Ring Theory (part 2): Rings (by Evan Dummit, 2017, v. 1.00)

Contents

2 Rings

Our goal in this chapter is to define the algebraic structure known as a ring, which is a set with two binary operations (addition and multiplication) satisfying certain axioms. We then study a variety of examples of rings and explore their basic properties. Much of our discussion parallels our treatment of $\mathbb{Z}$, the most fundamental example of a ring. We will then discuss at length an important class of rings known as polynomial rings, which generalize the idea of a "polynomial with real coefficients" familiar from elementary algebra. Our discussion of polynomials will again parallel our treatment of $\mathbb{Z}$: we will examine the Euclidean algorithm, unique factorization, and modular arithmetic in the polynomial setting.

2.1 Rings

- In the previous chapter, we studied two sets: $\mathbb{Z}$ (the integers) and $\mathbb{Z}/m\mathbb{Z}$ (the integers modulo m), and found that they had a number of algebraic properties in common. Our present goal is to "abstract" those properties of $\mathbb{Z}$ and $\mathbb{Z}/m\mathbb{Z}$ that give them much of their underlying structure, in order to study more general systems that behave like $\mathbb{Z}$ and $\mathbb{Z}/m\mathbb{Z}$.

2.1.1 The Formal Definition of a Ring

- Definition: A ring is any set R having two (closed) binary operations $+$ and $\cdot$ that satisfy the six axioms [R1]-[R6]:

  [R1] The operation $+$ is associative: $a + (b + c) = (a + b) + c$ for any elements a, b, c in R.

  [R2] The operation $+$ is commutative: $a + b = b + a$ for any elements a, b in R.

  [R3] There is an additive identity 0 satisfying $a + 0 = a$ for all a in R.

[R4] Every element a has an additive inverse $-a$ satisfying $a + (-a) = 0$.

[R5] The operation $\cdot$ is associative: $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for any elements a, b, c in R.

[R6] The operation $\cdot$ distributes over $+$: $a \cdot (b + c) = a \cdot b + a \cdot c$ and $(a + b) \cdot c = a \cdot c + b \cdot c$ for any elements a, b, c in R.

- Certain rings will also possess additional properties. These properties arise often enough that we give them names:
- Definition: If a ring satisfies axiom [R7], we say it is a commutative ring.

[R7] The operation $\cdot$ is commutative: $a \cdot b = b \cdot a$ for any elements a, b in R.

- Definition: If a ring satisfies axiom [R8], we say it is a ring with identity (or a "ring with 1").

[R8] There is a multiplicative identity $1 \neq 0$, satisfying $1 \cdot a = a = a \cdot 1$ for all a in R.

- Definition: If a ring with identity further satisfies the axiom [D], it is called a division ring. A commutative division ring is called a field.

[D] Every nonzero a in R has a multiplicative inverse a^{-1} satisfying $a \cdot a^{-1} = 1 = a^{-1} \cdot a$.

2.1.2 Examples of Rings

- Here is a list of examples (and non-examples) of rings[1]:
- Example: The integers $\mathbb{Z}$ are a commutative ring with identity.
- Example: The set of even integers is a commutative ring that does not have an identity.
  - The properties [R1]-[R7] all follow from their counterparts in $\mathbb{Z}$: [R3] follows because 0 is an even integer, and [R4] follows because n is an even integer if and only if $-n$ is an even integer.
  - This ring does not have a multiplicative identity because there is no solution to $2n = 2$ inside the set of even integers.
- Non-Example: The set of odd integers is not a ring.
  - The problem is that, although multiplication of two odd integers does return an odd integer, the sum of two odd integers is not odd: thus, the operation $+$ is not defined on the set of odd integers.
- Example: The set $\mathbb{Z}/m\mathbb{Z}$ of residue classes modulo m form a commutative ring with identity.
  - Furthermore, if p is a prime, we know that all of the nonzero residue classes modulo p are invertible, meaning that $\mathbb{Z}/p\mathbb{Z}$ is a field.
  - Indeed, the only residue classes that are invertible modulo m are those relatively prime to m, so if m is not prime, then $\mathbb{Z}/m\mathbb{Z}$ is not a field.
- Example: The rational numbers $\mathbb{Q}$, the real numbers $\mathbb{R}$, and the complex numbers $\mathbb{C}$ are all examples of fields.
  - Recall that $\mathbb{C}$ is the set of numbers of the form $a + bi$, where a and b are real numbers and $i^2 = -1$.
  - Addition and multiplication in $\mathbb{C}$ are as follows: $(a+bi)+(c+di) = (a+c)+(b+d)i$, and $(a+bi)\cdot(c+di) = ac + adi + bci + bdi^2 = (ac - bd) + (ad + bc)i$.
- Example: The set of complex numbers of the form $a + bi$ where $a, b \in \mathbb{Z}$ are a commutative ring with identity.
  - This ring is denoted $\mathbb{Z}[i]$ (read as: "$\mathbb{Z}$ adjoin i") and is also often called the Gaussian integers.

[1] For brevity, when we do not specify the operations $+$ and $\cdot$, they are always assumed to be the standard addition and multiplication operations on the corresponding sets.

- The properties [R1]-[R8] all follow from their counterparts in $\mathbb{C}$: [R3] follows because $0 = 0 + 0i$, and [R4] follows because we have $-(a+bi) = (-a) + (-b)i$.

- Example: The set of real numbers of the form $a + b\sqrt{2}$ where $a, b \in \mathbb{Z}$ are a commutative ring with identity.
  - This ring is denoted $\mathbb{Z}[\sqrt{2}]$. The addition and multiplication are defined in a similar way as for the complex numbers and Gaussian integers: $(a+b\sqrt{2}) + (c+d\sqrt{2}) = (a+c) + (b+d)\sqrt{2}$, and $(a+b\sqrt{2}) \cdot (c+d\sqrt{2}) = ac + ad\sqrt{2} + bc\sqrt{2} + 2bd = (ac+2bd) + (ad+bc)\sqrt{2}$.

- We can also see the structure of a ring in collections of functions:

- Example: If S is any set and A is any ring, the collection R of functions $f : S \to A$, with operations $(f+g)(x) = f(x) + g(x)$ and $(f \cdot g)(x) = f(x)g(x)$, forms a ring.
  - Thus, for example, if A is the set of real numbers, with $f(x) = x^2$ and $g(x) = 3x^2$, then $(f+g)(x) = 4x^2$ and $(fg)(x) = 3x^4$.
  - Ultimately, each of the properties [R1]-[R6] follows from the corresponding property of A. The additive identity is the "identically-zero function" 0_S that is 0 on each element of S, and the additive inverse $-f$ of f is defined as $(-f)(x) = -f(x)$ for each x in S.
  - If A is commutative, then it is easy to see that R will also be commutative. Likewise, if A has a 1, then the "identically-1 function" 1_S that is 1 on each element of S, is a multiplicative identity in R.

- Example: The collection R of continuous real-valued functions $f : \mathbb{R} \to \mathbb{R}$, with operations $(f+g)(x) = f(x) + g(x)$ and $(f \cdot g)(x) = f(x)g(x)$, forms a commutative ring.
  - The operations are well-defined because the sum and product of two continuous functions is continuous.
  - The remaining properties [R1]-[R6] follows from the same observations as in the example above.

- So far, all of the rings we have listed are commutative. Here are a few that are not:

- Example: The set of 2×2 matrices $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ with real number entries, denoted $M_{2\times 2}(\mathbb{R})$, forms a noncommutative ring with identity.
  - Explicitly, the operations in this ring are $\begin{bmatrix} a & b \\ c & d \end{bmatrix} + \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} a+e & b+f \\ c+g & d+h \end{bmatrix}$ and $\begin{bmatrix} a & b \\ c & d \end{bmatrix} \cdot \begin{bmatrix} e & f \\ g & h \end{bmatrix} = \begin{bmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{bmatrix}$.
  - It is a straightforward algebraic computation to verify axioms [R1]-[R6]: the additive identity is the zero matrix $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, and the additive inverse of $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ is of course $\begin{bmatrix} -a & -b \\ -c & -d \end{bmatrix}$.
  - The multiplicative identity is the famous "identity matrix" $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$.
  - However, the 2×2 matrices are not a commutative ring, since (for example) we have $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ while $\begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$.
  - Remark: More generally, for any integer $n \geq 2$, the set of $n \times n$ matrices with entries from any field F, denoted $M_{n\times n}(F)$, forms a noncommutative ring with identity.

- Example: The set $\mathbb{H}$ of real quaternions $a + bi + cj + dk$, for real numbers a, b, c, d and "imaginary units" i, j, k satisfying the relations $i^2 = j^2 = k^2 = -1$, $ij = -ji = k$, $jk = -kj = i$, and $ki = -ik = j$, form a noncommutative ring with identity.
  - This ring was first characterized by William Rowan Hamilton in 1843 (whence the name $\mathbb{H}$), and is, historically speaking, one of the first examples of a noncommutative ring.

- The addition and multiplication operations are defined similarly to those in the complex numbers: addition works "componentwise", so that $(a+bi+cj+dk)+(a'+b'i+c'j+d'k)=(a+a')+(b+b')i+(c+c')j+(d+d')k$.
- Multiplication is defined using the distributive law and the relations listed above, taking care to keep the terms in the proper order when multiplying. (The real number coefficients commute with the "imaginary units" i, j, and k.)
- Thus, for example, we have

$$\begin{aligned}(1+i-k)\cdot(2+3i+j) &= (1+i-k)\cdot 2+(1+i-k)\cdot 3i+(1+i-k)\cdot j \\ &= (2+2i-2k)+(3i-3-3j)+(j+k+i) \\ &= -1+6i-2j-k.\end{aligned}$$

- In fact, the real quaternions are a division ring: one may verify that $(a+bi+cj+dk)(a-bi-cj-dk)=a^2+b^2+c^2+d^2$, and so the nonzero quaternion $a+bi+cj+dk$ has a multiplicative inverse $\dfrac{a-bi-cj-dk}{a^2+b^2+c^2+d^2}$.
- The quaternions originally arose in the study of 3-dimensional geometry. As a hint of this connection, we will note that under the standard notation for the coordinate vectors in 3-space, namely with $\mathbf{i}=\langle 1,0,0\rangle$, $\mathbf{j}=\langle 0,1,0\rangle$, and $\mathbf{k}=\langle 0,0,1\rangle$, then $\mathbf{i}\times\mathbf{j}=\mathbf{k}=-\mathbf{j}\times\mathbf{i}$, and similarly for the other possible cross products.

- Example: If V is a vector space of dimension larger than 1, the set $\mathcal{L}(V,V)$ of linear transformations from V to V is a noncommutative ring with 1 under the operations of function addition and function composition: $(S+T)(\mathbf{v})=S\mathbf{v}+T\mathbf{v}$ and $(ST)\mathbf{v}=S(T\mathbf{v})$.
  - This ring is not commutative because $ST\neq TS$ in general (since linear transformations generally do not commute with one another, in the same way that matrices do not). The multiplicative identity is the "identity transformation" with $I(\mathbf{v})=\mathbf{v}$ for every $\mathbf{v}$ in V.
- As a final observation, we remark that if we have a set with an addition operation, we can make it into a ring in a trivial way. Two examples are as follows:
- Example: If S is $\mathbb{Z}$, $\mathbb{Q}$, or $\mathbb{R}$, with $+$ taken to be normal addition, and $\cdot$ defined so that $a\cdot b=0$ for every a and b, then S is a commutative ring.
  - All of the multiplicative axioms immediately reduce to the true statement $0=0$. Of course, this ring has no multiplicative identity.
  - Remark (for those who like group theory): More generally, if G is any abelian group with operation $+$, then we may make G into a ring with "trivial multiplication" by setting $a\cdot b=0$ for every $a,b\in G$.
- Example: The set $R=\{0\}$, with operations $0+0=0$ and $0\cdot 0=0$, is a commutative ring.
  - All of the axioms follow trivially. In fact, this ring even has a multiplicative identity!
  - This ring is known as the trivial ring, and is the only ring where $1=0$.

