Number Theory (part 1): The Integers (by Evan Dummit, 2020, v. 2.05)

Contents

1 The Integers

One of the most foundational objects in mathematics is the integers, as they are used the basis and reference point for many other topics in mathematics. Our goal in this chapter is to define the integers axiomatically and to develop some basic properties of primes and divisibility. We then introduce the general notion of a commutative ring with 1 along with some other number systems, in order to contrast their properties with those of the integers.

1.1 The Integers, Axiomatically

- We are all at least a little bit familiar with the integers $\mathbb{Z}$, consisting of the positive integers $\mathbb{Z}_+$ (1, 2, 3, 4, ...), along with their negatives (−1, −2, −3, −4, ...) and zero (0).
    - There are two natural binary arithmetic operations defined on the integers, namely addition (+) and multiplication (·), along with the unary operation of negation (−).
    - But it is not so easy to prove things about the integers without a solid set of properties to work from.

1.1.1 Definition of the Integers

- "Definition": The integers are a set $\mathbb{Z}$ along with two (closed) binary[1] operations + and ·, obeying the following properties[2]:

[I1] The operation + is associative: $a + (b + c) = (a + b) + c$ for any integers a, b, c.

[1] The definition of a binary operation means that for any two integers a and b, the symbols $a + b$ and $a \cdot b$ are always defined and are integers. Some authors list these properties explicitly as part of their list of axioms.

[2] To be a proper definition, we would also need to establish that there actually is a set with operations obeying these properties, which turns out to be rather difficult. But there are various constructions for $\mathbb{Z}$ using set theory, which we will not detail here.

[I2] The operation $+$ is commutative: $a + b = b + a$ for any integers a, b.

[I3] There is an additive identity 0 satisfying $a + 0 = a$ for all integers a.

[I4] Every integer a has an additive inverse $-a$ satisfying $(-a) + a = 0$.

[I5] The operation $\cdot$ is associative: $a \cdot (b \cdot c) = (a \cdot b) \cdot c$ for any integers a, b, c.

[I6] The operation $\cdot$ is commutative: $a \cdot b = b \cdot a$ for any integers a, b.

[I7] There is a multiplicative identity $1 \neq 0$ satisfying $1 \cdot a = a$ for all integers a.

[I8] The operation $\cdot$ distributes over $+$: $a \cdot (b + c) = a \cdot b + a \cdot c$ for any integers a, b, c.

Furthermore, there is a subset of $\mathbb{Z}$, namely the positive integers $\mathbb{Z}_+$, such that

[N1] For every $a \in \mathbb{Z}$, precisely one of the following holds: $a \in \mathbb{Z}_+$, $a = 0$, or $(-a) \in \mathbb{Z}_+$.

[N2] The set $\mathbb{Z}_+$ is closed under $+$ and $\cdot$: for any $a, b \in \mathbb{Z}_+$, both $a + b$ and $a \cdot b$ are in $\mathbb{Z}_+$.

[N3] Every nonempty subset S of $\mathbb{Z}_+$ contains a smallest element: that is, an element $x \in S$ such that if $y \in S$, then either $y = x$ or $y - x \in \mathbb{Z}_+$.

- Remark: The axiom [N3] is called the well-ordering axiom. It is the axiom that differentiates the integers from other number systems such as the rational numbers or the real numbers, both of which obey all of the other axioms.

1.1.2 Basic Arithmetic

- Using the axioms for $\mathbb{Z}$, we can establish all of the properties of basic arithmetic. Doing this is not especially difficult once the basic idea is identified (namely, invoking the axioms judiciously, along with some case analysis). Here are some examples:
- Proposition (Basic Arithmetic): Inside the integers $\mathbb{Z}$, the following properties hold:

1. The additive and multiplicative identities are unique.
   - Proof: Suppose we had two additive identities 0_A and 0_B. Then by axioms [I2] and [I3], we may write $0_A = 0_A + 0_B = 0_B + 0_A = 0_B$, and therefore $0_A = 0_B$.
   - In a similar way, if we had two multiplicative identities 1_A and 1_B, then by axioms [I6] and [I7], we may write $1_A = 1_A \cdot 1_B = 1_B \cdot 1_A = 1_B$, and therefore $1_A = 1_B$.
2. Addition possesses a cancellation law: if $a + b = a + c$, then $b = c$.
   - Proof: By axioms [I1], [I3], and [I4], we have $b = 0 + b = [(-a) + a] + b = (-a) + (a + b) = (-a) + (a + c) = [(-a) + a] + c = 0 + c = c$.
3. Additive inverses are unique.
   - Proof: Suppose a had two additive inverses b and c. Then we would have $a + b = 0 = a + c$ by [I2] and [I4], and therefore by the cancellation law (2) we would have $b = c$.
4. For all $a \in \mathbb{Z}$, $0 \cdot a = 0$, $(-1) \cdot a = -a$, and $-(-a) = a$.
   - Proof: For any element a, by [I3] and [I8] we have $0 \cdot a + 0 = 0 \cdot a = (0 + 0) \cdot a = 0 \cdot a + 0 \cdot a$. Then by the cancellation law (2), we obtain $0 = 0 \cdot a$.
   - For the second statement, by the above along with [I3], [I7], and [I8] we have $0 = 0 \cdot a = [1 + (-1)] \cdot a = 1 \cdot a + (-1) \cdot a = a + (-1) \cdot a$. Then by the uniqueness of additive inverses (3), we see $(-1) \cdot a = -a$.
   - For the last statement, observe that by definition, $-(-a)$ is the element which when added to $-a$ yields 0. But since $a + (-a) = 0 = (-a) + a$ by definition and [I2], by the uniqueness of the additive inverse (3) we conclude $-(-a) = a$.
5. For any a and b, $-(a + b) = (-a) + (-b)$, $(-a) \cdot b = -(a \cdot b) = a \cdot (-b)$, and $(-a) \cdot (-b) = a \cdot b$.
   - Proof: For the first statement, observe that by [I1]-[I4], we have $[a + b] + [(-a) + (-b)] = [a + [b + (-b)]] + (-a) = (a + 0) + (-a) = a + (-a) = 0$, and so by the uniqueness of the additive inverse (3) we see $-(a + b) = (-a) + (-b)$.
   - For the second statement, by (4) and [I6] we have $(-a) \cdot b = [(-1) \cdot a] \cdot b = (-1) \cdot (a \cdot b) = -(a \cdot b)$. The other part of the second statement and the third statement follow by essentially the same argument.

