} \\
u^8 &= 2 + 3\sqrt{2} & u^9 &= 0 + 3\sqrt{2} & u^{10} &= 1 + 1\sqrt{2} & u^{11} &= 4 + 3\sqrt{2} \\
u^{12} &= 4 + 0\sqrt{2} & u^{13} &= 3 + 4\sqrt{2} & u^{14} &= 4 + 1\sqrt{2} & u^{15} &= 0 + 1\sqrt{2} \\
u^{16} &= 2 + 2\sqrt{2} & u^{17} &= 3 + 1\sqrt{2} & u^{18} &= 3 + 0\sqrt{2} & u^{19} &= 1 + 3\sqrt{2} \\
u^{20} &= 3 + 2\sqrt{2} & u^{21} &= 0 + 2\sqrt{2} & u^{22} &= 4 + 4\sqrt{2} & u^{23} &= 1 + 2\sqrt{2}
\end{align*}
$$

To add elements we prefer the form $a + b\sqrt{2}$, adding coefficients of basis elements 1 and $\sqrt{2}$ mod 5, while to multiply elements we either need to remember $\sqrt{2}^2 = 2$ or work with elements as powers of $u$, simply adding exponents mod 24.

Note that $\mathbb{Z}_5[\sqrt{2}] \cong GF(25)$ is the splitting field of $f(x) = x^{25} - x$ over $\mathbb{Z}_5$, but $f$ factors over $\mathbb{Z}_5$ as

$$
f(x) = x^{25} - x = x(x - 1)(x - 2)(x - 3)(x - 4)(x^2 - 2)(x^2 - 3) \\
(x^2 + x + 1)(x^2 + x + 2)(x^2 + 2x + 3)(x^2 + 2x + 4) \\
(x^2 + 3x + 3)(x^2 + 3x + 4)(x^2 + 4x + 1)(x^2 + 4x + 2)
$$

73
where each of the quadratic factors are irreducible. Extending by a root of any one of these quadratic factors gives an isomorphic dimension 2 extension of \( \mathbb{Z}_5 \).

### 3.11 An application of finite fields

As interesting and useful as finite fields are mathematically, the arithmetic of finite fields has nonmathematical applications. One such application is in the area of error correcting codes.

Suppose you have a message to be transmitted or recorded. Any medium of transmission or recording the message will be subject to noise, random fluctuations that might cause parts of the message to be corrupted. If you want to be sure the message can be accurately read, you would want to expand the message with additional information that might allow you to recover corrupted parts of the message.

Thus for example, if your satellite is sending important data to a ground station over a great distance, it might not be practical to resend data that was corrupted in transmission. Instead it would be worthwhile to send the data encoded in a form that allowed such errors in transmission to be corrected and the original data recovered. Similarly, if you record data on a compact disc, a scratch across the disc might make part of the disc unreadable. By encoding the data with enough error correcting information to fill in such gaps, you can increase the likelihood that even a damaged disc will still contain enough information to recover the data.

Suppose messages and encoded messages are written in letters from a set \( X \), a message of length \( n \) is a sequence from \( X^n \). The set \( X \) might consist of 0 and 1 in typical on-off binary data, or thought of as groups of 8 or more such bits, or perhaps a set of amplitude-phase combinations on some transmission channel. And encoding is an injective map of \( X^n \) into \( X^m \) for some \( m \geq n \), of messages into the encoded form for transmission. We need the map to be injective in order that a received encoded message can be decoded and the original message recovered. If transmission were perfect then we could simply take \( m = n \) and encode by the identity function. An error in transmission might however change one or more of the letters of the encoded message. If changing some of the letters of the encoded message results in the encoded form of another message, then we would assume that the message sent was that other message and might never know that the message had been corrupted. Thus it is advantageous if any two elements
the image of the encoding map differ in at least some minimum number of letters, making it unlikely that one valid encoded message is corrupted to give another valid encoded message. When an invalid encoded message is received, it may be that there is a unique valid encoded message differing from the received message in the fewest number of places, and we might then reasonably say that this valid message is the one most likely to have been transmitted (fewest number of letters that were misread because of noise), and so decode the invalid message as this nearest valid message.