2.1.3 Basic Properties of Rings, Zero Divisors, Units, Integral Domains

- Our immediate goal in discussing rings is to study properties of arithmetic in $\mathbb{Z}$ and $\mathbb{Z}/m\mathbb{Z}$ that generalize to arbitrary rings. To this end, we begin by establishing a number of basic properties of ring arithmetic.
  - As in $\mathbb{Z}$, we define the binary operation of subtraction by setting $a-b=a+(-b)$. We also often use implicit multiplication, and drop the $\cdot$ notation.
  - We can define scaling of a ring element a by a positive integer as repeated addition: $na=\underbrace{a+a+a+\cdots+a}_{n\text{ terms}}$.

    By associativity of addition, this notation is well-defined. In a ring with 1, this notation coincides with the product of ring elements $n\cdot a$, but (as we would desire) it is true that $n\cdot a=na$.
  - We can also define exponentiation of a ring element a as $a^k=\underbrace{a\cdot a\cdot a\cdot\cdots\cdot a}_{k\text{ terms}}$, for any positive integer k.

    By associativity of multiplication, this notation is well-defined.

- Proposition (Basic Arithmetic): Let R be an arbitrary ring. The following properties hold in R:

  1. The additive identity 0 is unique, as is the multiplicative identity 1 (if R has a 1).
     - Proof: Suppose that 0_a and 0_b were both additive identities. Then by [R2], [R3], and the hypotheses, $0_a = 0_a + 0_b = 0_b + 0_a = 0_b$. A similar argument with [R8] shows that the multiplicative identity is unique, if it exists.
  2. Addition has a cancellation law: for any $a, b, c \in R$, if $a + b = a + c$, then $b = c$.
     - Proof: By [R1]-[R4], $b = 0+b = [(-a)+a]+b = (-a)+[a+b] = (-a)+[a+c] = [(-a)+a]+c = 0+c = c$.
  3. Additive inverses are unique.
     - Proof: Suppose that b and c were both additive inverses of a. Then $a+b = 0 = a+c$, so by property (2), $b = c$.
  4. For any $a \in R$, $0 \cdot a = 0 = a \cdot 0$.
     - Proof: Let b be any element of R. By [R3], [R5] and [R6], we have $b\cdot a+0\cdot a = (b+0)\cdot a = b\cdot a = b\cdot a+0$. Then by property (2), we conclude $0 \cdot a = 0$. A similar argument using distribution on the right shows that $a \cdot 0 = 0$ also.
  5. For any $a \in R$, $-(-a) = a$.
     - Proof: By definition, $-(-a)$ has the property that $(-a) + [-(-a)] = 0$. But by [R2] applied to [R4], we also know $(-a) + a = 0$, so by property (3), we conclude $-(-a) = a$.
  6. If R has a 1, then for any $a \in R$, $(-1) \cdot a = -a = a \cdot (-1)$.
     - Proof: By [R4], [R6], [R8], and the previous property, we have $0 = 0\cdot a = [1+(-1)]\cdot a = 1\cdot a+(-1)\cdot a = a + (-1) \cdot a$. Therefore, $(-1) \cdot a$ is an additive inverse of a, so by property (3), we see $(-1) \cdot a = -a$. In a similar way, we can see that $a \cdot (-1) = -a$.
  7. For any $a, b \in R$, $-(a + b) = (-a) + (-b)$.
     - Proof: By [R1]-[R4], observe that $(b + a) + [(-a) + (-b)] = [b + (a + (-a)] + (-b) = [b + 0] + (-b) = b + (-b) = 0$. Thus, by (3), we conclude that $(-a) + (-b)$ is the additive inverse of $b + a = a + b$.
  8. For any $a, b \in R$, $(-a) \cdot b = -(a \cdot b) = a \cdot (-b)$, and $(-a) \cdot (-b) = a \cdot b$.
     - Proof: Observe that $a \cdot b + (-a) \cdot b = [a + (-a)] \cdot b = 0 \cdot b = 0$ by [R4], [R6], and property (4). Thus, $(-a)\cdot b$ is an additive inverse of $a\cdot b$, so by property (3), it is equal to $-(a\cdot b)$. A similar argument shows that $a\cdot(-b) = -(a\cdot b)$. For the last statement, observe that $(-a)\cdot(-b) = -[a\cdot(-b)] = -(-[a\cdot b]) = a\cdot b$ by the first two statements and property (5).
  9. For any positive integers m and n and any $a \in R$, $ma + na = (m + n)a$, $m(na) = (mn)a$, $a^{m+n} = a^m a^n$, and $a^{mn} = (a^m)^n$.
     - Proof: By definition and [R1], $(m + n)a = \underbrace{a + a + \cdots + a}_{m+n \text{ terms}} = \underbrace{a + a + \cdots + a}_{m \text{ terms}} + \underbrace{a + a + \cdots + a}_{n \text{ terms}} = ma + na$.
     - In a similar way, if $b = na$ then by [R1], $(mn)a = \underbrace{a + a + \cdots + a}_{mn \text{ terms}} = \underbrace{b + b + \cdots + b}_{m \text{ terms}} = mb = m(na)$.
     - The other two properties follow in the same way, using multiplication in place of addition.

- An important property of $\mathbb{Z}$ that does *not* hold in general rings is the statement that $ab = 0$ implies $a = 0$ or $b = 0$.
  - Indeed, we have already seen examples of situations in $\mathbb{Z}/m\mathbb{Z}$ where m is not prime that we can have situations where $ab = 0$ but $a, b \neq 0$: for example, in $\mathbb{Z}/6\mathbb{Z}$, we have the equality $\overline{2} \cdot \overline{3} = \overline{0}$.

- Inversely, it is also possible for a general ring to contain many elements that have multiplicative inverses (unlike in $\mathbb{Z}$, where the only elements with multiplicative inverses are 1 and -1).

- Definition: In a ring R, we say that an element a is a zero divisor if $a \neq 0$ and there exists a nonzero $b \in R$ such that $ab = 0$ or $ba = 0$. (Note in particular that 0 is *not* a zero divisor!)

- Definition: In a ring R with $1 \neq 0$, we say that an element a is a unit if there exists a $b \in R$ such that $ab = 1 = ba$. The set of units in R is denoted $R^\times$.
  - Example: In $\mathbb{Z}$, there are no zero divisors, and the units are ± 1.
  - Example: In $\mathbb{Z}/m\mathbb{Z}$, the units are the residue classes relatively prime to m, while the zero divisors are the nonzero classes having a nontrivial common divisor with m. In particular, every nonzero residue is either a unit or a zero divisor.
  - Example: In a field (or more generally, in a division ring), every nonzero element is a unit. Indeed, a ring R is a division ring precisely when every nonzero element is a unit.
  - Example: In the matrix ring $M_{2\times 2}(\mathbb{R})$, the units are precisely the invertible matrices (namely, those with nonzero determinant). It is a somewhat more difficult task to characterize the zero divisors, but it can be shown that every non-invertible nonzero matrix is a zero divisor.
  - Example: In the ring $\mathbb{Z}[\sqrt{2}]$, the integers 1 and -1 are units, but the element $\sqrt{2}+1$ is also a unit, because $(\sqrt{2}+1)\cdot(\sqrt{2}-1) = 1$. Note that $\mathbb{Z}[\sqrt{2}]$ is not a field, however, because $\sqrt{2}$ is not a unit.
  - Example: In the ring of real-valued functions on a set S, a function f is a unit if it is nonzero everywhere (since its inverse is $(1/f)(x) = \dfrac{1}{f(x)}$). Inversely, if f is zero at any element $s \in S$, then the function $g(x) = \begin{cases} 1 & \text{if } x = s \\ 0 & \text{if } x \neq s \end{cases}$ has the property that fg is identically zero, so f is a zero divisor.
  - If R is commutative, then of course the two zero divisor conditions $ab = 0$ and $ba = 0$ are equivalent.
  - However, there exist examples of elements a, b in noncommutative rings where $ab = 0$ but $ba \neq 0$: for example, if $A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$ and $B = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$, then $AB = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$ while $BA = \begin{bmatrix} 3 & 3 \\ -3 & -3 \end{bmatrix}$.
  - In a similar way, if R is commutative then the two unit conditions $ab = 1$ and $ba = 1$ are equivalent, but there exist examples of elements a, b in noncommutative rings where $ab = 1$ but $ba \neq 1$.[2]

- Here are a few basic properties of units and zero divisors:

- Proposition (Units and Zero Divisors): Let R be a ring with $1 \neq 0$.

  1. The multiplicative inverse of a unit is unique.
     - Proof: If a is a unit with $ab = 1 = ba$ and also $ac = 1 = ca$, then $b = b(ac) = (ba)c = c$.
  2. The product of two units is a unit, as is the multiplicative inverse of a unit.
     - Proof: If a is a unit with $ab = 1 = ba$, then by definition b is also a unit.
     - If c is another unit with $cd = 1 = dc$, then $(ac)(db) = a(cd)b = a1b = ab = 1$ and likewise $(db)(ac) = 1$ as well, so the inverse of ac is db.
  3. A unit can never be a zero divisor in R.
     - Proof: Suppose first that a is a unit and that $xa = 0$ for some $x \neq 0$.
     - Then by assumption, there is a b such that $ab = 1$, so then $x = x(ab) = (xa)b = 0b = 0$, contradicting the assumption that $x \neq 0$.
     - In the same way, if $ax = 0$ for some $x \neq 0$, then if $ba = 1$ then $x = (ba)x = b(ax) = b0 = 0$, again a contradiction.

  - Remark (for those who like group theory): Together with the observation that 1 is a unit, parts (1) and (2) imply that the set of units $R^\times$ forms a group under multiplication. (For this reason $R^\times$ is usually called the "group of units" of the ring R.)

- We can adapt the concept of divisibility directly into the setting of a general commutative ring R:

- Definition: If R is a commutative ring and $a, b \in R$, we say that $a|b$ if there exists some $k \in R$ such that $b = ak$.

[2] One example of this phenomenon is in the ring $\mathcal{L}(V,V)$ where $V = \{a_1, a_2, \dots\}$ is the vector space of infinite sequences of real numbers. If L is the "left shift operator" with $L(a_1, a_2, a_3, \dots) = (a_2, a_3, \dots)$ and R is the "right shift operator" with $R(a_1, a_2, a_3, \dots) = (0, a_1, a_2, a_3, \dots)$, then $RL = I$ but $LR \neq I$.

- Example: In $\mathbb{Z}[i]$, the element $2+i$ divides 5, because $5 = (2+i)(2-i)$.
- Example: In $\mathbb{Z}/6\mathbb{Z}$, the element $\overline{4}$ divides $\overline{2}$, because $\overline{2} = \overline{2} \cdot \overline{4}$.
- Notice that if u is any unit in R, then $d|a$ is equivalent to $(ud)|a$.
- Warning: If R does not have a 1, bizarre things can occur with divisibility. For example, let R be the ring consisting of the even integers. Then in this ring, it is not the case that $2|6$, because there is no even integer k such that $6 = 2k$. Indeed, it is not even the case that $2|2$ in this ring!
- Many of the properties of divisibility (particularly those regarding gcds) do not hold in general rings; we will postpone a more careful analysis of these properties until later.

- Definition: If R is a commutative ring with 1 and $a' = ua$ for some unit u, we say that a and a' are associates.
  - Notice that if a and a' are associates, then $a|a'$ and $a'|a$. For this reason, associates have very similar divisibility properties to one another.
  - Example: In $\mathbb{Z}$, 2 and -2 are associates.

- We give a special name to the class of commutative rings having no zero divisors, attesting to their similarity to $\mathbb{Z}$:

- Definition: A commutative ring with $1 \neq 0$ having no zero divisors is called an integral domain (or often, just a "domain"). Equivalently, R is an integral domain if R is commutative with $1 \neq 0$, and where $ab = 0$ implies $a = 0$ or $b = 0$.
  - The integers are an integral domain, as is any field.
  - More generally, any ring that is a subset of a field (such as the Gaussian integers $\mathbb{Z}[i]$) is an integral domain. In fact, as we will describe later, the converse turns out to be true as well: any integral domain arises naturally as a subset of a field.