6. For any a and b, $b \cdot a = a + (b-1) \cdot a$. Thus, $2 \cdot a = a + a$, $3 \cdot a = a + (a + a)$, and so forth.
   - Proof: We have $b \cdot a = [1 + (b-1)] \cdot a = 1 \cdot a + (b-1) \cdot a = a + (b-1) \cdot a$ by [I7] and [I8].
   - The second statement follows from this along with the observation that $1 \cdot a = a$.
7. The multiplicative identity $1 \in \mathbb{Z}_+$.
   - Proof: By [N1], one of the following things holds: either $1 \in \mathbb{Z}_+$ (in which case we are done), or $1 = 0$ (this is impossible because by [I7], $1 \neq 0$), or $-1 \in \mathbb{Z}_+$.
   - If $-1 \in \mathbb{Z}_+$, then by [N2], we would see that $(-1) \cdot (-1) \in \mathbb{Z}_+$, and by (6) we have $(-1) \cdot (-1) = 1$, so we would have $1 \in \mathbb{Z}_+$. In all cases $1 \in \mathbb{Z}_+$ so we are done.
8. If $ab = 0$, then $a = 0$ or $b = 0$.
   - Proof: If $a, b \in \mathbb{Z}_+$ then $ab \in \mathbb{Z}_+$ and so $ab \neq 0$. If $a, -b \in \mathbb{Z}_+$ or $-a, b \in \mathbb{Z}_+$ then $-(ab) \in \mathbb{Z}_+$ by (5) and (4), and if $-a, -b \in \mathbb{Z}_+$ then $ab \in \mathbb{Z}_+$ also by (5) and (4).
   - Thus, the only case in which $ab = 0$ is the case where $a = 0$ or $b = 0$, as claimed.

- It is quite tedious to write every proof using only properties of the axioms, so from this point forward we will revert to using more standard notation and language.
  - However, it is worthwhile noting that we could (if we wanted to) always reduce every proof down to a series of statements each of which is an application of one of the axioms.
  - From this viewpoint, our intermediate results (our propositions, lemmas, theorems, and so forth) consist of a sequence of applications of the axioms that we can invoke in any situation where the hypotheses apply, and that yield the claimed result.
  - In this way, we can "build up" from the axiomatic foundation, by first proving very basic properties, and then using those results to prove more complicated properties, and so forth, until we have established substantial results.
  - As a matter of course, most mathematicians do not dwell much on foundational questions, and instead take for granted all of the basic properties of numbers and arithmetic that we will examine closely.
  - But, at least in principle, every mathematical proof can be reduced down to a sequence of axiomatic calculations. This idea is actually the foundation of automated theorem provers, which are computer programs that can construct and verify mathematical proofs down to the axiomatic level.
  - We cluster these statements together to make them more readable and (vastly!) more understandable to human readers.

- We can also define some other basic arithmetic properties of the integers:

- Definition: We can define the binary operation of subtraction in terms of addition and negation by setting $a - b = a + (-b)$.
  - Notice that this operation is well-defined (i.e., the definition makes sense and there is no ambiguity), because $-b$ is unique as we showed above.

- Definition: We define the order relation $<$ (less than) by saying $a < b$ if and only if $b - a \in \mathbb{Z}_+$. We also define $b > a$ (greater than) to mean the same thing, and likewise write $a \leq b$ to mean $a < b$ or $a = b$, and $a \geq b$ to mean $a > b$ or $a = b$.
  - The axioms [N1] and [N2] ensure that these symbols all behave in the way we expect inequality symbols to behave.
  - Explicitly, [N1] implies that for any integers a and b, exactly one of $a < b$, $a = b$, or $b < a$ holds, because the integer $b - a$ is either positive, zero, or negative (respectively).
  - Also, [N2] implies that for any a, b, c with $a < b$ and $b < c$, then $a < c$, because if $b - a$ and $c - b$ are positive, then their sum $(c - b) + (b - a) = c - a$ is also positive.
  - Finally, [N2] also implies that for any a, b, c with $a < b$ and $0 < c$, then $ac < bc$, since $b - a$ and $c - 0 = c$ are positive and thus have positive product.