For example, we might encode a message \((a_1, a_2, \ldots, a_n)\) by transmitting this message three times, the encoded transmission of three times the length being \((a_1, a_2, \ldots, a_n, a_1, a_2, \ldots, a_n, a_1, a_2, \ldots, a_n)\). We can check whether the received message has this form, and if not, guess that if two out of three of the letters in corresponding positions agree, there was an error in the third position, and interpret the original message in that position to be the letter in the other two corresponding positions. Thus any single error can be corrected, but two errors, if they happen to occur in corresponding positions, might lead to a message with three different letters in corresponding positions, or perhaps two of three of these letters equal to the wrong letter in that position.

Now any encoding scheme will fail to decode correct messages if enough letters of the encoded messages are misread due to noise (unless there is only one possible message). The “three times” encoding can correct any single error, but may fail for two errors. This encoding is also less than efficient in that the encoded message is three times the length of the message. A better scheme would use fewer extra letters and distribute redundant information about the original message throughout the encoded message so that errors in a number of letters would still leave enough information to recover the original message. Reed-Solomon codes are such error correcting schemes.

Now here is were the arithmetic properties of finite fields come into play. Suppose we take \(X\) to be a finite field \(F\) of \(N\) elements. Let \(u\) be a generator for the group of units of \(F\), so we can list the elements of \(F\) as \(0, 1, u, u^2, \ldots, u^{N-1}\). For any \(s < N/2\), an error correcting code which can correct any \(s\) errors is defined by taking a message \((a_1, a_2, \ldots, a_{N-2s})\) of length \(N - 2s\) encoded as the message

\[
(b_1, b_2, \ldots, b_N) = (f_a(0), f_a(1), f_a(u), f_a(u^2), \ldots, f_a(u^{N-1}))
\]
of length $N$ where we take

$$f_a(x) = a_1 + a_2 x + a_3 x^2 + \ldots + a_{N-2s} x^{N-2s-1} \in F[x]$$

The key to decoding messages that have as many as $s$ letters changed is the following theorem.

**Theorem 3.11.1** Suppose $F$ is a field with at least $k$ elements. Suppose also that $x_1, x_2, \ldots, x_k \in F$ are $k$ distinct elements of $F$ and let $b_1, b_2, \ldots, b_k \in F$ be any $k$ values. Then there is a unique polynomial $f(x) \in F[x]$ of degree at most $k-1$, such that $f(x_i) = b_i$ for each $i$.

**Proof:** The polynomial we want is the interpolating polynomial with $f(x_i) = b_i$ for each $i$. Let $f(x) = a_1 + a_2 x + a_3 x^2 + \ldots + a_k x^{k-1}$. Then the conditions $f(x_i) = b_i$ for each $i$, constitute a system of $k$ linear equations in the $k$ unknowns $a_1, a_2, \ldots, a_k$ which can always be solved for the coefficient $a_i$. Alternatively, for each $i$ let

$$\ell_i(x) = \frac{(x-x_1)(x-x_2)\ldots(x-x_i)\ldots(x-x_k)}{(x_i-x_1)(x_i-x_2)\ldots(x_i-x_i)\ldots(x_i-x_k)}$$

where we write a hat over the factors that are to be omitted from the numerator and denominator of this fraction. Since the $x_i$ are distinct the denominator of the fraction is nonzero and so has a multiplicative inverse in $F$. Thus $\ell_i(x)$ is a degree $k-1$ polynomial and we can easily check that $\ell_i(x_i) = 1$ but $\ell_i(x_j) = 0$ for $j \neq i$. Now let

$$f(x) = b_1 \ell_1(x) + b_2 \ell_2(x) + \ldots + b_k \ell_k(x)$$

then $f(x_i) = b_i$ so $f(x)$ is an appropriate interpolating polynomial written in terms of Lagrange polynomials $\ell_i(x)$ for the set of points $x_i$.