- Integral domains possess various fundamental properties:

- Proposition (Cancellation in Domains): Suppose R is an integral domain. Then multiplication in R has a cancellation law: if $a \neq 0$ and $ab = ac$, $b = c$.
  - Proof: Suppose that $ab = ac$: then $a(b-c) = 0$, so since R is a domain we either have $a = 0$ or $b - c = 0$. Thus, if $a \neq 0$, we have $b - c = 0$ so that $b = c$.

- Corollary: If R is a finite integral domain, then R is a field.
  - Proof: Let a be any nonzero element of R, and consider the set $\{a, a^2, a^3, \dots, a^n, \dots\}$. Since R is finite, two of the elements of this set must be equal: say $a^j = a^{j+k}$ for some positive integers j and k.
  - Then $a^j = a^{j+k}$ implies $a^j(a^k - 1) = 0$, and then since $a \neq 0$, we see $a^j \neq 0$. Thus, $a^k - 1 = 0$, so that $a \cdot a^{k-1} = 1$, meaning that a^{k-1} is the multiplicative inverse of a.

- As a final remark, we note that there is a generalization of the above result due to Wedderburn:

- Theorem (Wedderburn): A finite division ring is a field.
  - We will not prove this theorem, although the methods of the proof are comparatively elementary.

2.1.4 Subrings and Cartesian Products

- A number of the examples of rings we described earlier arise naturally as subsets of other rings. We can easily describe this phenomenon in general:

- Definition: If R is a ring, we say a subset S of R is a subring if it also possesses the structure of a ring, under the same operations as R.
  - Observe that if S is a subset of a ring, in order for the operations $+$ and $\cdot$ to be well-defined binary operations on S, it must be the case that $a + b$ and $a \cdot b$ are elements of S, for any a and b in S.

- Next, observe that axioms [R1], [R2], [R5], and [R6] in S automatically follow from the corresponding properties of R.
- In order for [R3] to hold, we must have an additive identity 0_S in S with the property that $a + 0_S = a$ for every a in S. However, by the additive cancellation law in R, since $a + 0_R = a = a + 0_S$, we see that $0_S = 0_R$: in other words, S must contain the zero element of R.
- Finally, in order for [R4] to hold, we require that for every $a \in S$, its additive inverse $(-a)$ must also be in S.

- By employing subtraction, we can in fact combine two of these verifications:

- Proposition (Subring Criterion): A subset S of R is a subring if only if S contains the zero element of R and, for any $a, b \in S$, the elements $a - b$ and ab are also in S.
  - Proof: If S is a subring, then as noted above S must contain the zero element of R and for any $a, b \in S$, we must have $a - b$ and ab in S.
  - Conversely, suppose S contains 0 and that $a - b$ and ab are also in S for any $a, b \in S$. By setting $a = 0$ we see that $0 - b = -b$ is in S, and then by setting $b = -c$ we see that $a - (-c) = a + c$ is in S.
  - Therefore, S contains 0 and is closed under addition, multiplication, and taking additive inverses. By the observations above, S is therefore a subring.

- Using the subring criterion, we can construct many more examples of rings.

- Example: $\mathbb{Z}$ is a subring of $\mathbb{Q}$, which is a subring of $\mathbb{R}$, which is a subring of $\mathbb{C}$.

- Example: The trivial ring $\{0\}$ is a subring of any ring.

- Example: The even integers $2\mathbb{Z}$ are a subring of $\mathbb{Z}$, as are (more generally) the integer multiples of n, written $n\mathbb{Z}$.
  - In fact, every subring of $\mathbb{Z}$ is of the form $n\mathbb{Z}$ for some integer n (the "trivial" subring $\{0\}$ corresponds to the case $n = 0$).
  - To see this, suppose S is a subring of $\mathbb{Z}$, and let T be the set of positive elements in this subring. If T is empty, then $S = \{0\}$, and otherwise, T must have a minimal element n by the well-ordering principle. Then S contains $n\mathbb{Z}$.
  - We claim any element of S must be a multiple of n, so that $S = n\mathbb{Z}$: by the division algorithm, if S contained an integer not divisible by n, the remainder upon dividing a by n would be a positive element of S smaller than n, contradiction. Thus, $S = n\mathbb{Z}$.

- Example: The set of rational numbers having denominator equal to a power of 2 (i.e., that are of the form $\dfrac{n}{2^k}$ for an integer n and nonnegative integer k), forms a subring of $\mathbb{Q}$.

- Example: The set of upper-triangular 2×2 matrices with real entries (i.e., those of the form $\begin{bmatrix} a & b \\ 0 & c \end{bmatrix}$) forms a subring of $M_{2\times 2}(\mathbb{R})$.

- Example: The set of matrices of the form $\begin{bmatrix} a & a \\ a & a \end{bmatrix}$, for a real number a, forms a subring of $M_{2\times 2}(\mathbb{R})$.
  - It is straightforward to see that this ring is commutative and even has a multiplicative identity, namely, the matrix $\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix}$.
  - This example illustrates an important point: a subring of a ring R with 1 may also be a ring with 1, but the multiplicative identity in S may not be the same multiplicative identity as in R.

- Example: The set of differentiable real-valued functions is a subring of the ring of continuous real-valued functions, which is in turn a subring of the ring of all real-valued functions.

- To make these observations, we need only observe that the difference and product of differentiable functions are differentiable, and likewise the difference and product of continuous functions are continuous.

- We can also construct new rings using Cartesian products.
    - Recall that if S and T are sets, the Cartesian product $S \times T$ is the set of ordered pairs (s, t) where $s \in S$ and $t \in T$.

- Proposition (Cartesian Products of Rings): If A and B are rings, then the Cartesian product $A \times B$ is also a ring, with operations performed componentwise: $(a_1, b_1) + (a_2, b_2) = (a_1 + a_2, b_1 + b_2)$ and $(a_1, b_1) \cdot (a_2, b_2) = (a_1 a_2, b_1 b_2)$.
    - More generally, the Cartesian product of a collection of rings R_i indexed by the set I is also a ring, again with operations performed componentwise.
    - Proof: Each of the properties [R1]-[R6] follows from the corresponding properties of A and B. The additive identity is $(0, 0)$, and additive inverses are given by $-(a, b) = (-a, -b)$.
    - Note that if A and B are commutative, then so is $A \times B$; likewise, if A and B have a 1, then $(1_A, 1_B)$ is the multiplicative identity in $A \times B$.

- Example: The ring $(\mathbb{Z}/2\mathbb{Z}) \times (\mathbb{Z}/3\mathbb{Z})$ is a commutative ring with 1.
    - This ring has six elements: $(0, 0)$, $(0, 1)$, $(0, 2)$, $(1, 0)$, $(1, 1)$, and $(1, 2)$. The additive identity is $(0, 0)$ and the multiplicative identity is $(1, 1)$.
    - Operations are performed "modulo 2" in the first coordinate and "modulo 3" in the second coordinate, so for example we have $(1, 1) + (0, 2) = (1, 0)$ and $(1, 2) \cdot (0, 2) = (0, 1)$.

2.1.5 Quadratic Fields and Quadratic Integer Rings

- We conclude our discussion with a brief analysis of an important class of rings known as the quadratic integer rings.

- Definition: Let D be a squarefree integer not equal to 1. The quadratic field $\mathbb{Q}(\sqrt{D})$ is the set of complex numbers of the form $a + b\sqrt{D}$, where a and b are rational numbers.
    - Remark: An integer is squarefree if it is not divisible by the square of any prime, and not equal to 1. We lose nothing here by assuming that D is a squarefree integer, since two different integers differing by a square would generate the same set of complex numbers $a + b\sqrt{D}$.
    - The arithmetic in $\mathbb{Q}(\sqrt{D})$ is as follows: $(a+b\sqrt{D})+(c+d\sqrt{D}) = (a+c)+(b+d)\sqrt{D}$, and $(a+b\sqrt{D})(c+d\sqrt{D}) = (ac + Dbd) + (ad + bc)\sqrt{D}$.
    - Since $\mathbb{Q}(\sqrt{D})$ is clearly closed under subtraction and multiplication, and contains $0 = 0 + 0\sqrt{D}$, it is a subring of $\mathbb{C}$ and hence an integral domain, since it contains 1.
    - It is in fact a field (justifying the name "quadratic field") because we can write $(a+b\sqrt{D})^{-1} = \dfrac{a - b\sqrt{D}}{a^2 - Db^2}$, and $a^2 - Db^2 \neq 0$ provided that a and b are not both zero because $\sqrt{D}$ is irrational by the assumption that D is squarefree and not equal to 1.

- Definition: The field norm $N : \mathbb{Q}(\sqrt{D}) \to \mathbb{Q}$ is defined to be the function $N(a + b\sqrt{D}) = a^2 - Db^2 = (a + b\sqrt{D})(a - b\sqrt{D})$.
    - The fundamental property of this field norm is that it is multiplicative: $N(xy) = N(x)N(y)$ for two elements x and y in $\mathbb{Q}(\sqrt{D})$, as can be verified by writing out both sides explicitly and comparing the results.
    - The field norm provides a measure of "size" of an element of $\mathbb{Q}(\sqrt{D})$, in much the same way that the complex absolute value measures the "size" of a complex number. In fact, if $D < 0$, then the field norm of an element $a + b\sqrt{D}$ is the same as the square of its complex absolute value.

- A fundamental subring of the quadratic field $\mathbb{Q}(\sqrt{D})$ is its associated "quadratic integer ring".
  - ○ The most obvious choice for an analogy of the integers $\mathbb{Z}$ inside $\mathbb{Q}(\sqrt{D})$ would be the set $\mathbb{Z}[\sqrt{D}] = \{a + b\sqrt{D} : a, b \in \mathbb{Z}\}$, which is easily seen to be a subring.
  - ○ However, notice that if $D \equiv 1 \pmod 4$, then the slightly larger subset $\mathbb{Z}[\frac{1+\sqrt{D}}{2}] = \{a + b\frac{1+\sqrt{D}}{2} : a, b \in \mathbb{Z}\}$ is actually also a subring: closure under subtraction is obvious, and for multiplication we can write $(a + b\frac{1+\sqrt{D}}{2})(c + d\frac{1+\sqrt{D}}{2}) = (ac + \frac{D-1}{4}bd) + (ad + bc + bd)\frac{1+\sqrt{D}}{2}$.
  - ○ One reason that this slightly larger set turns out to give a slightly better analogy for the integers $\mathbb{Z}$ when $D \equiv 1 \pmod 4$ is that the number $\frac{1+\sqrt{D}}{2}$ satisfies a polynomial with integer coefficients and leading coefficient 1: explicitly, it is a root of $x^2 - x + \frac{1-D}{4} = 0$.

- Definition: The ring of integers $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ in the quadratic field $\mathbb{Q}(\sqrt{D})$ is defined to be $\mathbb{Z}[\sqrt{D}]$ if $D \equiv 2$ or $3 \pmod 4$, and is $\mathbb{Z}[\frac{1+\sqrt{D}}{2}]$ if $D \equiv 1 \pmod 4$. Each of these rings is an integral domain.
  - ○ We have already briefly mentioned two examples of these rings: the Gaussian integers $\mathbb{Z}[i]$ corresponding to $D = -1$, and the ring $\mathbb{Z}[\sqrt{2}]$ corresponding to $D = 2$.
  - ○ For $D \equiv 2, 3 \pmod 4$, observe that $N(a + b\sqrt{D}) = a^2 - Db^2$ is an integer for every $a + b\sqrt{D} \in \mathcal{O}_{\sqrt{D}}$.
  - ○ Likewise, if $D \equiv 1 \pmod 4$, we have $N(a + b\frac{1+\sqrt{D}}{2}) = a^2 + ab + \frac{1-D}{4}b^2$ is also an integer for every $a + b\frac{1+\sqrt{D}}{2} \in \mathcal{O}_{\sqrt{D}}$.
  - ○ Thus, the field norm N is always integer-valued on $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$. We can in fact use it to determine whether a given element is a unit:

- Proposition (Characterizing Units in $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$): An element r in the ring $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ is a unit if and only if $N(r) = \pm 1$.
  - ○ Proof: Suppose $r = a + b\sqrt{D}$ and let $\overline{r} = a - b\sqrt{D}$, so that $N(r) = r\overline{r}$. (Note that $\overline{r} = 2a - r$, so that even when $D \equiv 1 \pmod 4$, so that a and b are possibly half-integers, we see that $\overline{r}$ is still in $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$.)
  - ○ If $N(r) = \pm 1$, then we see that $r\overline{r} = \pm 1$, so (by multiplying by -1 if necessary) we obtain a multiplicative inverse for r.
  - ○ Conversely, suppose r is a unit and $rs = 1$. Taking norms yields $N(r)N(s) = N(rs) = 1$. Since $N(r)$ and $N(s)$ are both integers, we see that $N(r)$ must either be 1 or -1.