- A seemingly obvious, yet bizarrely important, property of the integers is the following result:
- Proposition: There are no integers between 0 and 1.
  - Observe that this proposition must rely on the well-ordering axiom, because all of the other axioms also apply to the rational and real numbers (which certainly do have elements between 0 and 1).
  - Proof: Let $S = \{r \in \mathbb{Z} : 0 < r < 1\}$ be the set of all integers between 0 and 1. If S is empty, we are done, so assume $S \neq \emptyset$.
  - By the well-ordering axiom [N3], S has a minimal element r.
  - Now observe that since $0 < r < 1$, we have $0 < r^2 < r < 1$ by appropriate uses of [N1] and [N2].
  - But this is a contradiction, because r^2 is then a positive integer less than r, but r was assumed to be minimal.
  - Therefore, S cannot be nonempty, so $S = \emptyset$ as claimed.

1.1.3 Induction

- By using the well-ordering axiom, we can establish the validity of "proof by induction":
- Proposition (Proof by Induction): If S is a set of positive integers such that $1 \in S$, and $n \in S$ implies $(n+1) \in S$, then $S = \mathbb{Z}_+$ is the set of all positive integers.
  - Proof: Let $T = \mathbb{Z}_+ \backslash S$, the positive integers not in S. If T is empty, we are done, so assume $T \neq \emptyset$.
  - By the well-ordering axiom [N3], T has a minimal element r.
  - Since r is positive, there are three possibilities: $0 < r < 1$, $r = 1$, or $1 < r$.
  - Since there are no positive integers between 0 and 1, we cannot have $0 < r < 1$.
  - Furthermore, since $1 \in S$, we cannot have $r = 1$.
  - The only remaining possibility is that $1 < r$. But then $0 < r - 1$, so $r - 1$ is a positive integer.
  - Since $r - 1 < r$ and r is the minimal element of T, we see that $r - 1 \in S$.
  - But then the hypotheses on S imply $r \in S$, which is a contradiction since we assumed $r \in T$.
  - Hence $T = \emptyset$, so $S = \mathbb{Z}_+$ as claimed.
- Now we can invoke the result of the proposition to give a concrete procedure for mathematical induction.
  - Explicitly, suppose $P(n)$ is a proposition such that the "base case" $P(1)$ holds, and also such that the "inductive step" holds: namely, $P(n)$ implies $P(n+1)$ for all $n \geq 1$.
  - Then we claim that $P(k)$ is true for every positive integer k.
  - To show this fact, let S be the set of positive integers k such that $P(k)$ is true.
  - By hypothesis, $1 \in S$, and $n \in S$ implies $(n+1) \in S$.
  - Therefore, by our proposition, we conclude that S is the set of all positive integers, which is to say, $P(k)$ is true for all positive integers k.
- Example: Prove that $2^0 + 2^1 + 2^2 + \cdots + 2^{n-1} = 2^n - 1$ for every positive integer n.
  - We prove this by induction on n.
  - For the base case $n = 1$, we must show that $2^0 = 2^1 - 1$ which is clearly true.
  - For the inductive step, we are given that $2^0 + 2^1 + 2^2 + \cdots + 2^{n-1} = 2^n - 1$ and must show that $2^0 + 2^1 + 2^2 + \cdots + 2^n = 2^{n+1} - 1$.
  - By the inductive hypothesis, we can write

$$\begin{aligned} 2^0 + 2^1 + 2^2 + \cdots + 2^n &= [2^0 + 2^1 + 2^2 + \cdots + 2^{n-1}] + 2^n \\ &= [2^n - 1] + 2^n = 2^{n+1} - 1 \end{aligned}$$

  and therefore we see $2^0 + 2^1 + 2^2 + \cdots + 2^n = 2^{n+1} - 1$, as required.

- By induction, $2^0 + 2^1 + 2^2 + \cdots + 2^{n-1} = 2^n - 1$ for every positive integer n.

- Example: Prove that $1 + 3 + 5 + \cdots + (2n-1) = n^2$ for every positive integer n.
  - We prove this by induction on n.
  - For the base case $n = 1$, we must show that $1 = 1$ which is clearly true.
  - For the inductive step, we are given that $1 + 3 + 5 + \cdots + (2n-1) = n^2$ and must show that $1 + 3 + 5 + \cdots + (2n-1) + (2n+1) = (n+1)^2$.
  - By the inductive hypothesis, we can write
$$\begin{aligned} 1 + 3 + 5 + \cdots + (2n-1) + (2n+1) &= [1 + 3 + 5 + \cdots + (2n-1)] + (2n+1) \\ &= n^2 + 2n + 1 = (n+1)^2 \end{aligned}$$
    and therefore we see $1 + 3 + 5 + \cdots + (2n-1) + (2n+1) = (n+1)^2$, as required.
  - By induction, $1 + 3 + 5 + \cdots + (2n-1) = n^2$ for every positive integer n.

- Another flavor of induction is called "complete induction" or "strong induction": rather than assuming the immediately previous case, we assume *all* of the previous cases: the inductive step is now that $P(1)$, $P(2)$, ..., $P(n)$ collectively imply $P(n+1)$.
  - It may seem like we are assuming extra information, but in fact strong induction and regular induction are logically equivalent.
  - The reason is that we can view any strong-induction proof as a regular-induction proof with a slightly different hypothesis.
  - Explicitly, if we define $Q(n)$ to be the proposition that all of the statements $P(1)$, $P(2)$, ... , $P(n)$ are true, then it is not hard to see that a strong-induction proof of the proposition $P(n)$ is the same as a standard-induction proof of the proposition $Q(n)$.
  - Thus, it is always allowable to assume the strong induction hypothesis when writing an induction proof (although in practice, one typically only does so when it is actually necessary).