Now if $f_1(x)$ and $f_2(x)$ are two polynomials with $f_1(x_i) = b_i = f_2(x_i)$ for each $i$, then $f_1(x) - f_2(x) \in F[x]$ is a polynomial of degree $k-1$ with $k$ distinct roots $x_i$ for each $i$. But a nonzero polynomial can have at most as many roots as its degree, hence $f_1(x) - f_2(x)$ is the zero polynomial and $f_1(x) = f_2(x)$. □

Now suppose a message $a$ of length $N - 2s$ is encoded as the sequence of $N$ values of the polynomial $f_a(x)$, a polynomial of degree at most $N - 2s - 1$. Suppose at most $s$ letters of the encoded message are received incorrectly.
Then there is a subset of $N - s$ letters of the received message which are all the correct values of the polynomial $f_a(x)$ at the corresponding elements of $F$, this polynomial can be determined uniquely by interpolating at any $N - 2s$ of the points and checked by evaluating at the remaining $s$ of the $N - s$ points. On the other hand, if we take any $N - s$ letters of the received message, then at least $N - 2s$ of these letters are the correct values of $f_a(x)$, so if there is any polynomial $f(x)$ of degree at most $N - 2s - 1$ agreeing with all $N - s$ values in this subset, then it agrees with $f_a(x)$ on a subset of $N - 2s$ values. But there is a unique polynomial of degree at most $N - 2s - 1$ with these $N - 2s$ values so $f(x) = f_a(x)$. That is, $f_a(x)$ is the only polynomial of degree at most $N - 2s - 1$ which agrees with the received message at as many as $N - s$ points.

For example, for a single error correcting code take $F = \mathbb{Z}_5$, $N = 5$, and $s = 1$. We can divide a long message over this five letter alphabet into blocks of three. A block $(a_1, a_2, a_3) \in \mathbb{Z}_5^3$ is encoded by defining a quadratic polynomial $f_a(x) = a_1 + a_2x + a_3x^2$ and taking the encoded message to be $(f_a(0), f_a(1), f_a(2), f_a(3), f_a(4))$. Even if any single letter of this message is changed, a polynomial $f(x)$ of degree 2 which agrees with four of these values agrees with $f_a(x)$ on at least 3 correct values in this message, but there is a unique polynomial of degree 2 having 3 specified values. As an example, to encode $(2, 0, 1)$ we evaluate $f_a(x) = 2 + x^2$ to get values $(2, 3, 1, 1, 3)$. Suppose the last letter is misread as 2 and the received message is $(2, 3, 1, 1, 2)$. Then we note that the first four values would be explained as values of the quadratic polynomial $f_a(x)$. On the other hand, the last three values agree with

$$f(x) = \frac{(x - 3)(x - 4)}{(2 - 3)(2 - 4)} + \frac{(x - 2)(x - 4)}{(3 - 2)(3 - 4)} + \frac{2(x - 2)(x - 3)}{(4 - 2)(4 - 3)} = 4 + 3x^2$$

but then $f(1) = 2$ and not 3 as received. No other subset of four of the five values can be fit by a quadratic polynomial either. We conclude that if there was at most one error, then the encoded message was $(2, 3, 1, 1, 3)$, values of the polynomial $f_a(x) = 2 + x^2$, and the original message was $(2, 0, 1)$.

The encoding of messages of length $3n$ as messages of length $5n$ with single error correction in any block of 5 is a big improvement over the “three times” encoding that triples the length of the message. Much better would be to take a large finite field and encode large blocks of the message so that a multiple number of errors anywhere in the block can be corrected, including errors in several successive letters of the encoded message. For example, taking $N = 2^8 = 256$ and $s = 28$ means that messages of 200 bytes can be
encoded in 256 bytes so that any 28 errors can be corrected (i.e., more than a 10% error rate in transmission could be corrected).