- Example: Find the units in $\mathbb{Z}[i]$ and $\mathbb{Z}[(1 + \sqrt{-3})/2]$.
  - ○ For $\mathbb{Z}[i]$, we have $D = -1$, so if $r = a + bi$ we see $N(r) = a^2 + b^2$. We must therefore solve $a^2 + b^2 = 1$ in $\mathbb{Z}$: there are clearly four solutions, corresponding to $r = \boxed{1,\ i,\ -1,\ -i}$.
  - ○ For $\mathbb{Z}[(1 + \sqrt{-3})/2]$, we have $D = -3$, so if $r = a + b\frac{1+\sqrt{-3}}{2}$ we see $N(r) = a^2 + ab + b^2$. We must therefore solve $a^2 + ab + b^2 = 1$ in $\mathbb{Z}$: by multiplying by 4 and completing the square, this equation is equivalent to $(2a + b)^2 + 3b^2 = 4$, which has six solutions corresponding to $r = \boxed{1,\ -1,\ \omega,\ -\omega,\ \omega^2,\ -\omega^2}$, where $\omega = \frac{1+\sqrt{-3}}{2}$ is seen to be a sixth root of unity satisfying $\omega^6 = 1$.

- In general, determining the full set of units in $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ is a nontrivial computation.
  - ○ When $D < 0$ it is not too difficult to see (by completing the square in a similar way to above when $D \equiv 1 \pmod 4$) that if $D \neq -1, -3$, then the only units in $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$ are ± 1.

- When $D > 0$, however, it turns out that there are always infinitely many units in $\mathcal{O}_{\mathbb{Q}(\sqrt{D})}$, and they are all of the form $\pm u^k$ (for some integer k) for a certain "primitive unit" u. Computing these units is essentially equivalent, per the proposition above, to solving the equation $a^2 - Db^2 = \pm 1$, known as Pell's equation.
- For example, it can be shown that the units in $\mathbb{Z}[\sqrt{2}]$ are $\pm(1+\sqrt{2})^k$, while the units in $\mathbb{Z}[\sqrt{3}]$ are $\pm(2+\sqrt{3})^k$ for integers k.

2.2 Polynomial Rings

- We now discuss a fundamental class of rings known as polynomial rings.
  - Polynomials with real coefficients (like $p(x) = 1+x^2$ or $q(x) = 3+\pi x^2$) are likely familiar from elementary algebra.
  - Unlike in elementary algebra, however, our polynomials will be "formal symbols" rather than functions. We will soon exploit the connection between polynomials and functions, but, as we will discuss, there are very important reasons for us to take a more abstract approach to polynomials than simply viewing them as functions.
- In this section, let F be a field.
  - For ease of notation, we will write $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ for p a prime, since (as we have discussed) the ring $\mathbb{Z}/p\mathbb{Z}$ is a field.
  - Frequently, $\mathbb{F}_p$ is called "the field with p elements". It is also occasionally denoted $GF(p)$, for "the Galois Field with p elements", since these fields were first extensively studied by Évariste Galois.
  - The uses of the word "the" are justified, as we will show later that (up to relabeling the elements), there is only one field with p elements for any prime p.

2.2.1 Definition and Basic Properties

- Definition: Let R be a ring and x be an indeterminate. A polynomial in x with coefficients in R consists of a formal sum $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$, for an integer $n \geq 0$ and where each element $a_i \in R$.
  - The term "indeterminate" is deliberately undefined in the definition above. A more concrete[3] (but vastly less intuitive) definition of polynomials can be given using Cartesian product, but we will not use it.
  - If $a_n \neq 0$, we say that the polynomial has degree n, and if $a_n = 1$ we say the polynomial is monic. (By convention, the degree of the zero polynomial 0 is $-\infty$.)
  - The leading term of the polynomial is its highest-degree term (i.e., $a_n x^n$) and its leading coefficient is the corresponding coefficient (i.e., a_n).
  - We will employ the traditional "function" notation for polynomials (e.g., by writing a polynomial as $p(x) = x^2 + 5$), and also often drop the variable portion (e.g., by referring to "the polynomial p") when convenient. We reiterate, however, that our polynomials are *not* functions, but rather formal sums.
- As is familiar from elementary algebra, the polynomials with coefficients in R have a natural ring structure:
- Definition: The polynomial ring $R[x]$ consists of the polynomials in x with coefficients in R, under the two ring operations of addition and multiplication, defined as follows:
  - Addition is defined "termwise":

$$(a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0) + (b_n x^n + b_{n-1}x^{n-1} + \cdots + b_1 x + b_0) = (a_n + b_n)x^n + (a_{n-1} + b_{n-1})x^{n-1} + \cdots + (a_0 + b_0).$$

[3] Specifically: inside the Cartesian product $\prod_{\mathbb{Z}_{\geq 0}} R = (r_0, r_1, r_2, \dots)$ indexed by the nonnegative integers, we define the "polynomials" to be the sequences all but finitely many of whose entries are zero, and interpret the sequence $(r_0, r_1, r_2, \dots, r_n, 0, 0, \dots)$ as the formal sum $r_0 + r_1 x + r_2 x^2 + \cdots + r_n x^n$. We can then define the operations of polynomial addition and multiplication solely in terms of these sequences.

  - Multiplication is defined first on "monomials" (polynomials with only one nonzero coefficient), via $(ax^n) \cdot (bx^m) = abx^{n+m}$, and then extended to arbitrary polynomials via the distributive laws. Explicitly, we have

    $$(a_0+a_1x+a_2x^2+\cdots+a_nx^n)\cdot(b_0+b_1x+b_2x^2+\cdots+b_mx^m) = a_0b_0+(a_1b_0+a_0b_1)x+(a_2b_0+a_1b_1+a_0b_2)x^2+\cdots+a_nb_mx^{n+m}$$

    where the coefficient of x^j in the product is given by $\displaystyle\sum_{k=0}^{j} a_k b_{j-k}$.
  - These operations are all defined, since R is a ring, and it is reasonably straightforward to see that each of the ring axioms hold: thus, $R[x]$ is a ring.

- Observe that if R is commutative, then so is $R[x]$; likewise, if R has a 1, then so does $R[x]$. Note also that R appears as a subring of $R[x]$ in the form of the "constant polynomials".
  - We will primarily be interested in the case where R is a commutative ring with 1, since polynomial rings over a noncommutative ring tend to have unusual ("pathological") properties with regard to factorization.

- Example: In $\mathbb{Z}[x]$, find $p(x) + q(x)$ and $p(x) \cdot q(x)$ for $p(x) = 2 + 3x$ and $q(x) = 3 + 2x$.
  - We have $p(x) + q(x) = \boxed{5 + 5x}$, while $p(x)q(x) = 6 + (4 + 9)x + 6x^2 = \boxed{6 + 13x + 6x^2}$.

- Example: In $\mathbb{F}_5[x]$, find $p(x) + q(x)$ and $p(x) \cdot q(x)$ for $p(x) = 2 + 3x$ and $q(x) = 3 + 2x$.
  - We have $p(x) + q(x) = 5 + 5x = \boxed{0}$, while $p(x)q(x) = 6 + (4 + 9)x + 6x^2 = \boxed{1 + 3x + x^2}$.

- Example: In $R[x]$ with $R = \mathbb{Z}/6\mathbb{Z}$, find $p(x) + q(x)$ and $p(x) \cdot q(x)$ for $p(x) = 2 + 3x$ and $q(x) = 3 + 2x$.
  - We have $p(x) + q(x) = \boxed{5 + 5x}$, while $p(x)q(x) = 6 + (4 + 9)x + 6x^2 = \boxed{x}$.

- Different coefficient rings R can cause substantial differences in the behavior of the resulting polynomial ring $R[x]$.
  - In the examples above, notice that when the ring R was $\mathbb{Z}$ or $\mathbb{Z}/5\mathbb{Z}$, the product of the two degree-1 polynomials $2 + 3x$ and $3 + 2x$ had degree 2, while for $R = \mathbb{Z}/6\mathbb{Z}$, the product of these two degree-1 polynomials only had degree 1.
  - Ultimately, the difference arises because the product of the leading coefficients (2 and 3) in $\mathbb{Z}/6\mathbb{Z}$ is equal to 0: the presence of zero divisors causes the degree to drop.
  - Another dramatic example is this phenomenon is that the product of the two degree-2 polynomials $(1 + 2x^2) \cdot (1 + 2x^2)$ in $\mathbb{Z}/4\mathbb{Z}$ is equal to 1 (which has degree 0).

- Proposition (Degrees in Polynomial Rings): If p and q are any polynomials in a polynomial ring $R[x]$, then $\deg(p + q) \leq \max(\deg p, \deg q)$, and $\deg(p \cdot q) \leq \deg p + \deg q$.
  - Proof: It is straightforward to verify that each claim holds if p or q is zero (in which case the left side of each inequality is $-\infty$). Now assume p and q are nonzero.
  - For $p + q$, observe that if there are no terms of degree n or higher in p or q, then there are no terms of degree n or higher in $p + q$ either.
  - For $p \cdot q$, observe that if $\deg p = n$ and $\deg q = m$, then every term in $p \cdot q$ has degree at most $m + n$.

- We can say substantially more when R is an integral domain:

- Proposition (Polynomials Over Integral Domains): If R is an integral domain, then $R[x]$ is also an integral domain. Furthermore, the units in $R[x]$ are precisely the units in R (as constant polynomials), and for any polynomials p and q in $R[x]$, we have $\deg(pq) = \deg p + \deg q$.
  - Proof: Since R is an integral domain, then R has no zero divisors: thus, if p has leading term a_nx^n and q has leading term a_mx^m, then the leading term of $p \cdot q$ will be $a_nb_mx^{m+n}$, which is nonzero since $a_n \neq 0$ and $b_m \neq 0$ by hypothesis.

- Thus, $\deg(pq) = \deg p + \deg q$. In particular, we see that $pq = 0$ only when $p = 0$ or $q = 0$, so $R[x]$ is an integral domain.
- Finally, each unit of R is clearly a unit of $R[x]$; conversely, if $pq = 1$, then $\deg p = \deg q = 0$, so p and q are constant polynomials: then by definition, they are units of R.