- Example: Prove that every positive integer can be written as the sum of one or more distinct powers of 2.
  - We will show this by (strong) induction on the integer, n.
  - We take the base case $n = 1$: clearly, $n = 2^0 = 1$ has the required property, as claimed.
  - For the inductive step, suppose that $n \geq 2$ and the result holds for any positive integer less than n.
  - If n is even, then $n/2$ is a positive integer with $n/2 < n$, so by the inductive hypothesis, $n/2$ can be written as the sum of one or more distinct powers of 2, say, $n/2 = 2^{a_1} + \cdots + 2^{a_d}$.
  - Then doubling all of the terms in this sum yields $n = 2^{a_1+1} + \cdots + 2^{a_d+1}$ so n is also the sum of distinct powers of 2, as required.
  - If n is odd, then $(n-1)/2$ is a positive integer with $(n-1)/2 < n$, so by the inductive hypothesis, $(n-1)/2$ can be written as the sum of one or more distinct powers of 2, say, $(n-1)/2 = 2^{a_1} + \cdots + 2^{a_d}$.
  - Then doubling all of the terms and adding 1 yields $n = 2^0 + 2^{a_1+1} + \cdots + 2^{a_d+1}$ so n is also the sum of distinct powers of 2, as required.

- As a final remark, we note that it is also possible to phrase induction arguments as "smallest counterexample" or "infinite descent" arguments.
  - The general idea is to work to show that $P(n)$ is true for all positive integers n by contradiction.
  - If $P(n)$ is not true for all positive integers n, then by the well-ordering axiom there must exist a minimal positive integer k such that $P(k)$ is false: this would be a "minimal counterexample".
  - If one can then prove that the existence of such a counterexample would imply the existence of a smaller counterexample (i.e., some smaller positive integer k' such that $P(k')$ is false), this would yield a contradiction.
  - Notice of course that the structure of this argument is equivalent to proof by induction (since both arguments invoke the well-ordering axiom as a way of showing that a set is equal to $\mathbb{Z}_+$).
  - In certain cases it can be easier to identify salient features of the induction argument by phrasing the problem in terms of smallest counterexamples. However, standard proof by induction tends to be more straightforward since it is a direct proof rather than a proof by contradiction.

1.2 Divisibility and the Euclidean Algorithm

- We have constructed three of the operations of standard arithmetic: $+$, $-$, and $\cdot$. We now discuss division.
  - One caveat with division is that, unlike addition, subtraction, and multiplication, the quotient of one integer by another (even if it is defined) need not be an integer.
  - Thus, instead of discussing division, we start by discussing divisibility

1.2.1 Divisibility and Division With Remainder

- Definition: If $a \neq 0$, we say that a divides b, written $a|b$, if there exists an integer k with $b = ka$. If $a|b$, we also say that b is divisible by a.
  - Examples: $2|4$ since $4 = 2 \cdot 2$, $(-7)|7$ since $7 = (-1) \cdot (-7)$, $13|1001$ since $1001 = 77 \cdot 13$, $6|0$ since $0 = 0 \cdot 6$, and $0|0$ since $0 = 2019 \cdot 0$.
  - If a does not divide b, we sometimes write $a \nmid b$. For example, $2 \nmid 3$ since there is no integer k with $3 = 2k$.
  - In the particular case of divisibility by 2, we say n is even if $2|n$. We will show (carefully) later that $2 \nmid n$ is equivalent to saying that $2|(n-1)$, which we take as the definition of odd.
- There are a number of basic properties of divisibility that follow from the definition and properties of arithmetic:
- Proposition (Properties of Divisibility): For any integers a, b, c, m, x, y, the following hold:
  1. If $a|b$, then $a|bc$ for any c.
  2. If $a|b$ and $b|c$, then $a|c$.
  3. If $a|b$ and $a|c$, then $a|(xb + yc)$ for any x and y.
  4. If $a|b$ and $b|a$, then $a = \pm b$.
  5. If $a|b$, and $a, b > 0$, then $a \leq b$.
  6. For any $m \neq 0$, $a|b$ is equivalent to $(ma)|(mb)$.
     - Proof: Each of these follows essentially directly from the definition of divisibility and the basic properties of arithmetic.
     - For example, (2) follows because $a|b$ and $b|c$ imply that there exist integers k and l such that $b = ka$ and $c = lb$, and thus $c = lb = (lk)a$: hence c is an integer times a, so $a|c$.
     - Likewise, (5) follows because if $a|b$ and a, b are positive, then $b = ka$ for some positive integer k. Since this means $1 \leq k$ because there are no integers between 0 and 1, we have $a \leq ka = b$, and so $a \leq b$.
- If $0 < b < a$ and b does not divide a, we can still attempt to divide a by b to obtain a quotient and remainder: this is a less-explicit version of the long-division algorithm familiar from elementary school. Formally:
- Theorem (Division With Remainder): If a is any integer and b is a positive integer, then there exist unique integers q and r such that $a = qb + r$ with $0 \leq r < b$. Furthermore, $r = 0$ if and only if $b|a$.
  - Proof: The second statement follows immediately from the first statement: if $r = 0$ then $a = qb$ so $b|a$, and if $b|a$ then $a = kb$ for some k; then the uniqueness of q and r implies that we must have $q = k$ and $r = 0$.
  - To show existence of q and r, let T be the intersection of the set $S = \{a + kb : k \in \mathbb{Z}\}$ with the nonnegative integers. If $a \geq 0$ then $a \in S$, while if $a < 0$ then $a - ab = a(1 - b) \in S$, so in either case T is nonempty.
  - Let r be the minimal element of T: then $0 \leq r$, and since $r - b$ is not in T by minimality, we also have $r < b$. But since $r \in S$, we must have $r = a - qb$ for some integer q. Thus, $a = qb + r$ where $0 \leq r < b$, as claimed.