One of the deficiencies of this type of encoding is that we need to know reliably where each received letter is supposed to lie in the sequence of values of $f_a(x)$. An error that caused a letter to be deleted from the sequence and not just changed could not be corrected (in the same way at least). More and more clever schemes of encoding and error correction have allowed more and more nearly the maximum theoretical amount of information to be transmitted along noisy channels. The mathematical support for devising such schemes is combinatorial and algebraic.
Final Exam

Due April 27, 5:00PM

Return your exam solutions to me in my office, SC1507, or place them in my box in the math office. In contrast to homework problems, do not discuss the problems with your classmates or others. You may use your class notes or book. Calculators are allowed but answers should be exact and simplified. Clarity of exposition in proofs is important; you should, once you’ve figured out an answer, have time to write it out clearly. I don’t intend to make these problems so hard it will take days to find solutions. I sometimes make mistakes on problems, so don’t waste a lot of time trying to solve problems that are too difficult. It will not be necessary to have solved all of these problems to get an A in the course. If you can’t solve a problem, explain what ideas you have, what definitions you need to use, an example you examined, or whatever else I might be able to use to assign part credit for what you do know about the problem and the progress you made. If you have questions you can stop by my office, or call or email me. Generally, I expect to be in my office each day this week from 10 to 3 except for lunch time.

Problem 1. Suppose $G$ is a cyclic group of order $n$ and $k$ is a positive integer. Show that there exists a subgroup $H$ of $G$ of order $k$ iff $k$ divides $n$. If $k$ divides $n$, show that $G$ has a unique subgroup $H$ of order $k$.

Problem 2. Suppose $G$ is a group. For $g \in G$ define a map $\phi_g : G \to G$ by $\phi_g(x) = gxg^{-1}$. Show that, for any $g \in G$, $\phi_g$ is group homomorphism and determine the kernel of $\phi_g$. Show that for any $g,h \in G$, $\phi_g \circ \phi_h = \phi_{gh}$. Conclude that the map $\psi : G \to \text{Sym}(G)$ defined by $\psi(g) = \phi_g$ is a group homomorphism and therefore the set of elements $g \in G$ such that $\phi_g$ is the identity map on $G$ is a normal subgroup of $G$.

Problem 3. Suppose $N$ is a normal subgroup of a group $G$, and $n$ is some positive integer. Prove that if, for every $x \in G/N$, $x^n = 1$ in $G/N$, then for every $g \in G$, $g^n \in N$.

Problem 4. Suppose $S$ and $T$ are subrings of a ring $R$. Prove that $S \cap T$ is a subring of $R$. Prove that if $S$ and $T$ are ideals of $R$ then $S \cap T$ is an ideal of $R$. 
Problem 5. Prove that if $R$ and $S$ are isomorphic rings, then $R[x]$ and $S[x]$ are isomorphic rings.

Problem 6. Suppose $F$ is a field and $p(x) \in F[x]$ is irreducible. Let $I = (p(x))$ be the ideal of $F[x]$ generated by $p(x)$ and let $E = F[x]/I$. Applying the Euclidean algorithm to $F[x]$, show that if $q(x)$ is any nonzero polynomial of degree less than the degree of $p(x)$, then there exist polynomials $s(x)$ and $t(x)$ in $F[x]$ such that, $1 = s(x)p(x) + t(x)q(x)$. Conclude then that $I + t(x)$ is the multiplicative inverse of $I + q(x)$ in $E$.

Problem 7. Suppose $F$ is a field of characteristic $p \neq 0$. Define a map $\phi : F \rightarrow F$ by $\phi(x) = x^p$. Show that $\phi$ is a ring homomorphism. Determine the set of elements of $F$ fixed by $\phi$, i.e., the set \{ $x \in F : \phi(x) = x$ \}.