2.2.2 The Division Algorithm in $F[x]$

- Much like $\mathbb{Z}$, polynomials with coefficients in the field F also possess a "long division" algorithm: the only difference is that we measure the "size" of a polynomial via its degree.
- Theorem (Division Algorithm in $F[x]$): If F is any field, and $a(x)$ and $b(x)$ are any polynomials in $F[x]$ with $b(x) \neq 0$, then there exist unique polynomials $q(x)$ and $r(x)$ such that $a(x) = b(x)q(x) + r(x)$, where $\deg(r) < \deg(b)$. Furthermore, $b|a$ if and only if $r = 0$.
  - The idea is simply to show the validity of "polynomial long division". The reason we require F to be a field is that we need to be able to divide by arbitrary nonzero coefficients to be able to perform the divisions. (Over $\mathbb{Z}$, for instance, we cannot divide x^2 by $2x$ and get a remainder that is a constant polynomial.)
  - For example, when we divide the polynomial $x^3 + x^2 + 3x + 5$ by the polynomial $x^2 + 3x + 1$ in $\mathbb{R}[x]$, we obtain the quotient $q(x) = x - 2$ and remainder $r(x) = 8x + 7$: indeed, we have $x^3 + x^2 + 3x + 5 = (x - 2)(x^2 + 3x + 1) + (8x + 7)$.
  - Proof: We prove this by induction on the degree n of $a(x)$. The base case is trivial, as we may take $q = r = 0$ if $a = 0$.
  - Now suppose the result holds for all polynomials $a(x)$ of degree $\leq n - 1$. If $\deg(b) > \deg(a)$ then we can simply take $q = 0$ and $r = a$, so now also assume $\deg(b) \leq \deg(a)$.
  - Write $a(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ and $b(x) = b_m x^m + \cdots + b_0$, where $b_m \neq 0$ since $b(x) \neq 0$.
  - Observe that the polynomial $a^\dagger(x) = a(x) - \dfrac{a_n}{b_m} x^{n-m} b(x)$ has degree less than n, since we have cancelled the leading term of $a(x)$. (Here we are using the fact that F is a field, so that $\dfrac{a_n}{b_m}$ also lies in F.)
  - By the induction hypothesis, $a^\dagger(x) = q^\dagger(x)b(x) + r^\dagger(x)$ for some $q^\dagger(x)$ and $r^\dagger(x)$ with $r^\dagger = 0$ or $\deg(r^\dagger) < \deg(b)$.
  - Then $a(x) = \left[q^\dagger(x) + \dfrac{a_n}{b_m} x^{n-m}\right] b(x) + r^\dagger(x)$, so $q(x) = q^\dagger(x) + \dfrac{a_n}{b_m} x^{n-m}$ and $r(x) = r^\dagger(x)$ satisfy all of the requirements.
  - For the uniqueness, suppose that $a = qb + r = q'b + r'$: then $r - r' = b(q' - q)$ has degree less than $\deg(b)$ but is also divisible by b, hence must be zero.
  - Finally, by definition if $r = 0$ then $b|a$, and conversely if $b|a$ then since r is unique we must have $r = 0$.
- The existence of this division algorithm in $F[x]$ allows us to adapt many results that hold in $\mathbb{Z}$ into this setting. First is the idea of a common divisor:
- Definition: If a and b are polynomials in $F[x]$, we say a polynomial d is a common divisor if $d|a$ and $d|b$.
  - Example: The polynomial $x+1$ is a common divisor of $x^2 - 1$ and $x^2 + 3x + 2$ in $\mathbb{R}[x]$, as is the polynomial $2x + 2$.
- We would naturally want to define the greatest common divisor to be the polynomial of largest degree dividing both a and b.
  - However, this polynomial is not unique: in the example above, it is easy to see that $x^2 - 1$ and $x^2 + 3x + 2$ do not have a common divisor of degree 2 (or larger), so both $x + 1$ and $2x + 2$ are common divisors of maximal degree.
  - Ultimately, $x + 1$ and $2x + 2$ are essentially the same (as far as divisibility goes), since they are associates.

- Definition: If a and b are polynomials in $F[x]$, not both zero, we say the polynomial d is a greatest common divisor of a and b if it a common divisor of a and b with the property that if d' is any other common divisor, then $d'|d$.

  - Under this definition, we can verify that both $x+1$ and $2x+2$ are gcds of x^2-1 and x^2+3x+2.

- This definition does not immediately imply that a gcd actually exists. But by adapting the Euclidean algorithm to this setting, we can give a procedure for computing the gcd (and in particular, implying that it exists and is unique) and for writing it as a linear combination:

- Algorithm (Euclidean Algorithm in $F[x]$): Given polynomials a and b in $F[x]$, not both zero, repeatedly apply the division algorithm as follows, until a remainder of zero is obtained:

$$\begin{aligned} a &= q_1 b + r_1 \\ b &= q_2 r_1 + r_2 \\ r_1 &= q_3 r_2 + r_3 \\ &\vdots \\ r_{k-1} &= q_k r_k + r_{k+1} \\ r_k &= q_{k+1} r_{k+1}. \end{aligned}$$

  Then the last nonzero remainder r_{k+1} is a gcd of a and b. Furthermore, by successively solving for the remainders and plugging in the previous equations, r_{k+1} (and thereby $\gcd(a,b)$) can be explicitly written as a linear combination of a and b.

  - Proof: First observe that the algorithm will eventually terminate with a zero remainder, because $\deg(b) > \deg(r_1) > \deg(r_2) > \cdots$, and the well-ordering principle dictates that there cannot exist an infinite decreasing sequence of nonnegative integers.
  - By an easy induction, we can see that if $d|a$ and $d|b$, then $d|r_j$ for each $j \geq 1$: thus, any common divisor of a and b must divide r_{k+1}.
  - Conversely, by another easy induction, r_{k+1} divides each r_j for each $j \geq 1$, and thus r_{k+1} divides both a and b.
  - Therefore, r_{k+1} divides both a and b, and any other common divisor also divides r_{k+1}: thus, r_{k+1} is a gcd of a and b.
  - The correctness of the algorithm for computing the gcd as a linear combination follows by an easy induction.

- If a and b are not both zero, we can make the gcd unique by additionally requiring that it be monic (i.e., have leading coefficient 1).

  - Explicitly, if d_1 and d_2 are both common divisors of a and b, then $d_1|d_2$ and $d_2|d_1$, so that $d_1 = sd_2$ and $d_2 = td_1$ for some polynomials s and t.
  - By comparing degrees, we see that $\deg(s) = \deg(t) = 0$, meaning that s and t must both be elements of $F^\times$.
  - In other words: any two common divisors of a and b must be multiples of one another. Thus, there is a unique gcd whose leading coefficient is 1.

- Example: Find "the" greatest common divisor $d(x)$ of the polynomials $p = x^6+2$ and $q = x^8+2$ in $\mathbb{F}_3[x]$, and then write the gcd as a linear combination of p and q.

  - We apply the Euclidean algorithm: we have

$$\begin{aligned} x^8+2 &= x^2(x^6+2) + (x^2+2) \\ x^6+2 &= (x^4+x^2+1)(x^2+2) \end{aligned}$$

    and so the last nonzero remainder is $\boxed{x^2+2}$.

- By back-solving, we see that $x^2+2 = \boxed{1\cdot(x^8+2) - x^2(x^6+2)}$.

- When performing the Euclidean algorithm in $F[x]$, the coefficients can often become quite large or complicated:

- Example: Find "the" greatest common divisor $d(x)$ of the polynomials $p = x^3+7x^2+9x-2$ and $q = x^2+4x$ in $\mathbb{R}[x]$, and then write the gcd as a linear combination of p and q.
  - We apply the Euclidean algorithm: we have

$$\begin{aligned} x^3+7x^2+9x-2 &= (x+3)(x^2+4x)+(-3x-2) \\ x^2+4x &= (-\frac{10}{9}-\frac{1}{3}x)(-3x-2)+(-20/9) \\ -3x-2 &= \frac{27x+6}{20}(-20/9) \end{aligned}$$

  and so the last nonzero remainder is $-20/9$. Thus, by rescaling, we see that the gcd is $\boxed{1}$.
  - By back-solving, we see that

$$\begin{aligned} -3x-2 &= 1\cdot(x^3+7x^2+9x-2)-(x+3)\cdot(x^2+4x) \\ -20/9 &= x^2+4x+(\frac{10}{9}+\frac{1}{3}x)(-3x-2) \\ &= (\frac{10}{9}+\frac{1}{3}x)\cdot(x^3+7x^2+9x-2)-(\frac{7}{3}+\frac{19}{9}x+\frac{1}{3}x^2)\cdot(x^2+4x) \end{aligned}$$

  and thus by rescaling, we obtain $1 = \boxed{(-\frac{1}{2}-\frac{3}{20})\cdot(x^3+7x^2+9x-2)+(\frac{21}{20}+\frac{19}{20}x+\frac{3}{20}x^2)\cdot(x^2+4x)}$.

2.2.3 Irreducible Polynomials and Unique Factorization

- We next develop the polynomial analogue of the prime factorization of an integer: namely, writing a polynomial as a product of irreducible factors, and showing that this factorization is essentially unique.
  - We would like to say that a polynomial is irreducible if it has no divisors of smaller positive degree. For technical reasons, however, we will instead phrase this idea using units:

- Definition: Let R be an integral domain. A nonzero element $r \in R$ is irreducible if it is not a unit and, for any "factorization" $p = bc$ with $b, c \in R$, one of b and c must be a unit.
  - Example: The irreducible elements of $\mathbb{Z}$ are precisely the prime numbers (and their negatives).
  - A ring element that is not irreducible and not a unit is called reducible: it can be written as $r = ab$ where neither a nor b is a unit.
  - Since the units in $F[x]$ are precisely the nonzero constant polynomials, we see that the irreducible polynomials in $F[x]$ are easily seen to be the ones that cannot be factored into a product of polynomials of smaller positive degree.
  - Example: Any polynomial of degree 1 is irreducible.
  - Example: The polynomial x^2+x+1 is irreducible in $(\mathbb{Z}/2\mathbb{Z})[x]$, since the only possible factorizations would be $x\cdot x$, $x\cdot(x+1)$, or $(x+1)\cdot(x+1)$, and none of these is equal to x^2+x+1.
  - Example: The polynomial x^4+4 is reducible in $\mathbb{Q}[x]$, since we can write $x^4+4 = (x^2+2x+2)(x^2-2x+2)$.
  - Example: The polynomial x^2+1 is irreducible in $\mathbb{R}[x]$, since there is no way to write it as the product of two linear polynomials with real coefficients.
  - Important Warning: Whether a given polynomial is irreducible depends on the ring $F[x]$ we are considering it as an element of. For example, x^2+1 is irreducible in $\mathbb{R}[x]$ but not in $\mathbb{C}[x]$, since we can write $x^2+1 = (x+i)(x-i)$ in $\mathbb{C}[x]$.

- The irreducible polynomials are the "building blocks under multiplication" in $F[x]$, much like the primes are in $\mathbb{Z}$, because every nonzero polynomial can be written as the product of irreducibles:

- Proposition (Factorization into Irreducibles): Every polynomial of positive degree in $F[x]$ can be written as a product of irreducible polynomials (where a "product" is allowed to have only one term).
  - Proof: We use strong induction on $n = \deg(p)$. The result clearly holds if $n = 1$, since any polynomial of degree 1 is irreducible.
  - Now suppose $n \geq 2$. If p is irreducible, we are done, so otherwise assume that p is reducible.
  - By definition, there exist polynomials a, b with $0 < \deg(a), \deg(b) < n$ with $p = ab$.
  - By the strong induction hypothesis, both a and b can be written as a product of irreducibles; multiplying these two products then gives p as a product of irreducibles.
- In order to show that the factorization into irreducibles is unique, we need the analogous divisibility property that we required in $\mathbb{Z}$:
- Proposition (Irreducibles are Prime in $F[x]$): If $p \in F[x]$ is irreducible and $p|ab$, then $p|a$ or $p|b$.
  - Proof: Suppose $p|ab$. If $p|a$, we are done, so suppose $p \nmid a$, and let d be a gcd of p and a.
  - By hypothesis, d divides p, so (since p is irreducible) either d is a unit, or $d = up$ for some unit u: however, the latter cannot happen, because then up (hence p) would divide a.
  - Hence d is a unit, say with inverse e.
  - By the Euclidean algorithm, we see that there exist x and y such that $xp + ya = d$.
  - Multiplying by be and regrouping the terms yields $(bce)p + ey(ab) = (de)b = b$. Since p divides both terms on the left-hand side, we conclude $p|b$.
- Now we can prove that the factorization into irreducibles is essentially unique up to reordering.
  - There is one additional wrinkle that we must address, however, which we illustrate with an example.
  - In $\mathbb{C}[x]$, we can write $x^2 + 1 = (x + i)(x - i) = (ix + 1)(-ix + 1)$.
  - It would seem that these are two different factorizations, but we should really consider them the same, because all we have done is moved some units around: $x + i = i(-ix + 1)$ and $x - i = (-i)(ix + 1)$.
  - We should declare that two factorizations are equivalent if the only differences between them are by reordering terms or moving units around, which is equivalent to replacing elements with associates.
- Theorem (Unique Factorization in $F[x]$): Every polynomial of positive degree in $F[x]$ can be written as a product of irreducible polynomials. Furthermore, this factorization is unique up to associates: if $p = r_1 r_2 \cdots r_d = q_1 q_2 \cdots q_k$, then $d = k$ and there is some reordering of the factors such that p_i is associate to q_i for each $1 \leq i \leq k$.
  - Proof: We proved the existence of a factorization above. For the uniqueness, we induct on the number of irreducible factors of $p = r_1 r_2 \cdots r_d$.
  - If $d = 0$, then p is a unit. If p had some other factorization $p = rc$ with r irreducible, then q would divide a unit, hence be a unit (impossible).
  - Now suppose $d \geq 1$ and that $r = r_1 r_2 \cdots r_k = q_1 q_2 \cdots q_d$ has two factorizations into irreducibles.
  - Since $r_1 | (q_1 \cdots q_d)$ and r_1 is irreducible, repeatedly applying the fact that r_1 irreducible and $r_1 | ab$ implies $r_1 | a$ or $r_1 | b$ shows that r_1 must divide q_i for some i.
  - Then $q_i = r_1 u$ for some u: then since q_i is irreducible (and r_1 is not a unit), u must be a unit, so r_1 and q_i are associates.
  - Cancelling then yields the equation $r_2 \cdots r_d = (uq_2) \cdots q_k$, which is a product of fewer irreducibles. By the induction hypothesis, such a factorization is unique up to associates. This immediately yields the desired uniqueness result for p as well.
- As a final remark, we will note that if we drop the requirement that the coefficient ring in $R[x]$ be a field, then we may lose unique factorization. (We will explore this topic in more detail in a subsequent chapter.)
  - For example, if $R = 2\mathbb{Z}$, then two different factorizations of $12x$ into "irreducible" factors are $2 \cdot 6x$ and $6 \cdot 2x$. (None of these terms can be factored any further, and none of the terms are units.)
  - As another example, if $R = \mathbb{Z}/8\mathbb{Z}$, one can write $x^2 + 1 = (x + 1)(x + 7) = (x + 3)(x + 5)$.

2.2.4 Polynomial Functions, Roots of Polynomials

- In elementary algebra, polynomials are examples of functions. We would like to extend this idea of "plugging values in" to a general polynomial in $R[x]$.

- Definition: If R is a ring and $p = a_0 + a_1 x + \cdots + a_n x^n$ is an element of $R[x]$, for any $r \in R$ we define the value $p(r)$ to be the ring element $a_0 + a_1 r + \cdots + a_n r^n \in R$.
  - Example: If $p = 1 + x^2$ in $\mathbb{C}[x]$, then $p(1) = 1 + 1^2 = 2$, and $p(i) = 1 + i^2 = 0$.
  - Example: If $p = 1 + x^2$ in $\mathbb{F}_5[x]$, then $p(0) = 1$, $p(1) = 2$, $p(2) = 0$, $p(3) = 0$, and $p(4) = 2$.
  - In this way, we can view a polynomial $p \in R[x]$ as a function $p : R \to R$, with $p(r) = a_0 + a_1 r + \cdots + a_n r^n$.
  - We observe that the "traditional" polynomial notation $p(x)$ is somewhat ambiguous: we may be considering $p(x)$ as a ring element in $R[x]$ (in which case "x" represents an indeterminate), or we may be viewing it as a function from R to R (in which case "x" represents the variable of the function).

- Example: If $p = x + x^2$ in $\mathbb{F}_2[x]$, observe that $p(0) = p(1) = 0$.
  - Thus, although p is not the zero polynomial in $\mathbb{F}_2[x]$ (since it has degree 2), as a function from $\mathbb{F}_2$ to $\mathbb{F}_2$ it is the identically zero function!
  - More generally, if R is any finite commutative ring with elements $r_1, r_2, \ldots, r_n$, then the polynomial $p(x) = (x - r_1)(x - r_2) \cdots (x - r_n)$ is the identically zero function from R to R.
  - Thus, in general, we cannot always uniquely specify a polynomial $p \in R[x]$ by describing its behavior as a function $p : R \to R$.

- It is straightforward to see from the definition that if p and q are any polynomials in $R[x]$ and r is any element of R, then $(p + q)(r) = p(r) + q(r)$. However, if R is not commutative, it is not necessarily the case that $(pq)(r) = p(r)\,q(r)$.
  - The issue is that in $R[x]$, the indeterminate x commutes with the coefficients of the polynomials, while the ring element r may not.
  - Here is an explicit example: if $R = \mathbb{H}$ is the ring of real quaternions $a + bi + cj + dk$, let $p(x) = ix$ and $q(x) = j$, so that $pq(x) = (ix)j = kx$.
  - However, we can see that $p(i) = -1$ and $q(i) = j$, while $pq(i) = ki = j$, so it is *not* true that $pq(i) = p(i)q(i)$.
  - For this reason, we usually want to assume that R is commutative when working with polynomial functions in $R[x]$: otherwise, polynomial multiplication (and hence factorization) will not necessarily behave well.

- To begin our study of polynomial functions, we start with a pair of observations that are likely familiar from elementary algebra:

- Proposition (Remainder/Factor Theorem): Let F be a field. If $p \in F[x]$ is a polynomial and $r \in F$, then the remainder upon dividing $p(x)$ by $x - r$ is $p(r)$. In particular, $x - r$ divides $p(x)$ if and only if $p(r) = 0$. (In this case we say r is a zero or a root of $p(x)$.)
  - Proof: Suppose $p(x) = a_0 + a_1 x + \cdots + a_n x^n$. Observe first that $(x^k - r^k) = (x - r)(x^{k-1} + x^{k-2} r + \cdots + x r^{k-2} + r^{k-1})$, so in particular, $x - r$ divides $x^k - r^k$ for all k.
  - Now we simply write $p(x) - p(r) = \displaystyle\sum_{k=0}^{n} a_k (x^k - r^k)$, and since $x - r$ divides each term in the sum, it divides $p(x) - p(r)$.
  - Since $p(r)$ is a constant, it is therefore the remainder after dividing $p(x)$ by $x - r$. The other statement is immediate from the uniqueness of the remainder in the division algorithm.

- We can also bound the number of zeroes that a polynomial can have:

- Proposition: Let F be a field. If $p \in F[x]$ is a polynomial of degree d, then p has at most d distinct roots in F.
  - Proof: We induct on the degree d. For $d = 1$, the polynomial is of the form $a_0 + a_1 x$ for $a_1 \neq 0$, which has exactly one root, namely $-a_0/a_1$.
  - Now suppose the result holds for all polynomials of degree $\leq d$ and let p be a polynomial of degree $d+1$.
  - If p has no zeroes we are obviously done, so suppose otherwise and let $p(r) = 0$. We can then factor to write $p(x) = (x-r)q(x)$ for some polynomial $q(x)$ of degree d.
  - By the induction hypothesis, $q(x)$ has at most d roots: then $p(x)$ has at most $d+1$ roots, because $(a-r)q(a) = 0$ only when $a = r$ or $q(a) = 0$ (since F is a field).
- The above results, while seemingly obvious, can fail spectacularly if the coefficient ring is not a field. Here are some especially distressing examples:
  - The quadratic polynomial $q(x) = x^2 - 1$ visibly has four roots modulo 8, namely $x = 1, 3, 5, 7$. Furthermore, $q(x)$ can be factored in two different ways: as $(x-1)(x-7)$ and as $(x-3)(x-5)$.
  - The linear polynomial $q(x) = x$, despite having degree 1, is not irreducible modulo 6: it can be written as the product $(2x+3)(3x+2)$. Furthermore, $q(x) = x$ has one zero (namely $x = 0$), even though its two factors $2x+3$ and $3x+2$ each have no zeroes modulo 6.

2.2.5 Irreducibility Criteria

- In general, it is not easy to determine when an arbitrary polynomial is irreducible. If the degree is small, however, this task can be performed by examining all possible factorizations. The following result is frequently useful:
- Proposition: If F is a field and $p \in F[x]$ has degree 2 or 3 and has no zeroes in F, then p is irreducible.
  - Proof: If $p(x) = a(x)b(x)$, taking degrees shows $\deg(p) = \deg(a) + \deg(b)$. Since a and b both have positive degree and $\deg(p)$ is 2 or 3, at least one of a and b must have degree 1. Then its root is also a root of $p(x)$. Taking the contrapositive gives the desired statement.
  - Example: Over $\mathbb{R}$, the polynomial $x^2 + 2x + 11$ has no roots (it is always positive, as can be seen by completing the square), so it is irreducible.
  - Example: Over $\mathbb{F}_2$, the polynomial $q(x) = x^3 + x + 1$ is irreducible: it has no roots since $q(0) = q(1) = 1$.
  - Example: Over $\mathbb{F}_5$, the polynomial $q(x) = x^3 + x + 1$ is irreducible: it has no roots since $q(0) = 1$, $q(1) = 3$, $q(2) = 1$, $q(3) = 1$, and $q(4) = 4$.
  - Note of course that a polynomial of larger degree can be reducible without having any zeroes: for example, $x^4 + 3x^2 + 2$ has no zeroes in $\mathbb{R}$, but it is still reducible: $x^4 + 3x^2 + 2 = (x^2+1)(x^2+2)$.
- For polynomials of larger degree, determining irreducibility can be a much more difficult task. For certain particular fields, we can say more about the structure of the irreducible polynomials.
- Theorem (Fundamental Theorem of Algebra): Every polynomial of positive degree in $\mathbb{C}[x]$ has at least one root. Therefore, the irreducible polynomials in $\mathbb{C}[x]$ are precisely the polynomials of degree 1, and so every polynomial in $\mathbb{C}[x]$ factors into a product of degree-1 polynomials.
  - The first statement of this theorem is a standard result from analysis over the complex numbers, and we take it for granted.
  - To deduce the second statement from the first, observe that if $p(x)$ is any complex polynomial of degree larger than 1, then by assumption it has at least one root r in $\mathbb{C}$, so we can write $p(x) = (x-r)q(x)$ for some other polynomial $q(x)$: then p is reducible.
  - Therefore, the irreducible polynomials in $\mathbb{C}[x]$ are precisely the polynomials of degree 1. The final statement follows from the characterization of irreducible polynomials, because every polynomial is a product of irreducibles.

- By exploiting the relationship between the real and complex numbers, we can also characterize the irreducible polynomials in $\mathbb{R}[x]$.

- Corollary (Irreducible Polynomials in $\mathbb{R}[x]$): The irreducible polynomials in $\mathbb{R}[x]$ are the degree-1 polynomials along with the degree-2 polynomials of the form $ax^2 + bx + c$ with $b^2 - 4ac < 0$.

  - Proof: Both classes of the listed polynomials are irreducible (note that the quadratics have no roots in $\mathbb{R}$ by the quadratic formula). Furthermore, an irreducible quadratic polynomial must have no roots, hence be of the given form.
  - Now suppose $p(x) \in \mathbb{R}[x]$ is irreducible. By the Fundamental Theorem of Algebra, $p(x)$ has a root $z = a + bi \in \mathbb{C}$. If z is real, then we can factor $p(x) = (x - z)s(x)$ for some polynomial s, which necessarily must have degree 0: then p has degree 1.
  - Otherwise, suppose z is not real, and consider its complex conjugate $\overline{z} = a - bi$; note that $z \neq \overline{z}$ since z is nonreal.
  - It follows from the facts that $\overline{z + w} = \overline{z} + \overline{w}$ and $\overline{zw} = \overline{z} \cdot \overline{w}$ that if $q(x)$ is any polynomial with real coefficients, then $\overline{q(z)} = q(\overline{z})$.
  - Therefore, we must have $p(\overline{z}) = \overline{p(z)} = \overline{0} = 0$, meaning that $\overline{z}$ is also a root of p. Therefore, in $\mathbb{C}[x]$, we can write $p(x) = (x - z)(x - \overline{z}) \cdot s(x)$ for some polynomial s.
  - But now note that $(x - z)(x - \overline{z}) = x^2 - 2az + (a^2 + b^2)$ is in $\mathbb{R}[x]$, so (by an easy application of the division algorithm) it must be the case that $x^2 - 2az + (a^2 + b^2)$ divides $p(x)$ in $\mathbb{R}[x]$. Since $p(x)$ is irreducible, there cannot be any other positive-degree factors, so $\deg(p) = 2$. Since we have eliminated all higher-degree possibilities, we are done.