- For uniqueness, suppose $qb + r = a = q'b + r'$ with $0 \leq r, r' < b$. Then $-b < r - r' < b$, but we can write $r - r' = b(q' - q)$, so dividing through by b yields $-1 < q' - q < 1$. But since $q' - q$ is an integer and there are no integers between 0 and 1 (or -1 and 0), it must be the case that $q' = q$ and $r' = r$.

- Example: Find the quotient and remainder upon dividing the number $a = 25$ by $b = 4$.
  - We compute the sets $S = \{\dots, -7, -3, 1, 5, \dots, 21, 25, 29, 33, \dots\}$, and $T = \{1, 5, 9, \dots, 21, 25, 29, 33, \dots\}$. The minimal element of T is $r = 1$, and then we obtain $q = \dfrac{a - r}{b} = 6$. And indeed, we have $25 = 6 \cdot 4 + 1$.
  - In practice, of course, we would not actually construct the sets S and T to determine q and r: we would just numerically compute $25/4$ and round down to the nearest integer to find q.

1.2.2 Greatest Common Divisors and Least Common Multiples

- We now discuss the idea of common divisors.

- Definition: If $d|a$ and $d|b$, then d is a common divisor of a and b. If a and b are not both zero, then there are only a finite number of common divisors: the greatest one is called the greatest common divisor, or gcd, and denoted by $\gcd(a, b)$.
  - Warning: Many authors use the notation (a, b) to denote the gcd of a and b; this stems from the notation used for ideals in ring theory. We will always write $\gcd(a, b)$ explicitly, since otherwise it is easy to confuse the gcd with an ordered pair (a, b).
  - Example: The positive divisors of 30 are 1, 2, 3, 5, 6, 10, 15, 30. The positive divisors of 42 are 1, 2, 3, 6, 7, 14, 21, 42. The common (positive) divisors are 1, 2, 3, and 6, and so $\gcd(30, 42) = 6$.

- Theorem (GCD as Linear Combination): If a and b are integers, not both zero, and $d = \gcd(a, b)$, then there exist integers x and y with $d = ax + by$: in fact, the gcd is the smallest positive such linear combination.
  - This theorem says that the greatest common divisor of two integers is an integral linear combination of those integers.
  - Proof: Without loss of generality assume $a \neq 0$, and let $S = \{as + bt : s, t \in \mathbb{Z}\} \cap \mathbb{Z}_+$.
  - Clearly $S \neq \emptyset$ since one of a and $-a$ is in S, so now let $l = ax + by$ be the minimal element of S.
  - We claim that $l|b$.
    - To see this, apply the division algorithm to write $b = ql + r$ for some $0 \leq r < l$.
    - Observe that $r = b - ql = b - q(ax + by) = a(-qx) + b(1 - qy)$ is a linear combination of a and b. It is not negative, but it also cannot be positive because otherwise it would necessarily be less than l, and l is minimal.
    - Hence $r = 0$, so $l|b$.
  - By a symmetric argument, $l|a$, and so l is a common divisor of a and b, whence $l \leq d$.
  - But now since $d|a$ and $d|b$ we can wrote $a = dk_a$ and $b = dk_b$ for some integers k_a and k_b, and then $l = ax + by = dk_a x + dk_b y = d(k_a x + k_b y)$.
  - Therefore $d|l$, so in particular $d \leq l$ since both are positive. Since $l \leq d$ as well, we conclude $l = d$.

- Corollary: If $l|a$ and $l|b$, then l divides $\gcd(a, b)$. In other words, the gcd of a and b is divisible by every other common divisor.
  - Proof: Since $l|a$ and $l|b$, l divides any linear combination of a and b: in particular, it divides the gcd.

- As an example: we saw above that the gcd of 30 and 42 is 6, and indeed we can see that $3 \cdot 30 - 2 \cdot 42 = 6$. The other common divisors are 1, 2, and 3, and indeed they all divide 6.