- It is more difficult to test whether a polynomial is irreducible in $\mathbb{Q}[x]$. A central idea is that we can reduce the problem of factoring in $\mathbb{Q}[x]$ to one of factoring in $\mathbb{Z}[x]$, the ring of polynomials with integer coefficients, by "clearing denominators".

  - Specifically, if p is any polynomial in $\mathbb{Q}[x]$, we may multiply p by the product of all the denominators of its coefficients (or their least common multiple) to obtain a polynomial in $\mathbb{Z}[x]$. Since every nonzero number is a unit in $\mathbb{Q}$, the factorization of this new polynomial, with integer coefficients, will be essentially the same as that of the original polynomial.
  - As an example, consider the problem of factoring $p(x) = 2x^3 + x^2 + \dfrac{2}{3}x + \dfrac{1}{3}$ in $\mathbb{Q}[x]$.
  - Since 3 is a unit in $\mathbb{Q}[x]$, we may equivalently ask about the factorization of $3p(x) = 6x^3 + 3x^2 + 2x + 1$ in $\mathbb{Z}[x]$.

- We start by proving the famous "rational root test", which allows us to determine whether a given polynomial in $\mathbb{Z}[x]$ has a rational root:

- Proposition (Rational Root Test): Suppose $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ is a polynomial in $\mathbb{Z}[x]$. Then any root r/s (in lowest terms) must have $r | a_0$ and $s | a_n$.

  - Proof: If r/s is a root of $p(x)$, then $a_n(r/s)^n + a_{n-1}(r/s)^{n-1} + \cdots + a_0 = 0$. Clearing denominators yields $a_n r^n + a_{n-1}r^{n-1}s + \cdots + a_1 r s^{n-1} + a_0 s^n = 0$.
  - Thus, by rearranging, we see that $a_n r^n = s(-a_{n-1}r^{n-1} - \cdots - a_0 s^{n-1})$, so s divides $a_n r^n$. But since s and r are relatively prime, this means s divides a_n.
  - In a similar way, since $a_0 s^n = r(-a_n r^{n-1} - \cdots - a_1 s^{n-1})$, we see that r divides $a_0 s^n$ hence a_0.

- This test allows us to make a finite list of possible rational roots for any polynomial with integer coefficients.

- Example: Show that the polynomial $p(x) = x^3 + ax + 1$ is irreducible in $\mathbb{Q}[x]$ for any integer $a \neq 0, -2$.

  - Since this polynomial has degree 3, we need only show that it has no roots in $\mathbb{Q}$.
  - By the rational root test, the only possible rational roots are ± 1, and since $p(1) = 2 + a$ and $p(-1) = a$, the conditions on a imply that p has no rational roots. Thus, p is irreducible.

- It seems natural to say that factorization of polynomials in $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$ are essentially "the same". However, we must be very careful with statements about factorizations when the coefficient ring is not a field.
  - For example, the polynomial $p(x) = 2x$ is irreducible in $\mathbb{Q}[x]$, since it has degree 1, but it is *not* irreducible in $\mathbb{Z}[x]$, since we may write $p(x) = 2 \cdot x$, and neither 2 nor x is a unit in $\mathbb{Z}[x]$.
  - The issue here is that 2 is a unit in $\mathbb{Q}$ but not in $\mathbb{Z}$, so that the factorization $2 \cdot x$ is a trivial factorization in $\mathbb{Q}[x]$, but is nontrivial in $\mathbb{Z}[x]$.
  - Ultimately, however, this is the only kind of thing that can go wrong:

- Theorem (Gauss's Lemma): If $p(x) \in \mathbb{Z}[x]$ has positive degree and is reducible in $\mathbb{Q}[x]$, then $p(x) = f(x)g(x)$ for some $f(x), g(x) \in \mathbb{Z}[x]$ of positive degree.
  - Proof: We say a polynomial in $\mathbb{Z}[x]$ is "primitive" if the gcd of its coefficients is equal to 1.
  - First, we observe that, in $\mathbb{Q}[x]$, any nonzero polynomial $a(x)$ is associate to a primitive polynomial in $\mathbb{Z}[x]$.
  - To see this, let d be the least common multiple of the denominators of $a(x)$: then $d \cdot a(x)$ is a polynomial in $\mathbb{Z}[x]$. Now let e be the greatest common divisor of the coefficients of $d \cdot a(x)$: then $\frac{d}{e} \cdot a(x)$ is a primitive polynomial in $\mathbb{Z}[x]$; since $\frac{d}{e}$ is a unit in $\mathbb{Q}$, this primitive polynomial is associate to $a(x)$.
  - Next, we claim that the product of two primitive polynomials is also primitive.
  - To see this, suppose that $a(x)b(x)$ is not primitive for some $a(x), b(x) \in \mathbb{Z}[x]$, with $a(x) = a_0 + a_1 x + \cdots + a_n x^n$ and $b(x) = b_0 + \cdots + b_m x^m$: then since $a(x)b(x)$ is not primitive, all of its coefficients are divisible by some prime s.
  - If there is at least one coefficient of each of $a(x)$ and $b(x)$ not divisible by s, suppose that a_i and b_j are the lowest-degree such coefficients. Then the degree-$(i+j)$ term of $a(x)b(x)$ is $a_0 b_{i+j} + \cdots + a_{i-1} b_{j+1} + a_i b_j + a_{i+1} b_{j-1} + \cdots + a_{i+j} b_0$, but by hypothesis each term except $a_i b_j$ is divisible by s. This is a contradiction, since this coefficient of $a(x)b(x)$ would then not be divisible by s.
  - Now, returning to the original problem, suppose that $p(x)$ is reducible in $\mathbb{Q}[x]$, say as $p(x) = f_0(x) g_0(x)$ with f_0 and g_0 both of positive degree.
  - By our first observation, both f_0 and g_0 are associate to a primitive polynomial: say, f, and g respectively.
  - Then (by rearranging the corresponding unit factors) we see that $d \cdot p(x) = e \cdot f(x) \cdot g(x)$ for some relatively prime integers d and e.
  - Now notice that since d and e are relatively prime, d must divide all coefficients of $f(x)g(x)$. But $f(x)g(x)$ is primitive by our second observation, so we must have $d = \pm 1$.
  - Then $p(x) = [ed^{-1} \cdot f(x)] \cdot g(x)$ is a nontrivial factorization of $p(x)$ over $\mathbb{Z}[x]$, as required.

- As we can see from the proof above, roughly speaking (up to shuffling around constant factors), factoring in $\mathbb{Q}[x]$ is the same as factoring in $\mathbb{Z}[x]$.
  - The advantage to working in $\mathbb{Z}[x]$, however, is that we can exploit properties of the integers to establish that factorizations cannot exist.

- Example: Show that the polynomial $p(x) = x^4 + x^3 - 2x^2 + x + 1$ is irreducible in $\mathbb{Q}[x]$.
  - First, by the rational root test, the only possible roots of this polynomial are ± 1, neither of which is a root.
  - Thus, if $p(x)$ were reducible, it would factor as a product of two quadratics. By moving factors of -1 around (as needed) such a factorization would have the form $p(x) = (x^2 + ax + b)(x^2 + cx + d)$.
  - By expanding and comparing coefficients, we see that $a + c = 1$, $b + ac + d = -2$, $ad + bc = 1$, and $bd = 1$.
  - The last equation gives $(b, d) = (1, 1)$ or $(-1, -1)$.
  - If $b = d = 1$ then we obtain the equations $a + c = 1$ and $ac = -4$, which has no integer solutions.
  - If $b = d = -1$ then we obtain $a + c = 1$, $ac = 0$, and $a + c = -1$, which has no solutions at all (integer or otherwise).

  - Therefore, $p(x)$ is irreducible, as claimed.

- Note, however, that a similar sort of analysis becomes very difficult in larger degree (and also becomes more difficult when the coefficients are large).

  - For example, to show in this manner that a polynomial of degree 7 is irreducible, one would need to verify that it has no roots, as well as no factorization into a product of polynomials of degree 2 and 5, or 3 and 4.

- For this reason, other irreducibility criteria have been developed. Here is one:

- Theorem (Eisenstein-Schönemann Criterion): Let $q(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ be a polynomial in $\mathbb{Z}[x]$. If each coefficient $a_0, a_1, \dots, a_{n-1}$ is divisible by a prime p, and a_0 is not divisible by p^2, then $q(x)$ is irreducible in $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$.

  - Proof: Suppose that $q(x) = b(x)c(x)$ were reducible in $\mathbb{Z}[x]$, with $b(x) = x^s + b_{s-1}x^{s-1} + \cdots + b_0$ and $c(x) = x^t + c_{t-1}x^{t-1} + \cdots + c_0$.
  - Since p divides $a_0 = b_0 c_0$, p divides at least one of these coefficients: without loss of generality, suppose $p|b_0$.
  - Now let b_i be the lowest-degree coefficient of $b(x)$ not divisible by p (there must be one, since $b_s = 1$ is not divisible by p): then we have $a_i = b_0 c_i + b_1 c_{i-1} + \cdots + b_{i-1}c_1 + b_i c_0$.
  - Since p divides a_i and also divides the terms $b_0 c_i$, $b_1 c_{i-1}$, ... , $b_{i-1}c_1$, it must divide $b_i c_0$. But since p does not divide b_i, we see that p divides c_0.
  - But then p divides both b_0 and c_0, meaning that p^2 divides $b_0 c_0 = a_0$. This is a contradiction, so there cannot exist any such factorization of $q(x)$.
  - Thus, $q(x)$ is irreducible in $\mathbb{Z}[x]$, and then by Gauss's lemma, q is irreducible in $\mathbb{Q}[x]$ as well.

- Example: By Eisenstein's criterion with $p = 2$, the polynomial $x^n - 2$ is irreducible in $\mathbb{Z}[x]$ for any positive integer n.

- Example: Show that the polynomial $q(x) = x^4 + x^3 - 3x^2 + x + 7$ is irreducible in $\mathbb{Q}[x]$.

  - We cannot apply Eisenstein's criterion to this polynomial directly.
  - However, notice that $q(x-1) = x^4 - 3x^3 + 6x + 3$, and this polynomial is irreducible by Eisenstein's criterion with $p = 3$.
  - It is then easy to see that any factorization of $q(x-1)$ would give a factorization of $q(x)$, and vice versa: therefore, the original polynomial $q(x)$ must also have been irreducible.

2.2.6 Polynomial Modular Congruences

- We now turn our attention to discussing modular congruences (and modular arithmetic) in $F[x]$.

- Our underlying definition of modular congruences and residue classes are exactly the same as over $\mathbb{Z}$:

- Definition: Let F be a field and $R = F[x]$. If $a, b, p \in R$, we say that a is congruent to b modulo p, written $a \equiv b \pmod{p}$, if $p|(b-a)$.

  - Example: In $\mathbb{R}[x]$, it is true that $x^2 \equiv x$ modulo $x - 1$, because $x - 1$ divides $x^2 - x = x(x-1)$.
  - Example: In $\mathbb{F}_2[x]$, it is true that $x^3 + x \equiv x + 1$ modulo $x^2 + x + 1$, because $(x^2 + x + 1)$ divides $(x^3 + x) - (x + 1) = (x + 1)(x^2 + x + 1)$.