- As another example: because $6 \cdot 24 - 11 \cdot 13 = 1$, we see that 24 and 13 have greatest common divisor 1, since their gcd must divide any linear combination. Having a gcd of 1 occurs often enough that we give this situation a name:

- Definition: If $\gcd(a,b) = 1$, we say a and b are relatively prime.
  - Examples: 24 and 13 are relatively prime. 2 and 5 are relatively prime. 15 and 16 are relatively prime.
  - Non-Example: 30 and 69 are not relatively prime, since they have the common divisor 3.
- Using all of the results we have shown above, we can collect a number of useful facts about greatest common divisors:
- Proposition (Properties of GCDs): If m, a, b, d are integers, then the following hold:
  1. If $m > 0$, then $m \cdot \gcd(a,b) = \gcd(ma, mb)$.
     - Proof: We have $\gcd(ma, mb) = \min_{x,y \in \mathbb{Z}} [\{max + mby\} \cap \mathbb{Z}_+] = m \cdot \min_{x,y \in \mathbb{Z}} [\{ax + by\} \cap \mathbb{N}] = m \cdot \gcd(a,b)$.
  2. If $d > 0$ divides both a and b, then $\gcd(a/d, b/d) = \gcd(a,b)/d$.
     - Proof: Applying (1) to a/d and b/d with $m = d$ yields $d \cdot \gcd(a/d, b/d) = \gcd(a,b)$, which is equivalent to the given statement.
  3. There exist integers x and y with $ax + by = 1$ if and only if $\gcd(a,b) = 1$.
     - Proof: If $\gcd(a,b) = 1$ then we showed above that there exist integers x and y with $ax + by = 1$.
     - For the other direction, any common divisor of a and b must divide $ax + by = 1$: hence the gcd must divide 1, which leaves only the possibility that it equals 1.
  4. If a and b are both relatively prime to m, then so is ab.
     - Proof: By the linear combination property of the gcd, there exist x_1, y_1, x_2, y_2 with $ax_1 + my_1 = 1$ and $bx_2 + my_2 = 1$.
     - Multiplying these two equations together and rearranging the results yields $ab(x_1x_2) + m(y_1bx_2 + y_2ax_1 + my_1y_2) = 1$, and this implies that ab and m are relatively prime.
  5. For any integer x, $\gcd(a,b) = \gcd(a, b + ax)$.
     - Proof: Observe that the set of linear combinations of a and b is the same as the set of integral linear combinations of a and $b + ax$.
  6. If $a|bc$ and a and b are are relatively prime, then $a|c$.
     - Proof: By (1), we have $\gcd(ac, bc) = c \cdot \gcd(a,b) = c$. Since $a|bc$ and $a|ac$, we see that a is a common divisor of ac and bc, and therefore divides the gcd, which is c. Thus $a|c$ as claimed.
- Dual to the notion of the greatest common divisor is the notion of the least common multiple:
- Definition: If $a|l$ and $b|l$, l is a common multiple of a and b. Among all (nonnegative) common multiples of a and b, the smallest such l is called the least common multiple of a and b.
  - Example: The least common multiple of 30 and 42 is 210.
  - The least common multiple is often mentioned in elementary school in the context of adding fractions (for finding the "least common denominator").
- The least common multiple has fewer nice properties than the gcd:
- Proposition (Properties of LCMs): If m, a, b are any positive integers, then the following hold:
  1. We have $m \cdot \text{lcm}(a,b) = \text{lcm}(ma, mb)$.
     - Proof: Since ma divides $\text{lcm}(ma, mb)$, we can write $\text{lcm}(ma, mb) = mk$ for some integer k.
     - Then $ma|mk$ and $mb|mk$, so a and b both divide k, whence $k \geq l$, where $l = \text{lcm}(ma, mb)$.
     - On the other hand, certainly ma and mb divide ml, so $ml \geq mk$. We must therefore have $l = k$, so $m \cdot \text{lcm}(a,b) = \text{lcm}(ma, mb)$ as claimed.
  2. If a and b are positive integers, then $\gcd(a,b) \cdot \text{lcm}(a,b) = ab$.
     - Proof: First suppose a and b are relatively prime, and let l be a common multiple. Since $a|l$ we can write $l = ak$ for some integer k: then since $b|ak$ and $\gcd(a,b) = 1$, we conclude by properties of divisibility that $b|k$, meaning that $k \geq b$ and thus $l \geq ab$. But clearly ab is a common multiple of a and b, so it is the least common multiple.
     - In the general case, let $d = \gcd(a,b)$. Then $\gcd(a/d, b/d) = 1$, so by (1) we see that $\text{lcm}(a/d, b/d) = ab/d^2$. Then $\gcd(a,b) \cdot \text{lcm}(a,b) = d \cdot d\,\text{lcm}(a/d, b/d) = ab$, as desired.

1.2.3 The Euclidean Algorithm

- One question left unanswered raised by our results so far is how can we actually compute the gcd, except by actually writing down lists of common divisors. (And also how can we compute the gcd as a linear combination of the original integers.) Both questions turn out to have a nice answer:
- Theorem (Euclidean Algorithm): Given integers $0 < b < a$, repeatedly apply the division algorithm as follows, until a remainder of zero is obtained:

$$\begin{aligned} a &= q_1 b + r_1 \\ b &= q_2 r_1 + r_2 \\ r_1 &= q_3 r_2 + r_3 \\ &\vdots \\ r_{k-1} &= q_{k+1} r_k + r_{k+1} \\ r_k &= q_{k+2} r_{k+1}. \end{aligned}$$