- Most of the basic properties of modular congruences in $\mathbb{Z}$ extend to $F[x]$ with little or no change:

- Proposition (Modular Congruences): Let F be a field and $R = F[x]$. If $a, b, c, d, p \in R$ and $p \neq 0$, then the following are true:

  1. $a \equiv a \pmod{p}$.

2. $a \equiv b \pmod{p}$ if and only if $b \equiv a \pmod{p}$.
3. If $a \equiv b \pmod{p}$ and $b \equiv c \pmod{p}$, then $a \equiv c \pmod{p}$.
4. If $a \equiv b \pmod{p}$ and $c \equiv d \pmod{p}$, then $a + c \equiv b + d \pmod{p}$.
5. If $a \equiv b \pmod{p}$ and $c \equiv d \pmod{p}$, then $ac \equiv bd \pmod{p}$.
6. If $a \equiv b \pmod{p}$, then $ac \equiv bc \pmod{pc}$ for any nonzero c.
7. If $a \equiv b \pmod{p}$ then $a^k \equiv b^k \pmod{p}$ for any positive integer k.
8. If $d|p$, then $a \equiv b \pmod{p}$ implies $a \equiv b \pmod{d}$.
    - We leave the proofs as exercises, as they are all identical to the corresponding proofs in $\mathbb{Z}$.

- We can now construct residue classes, again in exactly the same way:

- Definition: If $a, r \in F[x]$, the residue class of a modulo r, denoted $\overline{a}$, is the set $S = \{a + dr : d \in F[x]\}$ of all elements in $F[x]$ congruent to a modulo r.
    - Example: The residue class of 1 modulo x in $\mathbb{F}_2[x]$ is $\{1,\ 1 + x,\ 1 + x^2,\ 1 + x + x^2,\ 1 + x^3, \dots\}$.

- Here are a few fundamental properties of residue classes:

- Proposition (Properties of Residue Classes): Let $R = F[x]$ for F a field, and suppose p is a nonzero polynomial in $F[x]$. Then

1. If a and b are polynomials in $F[x]$, then $a \equiv b \pmod{p}$ if and only if $\overline{a} = \overline{b}$.
    - Proof: Identical to the proof over $\mathbb{Z}$.
2. Two residue classes modulo p are either disjoint or identical.
    - Proof: Identical to the proof over $\mathbb{Z}$.
3. The residue classes modulo p are precisely those of the form $\overline{r}$ where $\deg(r) < \deg(p)$.
    - Proof: By the division algorithm, for any polynomial a there exists a unique r with $\deg(r) < \deg(p)$ such that $a = qm + r$ with $q \in F[x]$.
    - Then $a \equiv r \pmod{p}$, and so every polynomial is congruent modulo p to precisely one polynomial r with $\deg(r) < \deg(p)$.
    - In other words, every polynomial is contained in exactly one of the residue classes $\overline{r}$ where $\deg(r) < \deg(p)$.
    - By property (2), we conclude that these are all the residue classes, and that they are disjoint.

- If F is an infinite field, then if $\deg(p) > 0$, there will always be infinitely many residue classes in $F[x]$ modulo $p(x)$.
    - However, when F is a finite field of cardinality $|F|$, then the residue classes are each represented by a unique polynomial in $F[x]$ of degree less than $\deg(p)$.
    - Such a polynomial has exactly $\deg(p)$ coefficients (for the terms of degree 0, 1, ... , $\deg(p) - 1$), and each coefficient has $|F|$ possible choices: thus, there are precisely $|F|^{\deg(p)}$ residue classes modulo $p(x)$.

- Example: List the residue classes in $\mathbb{F}_2[x]$ modulo x^2.
    - Each coefficient is either 0 or 1, and by the above result, the residue classes are precisely the polynomials of degree less than 2.
    - Thus, there are four residue classes in $\mathbb{F}_2[x]$ modulo x^2: $\overline{0}$, $\overline{1}$, $\overline{x}$, and $\overline{x+1}$.

- Definition: If F is a field and $R = F[x]$ with $p \in R$ nonzero, the set of residue classes modulo p is denoted R/pR (read as "R modulo pR").

- Now we can show that the residue classes in R/pR form a ring:

- Theorem (Modular Arithmetic in $F[x]$): Let $R = F[x]$ where F is a field, and let $p \in F[x]$ be nonzero. Then the residue classes in R/pR form a commutative ring with 1, under the operations $\overline{a} + \overline{b} = \overline{a+b}$ and $\overline{a} \cdot \overline{b} = \overline{ab}$.

- The proof of this fact is exactly the same as over $\mathbb{Z}$.
- Proof: First we must show that the addition and multiplication operations are well-defined: that is, if we choose different elements $a' \in \bar{a}$ and $b' \in \bar{b}$, the residue class of $a' + b'$ is the same as that of $a + b$, and similarly for the product.
- To see this, if $a' \in \bar{a}$ then $a' \equiv a \pmod{p}$, and similarly if $b' \in \bar{b}$ then $b' \equiv b \pmod{p}$.
- Then $a' + b' \equiv a + b \pmod{p}$, so $\overline{a' + b'} = \overline{a + b}$. Likewise, $a'b' \equiv ab \pmod{p}$, so $\overline{a'b'} = \overline{ab}$.
- Thus, the operations are well-defined.
- For the ring axioms [R1]-[R8], we observe that associativity, commutativity, and the distributive law follow immediately from the corresponding properties in $R = F[x]$. The additive identity in R/pR is $\bar{0}$, the additive inverse of $\bar{a}$ is $\overline{-a}$, and the multiplicative identity is $\bar{1}$.

- Like in $\mathbb{Z}$, we can (and will!) abuse notation and drop the bar notation, with the understanding that all of our calculations are to be considered "modulo p".

- When F is an infinite field, there will be infinitely many residue classes in R/pR, so we cannot sensibly write down addition and multiplication tables. However, we can certainly construct such tables when F is finite.

- Example: With $R = \mathbb{F}_2[x]$, here are the addition and multiplication tables for R/pR with $p = x^2$:

$+$	0	1	x	$x+1$
0	0	1	x	$x+1$
1	1	0	$x+1$	x
x	x	$x+1$	0	1
$x+1$	$x+1$	x	1	0

$\cdot$	0	1	x	$x+1$
0	0	0	0	0
1	0	1	x	$x+1$
x	0	x	0	x
$x+1$	0	$x+1$	x	1

- Notice that this ring has a zero divisor (namely x), and that the elements 1 and $x + 1$ are units. Notice that $p(x) = x^2$ is reducible in R, since it has the factorization $x^2 = x \cdot x$.

- Example: With $R = \mathbb{F}_2[x]$, here are the addition and multiplication tables for R/pR with $p = x^2 + x + 1$:

$+$	0	1	x	$x+1$
0	0	1	x	$x+1$
1	1	0	$x+1$	x
x	x	$x+1$	0	1
$x+1$	$x+1$	x	1	0

$\cdot$	0	1	x	$x+1$
0	0	0	0	0
1	0	1	x	$x+1$
x	0	x	$x+1$	1
$x+1$	0	$x+1$	1	x

- Notice that this ring is a field, since every nonzero residue class is a unit. Observe also that the polynomial $p(x) = x^2 + x + 1$ is irreducible in R, since it has no roots.

- Example: With $R = \mathbb{F}_3[x]$, here is the multiplication table for R/pR with $p = x^2 + 1$:

$\cdot$	0	1	2	x	$x+1$	$x+2$	$2x$	$2x+1$	$2x+2$
0	0	0	0	0	0	0	0	0	0
1	0	1	2	x	$x+1$	$x+2$	$2x$	$2x+1$	$2x+2$
2	0	2	1	$2x$	$2x+2$	$2x+1$	x	$x+2$	$x+1$
x	0	x	$2x$	2	$x+2$	$2x+2$	1	$x+1$	$2x+1$
$x+1$	0	$x+1$	$2x+2$	$x+2$	$2x$	1	$2x+1$	2	x
$x+2$	0	$x+2$	$2x+1$	$2x+2$	1	x	$x+1$	$2x$	2
$2x$	0	$2x$	x	1	$2x+1$	$x+1$	$2x+2$	$2x+1$	$d+2$
$2x+1$	0	$2x+1$	$x+2$	$x+1$	2	$2x$	$2x+2$	x	1
$2x+2$	0	$2x+2$	$x+1$	$2x+1$	x	2	$x+2$	1	$2x$

- Notice that this ring is a field, since every nonzero residue class is a unit. Observe also that the polynomial $p(x) = x^2 + 1$ is irreducible in R, since it has no roots.

- As suggested by the examples above (and by the analogy between primes in $\mathbb{Z}$ and irreducible polynomials in $F[x]$), we can characterize the units and zero divisors in R/pR based on whether they have a nontrivial common divisor with p:

- Theorem (Units and Zero Divisors in R/pR): Let F be a field and $R = F[x]$, with $p \in R$ nonzero. A nonzero residue class $\overline{r}$ in R/pR is a unit if and only if r and p are relatively prime, and is a zero divisor if and only if r and p have a common divisor of positive degree.
  - Proof: Let $d = \gcd(r, p)$; observe that this is well-defined on residue classes because $\gcd(r, p) = \gcd(r', p)$ for any $r' \in \overline{r}$.
  - If $d = 1$, then by the Euclidean algorithm, we can write $1 = c_r r + c_p p$ for some polynomials c_r, c_p. Then $\overline{c_r} \cdot \overline{r} = \overline{1}$, meaning that r is a unit in R/pR.
  - If d has positive degree, let $p = ds$ for some polynomial s with $0 < \deg(s) < \deg(p)$: then $\overline{s} \neq \overline{0}$ by the condition on the degree.
  - If $r = dt$ for some polynomial t, we then have $\overline{r} \cdot \overline{s} = \overline{t(ds)} = \overline{t} \cdot \overline{ds} = \overline{t} \cdot \overline{0} = \overline{0}$. Therefore, since $\overline{s} \neq \overline{0}$, we see that $\overline{r}$ is a zero divisor in R/pR.
- Per the proof of the theorem above, we can use the Euclidean algorithm in $F[x]$ to compute the multiplicative inverse of a unit:
- Example: For $R = \mathbb{F}_5[x]$, find the multiplicative inverse of $x^2 + 2$ modulo $x^3 + 1$.
  - First we apply the Euclidean algorithm in R:

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

    and so the gcd of $x^2 + 2$ and $x^3 + 1$ is 1.
  - By back-solving, we obtain

$$\begin{aligned} 3x + 1 &= (x^3 + 1) - x \cdot (x^2 + 2) \\ 1 &= (x^2 + 2) - (2x + 1)(3x + 1) = (2x^2 + x + 1)(x^2 + 2) - (2x + 1)(x^3 + 1) \end{aligned}$$

    and thus by reducing modulo $x^3 + 1$, we see that the multiplicative inverse of $x^2 + 2$ is $\boxed{2x^2 + x + 1}$.
- In analogy with the fact that $\mathbb{Z}/m\mathbb{Z}$ is a field precisely when m is prime, we also see that R/pR is a field precisely when p is irreducible:
- Corollary: Let F be a field and $R = F[x]$, with $p \in R$ have positive degree. Then R/pR is a field if and only if p is irreducible.
  - Proof: By the previous theorem, we see that if p is irreducible then every nonzero residue class modulo p is a unit. Furthermore, if $\deg(p) > 0$, then $\overline{1} \neq \overline{0}$, so R/pR is a field.
  - Inversely, if p is reducible, then (again as above) there are zero divisors in R/pR.
- By finding irreducible polynomials in $\mathbb{F}_p[x]$, we can use the corollary above to construct finite fields.
- Example: Construct a finite field with 27 elements.
  - Since $27 = 3^3$, we can construct a finite field with 27 elements as R/pR, where $R = \mathbb{F}_3[x]$ and p is an irreducible polynomial of degree 3.
  - One possible choice is the polynomial $p(x) = x^3 + 2x + 1$: it has no roots, since $p(0) = p(1) = p(2) = 1$, so it is irreducible.
  - Therefore, the ring R/pR is a field with $3^3 = 27$ elements, as required.

Well, you're at the end of my handout. Hope it was helpful.