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

  - Proof: First observe that the algorithm will eventually terminate, because $b > r_1 > r_2 > \cdots \geq 0$, and the well-ordering axiom dictates that we cannot have an infinite decreasing sequence of nonnegative integers.
  - We now claim that $\gcd(a, b) = \gcd(b, r_1)$: this follows because $\gcd(b, r_1) = \gcd(b, a - q_1 b) = \gcd(b, a)$ from the gcd properties we proved earlier.
  - Now by an easy induction, we claim that $\gcd(r_j, r_{j+1}) = \gcd(a, b)$ for each $0 \leq j \leq k$, where we set $r_0 = b$ and $r_{-1} = a$. The base case $j = 0$ follows from $\gcd(a, b) = \gcd(b, r_1)$ above, and the inductive step follows by applying the same argument to see $\gcd(r_j, r_{j+1}) = \gcd(r_{j+1}, r_{j+2})$.
  - We conclude that $\gcd(a, b) = \gcd(r_{k+1}, r_k) = r_{k+1}$ since r_{k+1} divides r_k. Hence, $\gcd(a, b)$ is the last nonzero remainder as claimed.
  - The correctness of the algorithm for computing the gcd also follows by an easy induction: explicitly, we show by induction on j that there exist integers x_j and y_j such that $r_j = x_j a + y_j b$ for all integers j with $0 \leq j \leq k + 1$.
  - The base cases $j = 0$ and $j = 1$ follow by writing $r_0 = b$ and $r_1 = a - q_1 b$ so we may take $x_0 = 0$, $y_0 = 1$, $x_1 = 1$, and $y_1 = -q_1$.
  - The inductive step follows by writing $r_{j-1} = q_{j+1} r_j + r_{j+1}$, so rearranging yields $r_{j+1} = r_{j-1} - q_{j+1} r_j = (x_{j-1} a + y_{j-1} b) - q_{j+1}(x_j a + y_j b) = (x_{j-1} - q_{j+1} x_j) a + (y_{j-1} - q_{j+1} y_j) b$ and thus we take $x_{j+1} = x_{j-1} - q_{j+1} x_j$ and $y_{j+1} = y_{j-1} - q_{j+1} y_j$.
  - By induction, we eventually obtain an expression $\gcd(a, b) = r_{k+1} = x_{k+1} a + y_{k+1} b$ as required.

- Example: Find the gcd of 133 and 98 using the Euclidean algorithm, and write the gcd explicitly as a linear combination of 133 and 98.

  - First, we use the Euclidean algorithm:

$$\begin{aligned} 133 &= 1 \cdot 98 + 35 \\ 98 &= 2 \cdot 35 + 28 \\ 35 &= 1 \cdot 28 + 7 \\ 28 &= 4 \cdot 7 \end{aligned}$$

    and so the gcd is $\boxed{7}$.
  - For the linear combination, we solve for the remainders:

$$\begin{array}{rclcccl} 35 &=& 133 - 1 \cdot 98 &=& & = & 1 \cdot 133 - 1 \cdot 98 \\ 28 &=& 98 - 2 \cdot 35 &=& 98 - 2 \cdot (133 - 1 \cdot 98) & = & -2 \cdot 133 + 3 \cdot 98 \\ 7 &=& 35 - 1 \cdot 28 &=& (1 \cdot 133 - 1 \cdot 98) - 1 \cdot (-2 \cdot 133 + 3 \cdot 98) & = & 3 \cdot 133 - 4 \cdot 98 \end{array}$$

    so we obtain $\boxed{7 = 3 \cdot 133 - 4 \cdot 98}$.

- In the example above, we could simply have written down all the divisors of each number, and computed the gcd by comparing those lists. However, if the numbers are large, this procedure becomes very inefficient in comparison to the Euclidean algorithm.
- Example: Find the gcd of 44773 and 8537 using the Euclidean algorithm, and use the results to write the gcd as an explicit linear combination.
  - Applying the Euclidean algorithm to $a = 44773$ and $b = 8537$ yields

$$\begin{aligned} 44773 &= 5 \cdot 8537 + 2088 \\ 8537 &= 4 \cdot 2088 + 185 \\ 2088 &= 11 \cdot 185 + 53 \\ 185 &= 3 \cdot 53 + 26 \\ 53 &= 2 \cdot 26 + 1 \\ 26 &= 26 \cdot 1 \end{aligned}$$

  - Hence the gcd is $\boxed{1}$. For the linear combination, we solve for the remainders:

$$\begin{array}{rcccc} 2088 & = & & = & 1 \cdot 44773 - 5 \cdot 8537 \\ 185 & = & 8537 - 4 \cdot 2088 & = & -4 \cdot 44773 + 21 \cdot 8537 \\ 53 & = & 2088 - 11 \cdot 185 & = & 45 \cdot 44773 - 236 \cdot 8537 \\ 26 & = & 185 - 3 \cdot 53 & = & -139 \cdot 44773 + 729 \cdot 8537 \\ 1 & = & 53 - 2 \cdot 26 & = & 323 \cdot 44773 - 1694 \cdot 8537 \end{array}$$

  and therefore we can take $s = \boxed{323}$ and $t = \boxed{-1694}$.

1.3 Primes and Unique Factorization

- Now that we have examined divisibility and common factors, we will examine one of the other fundamental properties of the integers, namely, the existence and uniqueness of prime factorizations.
- We begin by discussing prime numbers:
- Definition: If $p > 1$ is an integer, we say it is prime if there is no d with $1 < d < p$ such that $d|p$. (In other words, if p has no proper divisors.) If $n > 1$ is not prime, we say it is composite.
  - The first few primes are 2, 3, 5, 7, 11, 13, 17, 19, and so forth. 1 is neither prime nor composite.
  - Remark: In more advanced contexts, the following equivalent definition of a prime is often used instead: the integer $p > 1$ is prime if and only if $p|ab$ implies that $p|a$ or $p|b$.
- The prime numbers are often called the "building blocks under multiplication", because every positive integer can be written as the product of prime numbers in an essentially unique way. To prove this, we first show that there exists at least one such factorization:
- Proposition (Existence of Prime Factorizations): Every positive integer $n > 1$ can be written as a product of one or more primes (where a "product" is allowed to have only one term).
  - The representation of n as a product of primes is called the prime factorization of n. (For example, the prime factorization of 6 is $6 = 2 \cdot 3$.) We will show in a moment that it is unique up to reordering the terms.
  - Proof: We use strong induction on n. The result clearly holds if $n = 2$, since 2 is prime.
  - Now suppose $n > 2$. If n is prime, we are done, so assume that n is not prime, hence composite. By definition, there exists a d with $1 < d < n$ such that $d|n$: then n/d is an integer satisfying $1 < n/d < n$. By the strong induction hypothesis, both d and n/d can be written as a product of primes; multiplying these two products then yields n as a product of primes.

- To establish the uniqueness of prime factorizations, we require the following prime divisibility property:
- Proposition (Prime Divisibility): If a and b are integers and p is a prime number with $p|ab$, then $p|a$ or $p|b$.
    - Proof: If $p|a$ we are done, so assume that $p \nmid a$.
    - Consider $\gcd(a, p)$: it divides p, hence is either 1 or p, but it is not p because p does not divide a.
    - Therefore, $\gcd(a, p) = 1$, so a and p are relatively prime.
    - Then since $p|ab$ and a, p are relatively prime, we see that $p|b$.
- Theorem (Fundamental Theorem of Arithmetic): Every integer $n > 1$ can be factored into a product of primes, and this factorization is unique up to reordering of the factors.
    - Proof: Suppose by way of contradiction that there is a positive integer with two prime factorizations. By the well-ordering axiom, we may select the minimal such positive integer n with two different factorizations: $n = p_1 p_2 \cdots p_k = q_1 q_2 \cdots q_l$. If any of the primes p_i and q_i were equal, we could cancel the corresponding terms and obtain a smaller n, so $p_1 \neq q_j$ for any j with $1 \leq j \leq l$.
    - But since p_1 is prime and divides $q_1 q_2 \cdots q_l$, by repeated application of the previous proposition we see that p_1 must divide one of $q_1, q_2, \ldots, q_l$: say, q_i. But the only divisors of q_i are 1 and q_i, and p_1 cannot be either of them. This is a contradiction, and we are done.
- To save space, we group equal primes together when actually writing out the canonical prime factorization: thus, $12 = 2^2 \cdot 3$, $720 = 2^2 \cdot 3^2 \cdot 5$, and so forth. More generally, we often write the prime factorization in the form $n = \prod_{i=1}^{j} p_i^{n_i}$, where the p_i are some (finite) set of primes and the n_i are their corresponding exponents.
- Proposition (Divisibility and Factorizations): If $a = \prod_{i=1}^{j} p_i^{a_i}$ and $b = \prod_{i=1}^{j} p_i^{b_i}$ for distinct primes p_i, then $a|b$ if and only if $a_i \leq b_i$ for each i. In particular, $\gcd(a, b) = \prod_{i=1}^{j} p_i^{\min(a_i, b_i)}$ and $\text{lcm}(a, b) = \prod_{i=1}^{j} p_i^{\max(a_i, b_i)}$.
    - Proof: We observe that if $b = ak$ and $k = \prod_{i=1}^{j} p_i^{k_i}$, then $a_i + k_i = b_i$. Since all exponents are nonnegative, saying that such an integer k exists is equivalent to saying that $a_i \leq b_i$ for all i.
    - The statements about the gcd and lcm are then immediate, since the exponent of p_i in the gcd is the largest integer that is $\leq a_i$ and $\leq b_i$, which is simply the minimum of a_i and b_i, and the exponent of p_i in the lcm is the least integer that is $\geq a_i$ and $\geq b_i$, which is simply the maximum of a_i and b_i.
- One question we might have is: how many primes are there? The most basic answer to this question is that there are infinitely many primes:
- Theorem (Euclid): There are infinitely many prime numbers.
    - Proof: Suppose we have a list of primes $p_1, p_2, \ldots, p_k$, and consider $n = p_1 p_2 \cdots p_k + 1$.
    - Let q be any prime divisor of n (possibly n itself). We claim that q cannot equal any of the p_i, so suppose otherwise: then p_i divides $n = p_1 p_2 \cdots p_k + 1$.
    - Since p_i also divides $p_1 p_2 \cdots p_k$, we see that p_i therefore divides $n - p_1 p_2 \cdots p_k = 1$. But this is a contradiction. Hence q is a new prime not on the list.
    - Therefore, any finite list cannot contain all of the prime numbers, so there are infinitely many.
- One particularly famous application of prime factorizations is to show that $\sqrt{2}$ is irrational:
- Theorem (Irrationality of $\sqrt{2}$): The number $\sqrt{2}$ is irrational, which is to say, there do not exist integers m and n such that $\sqrt{2} = m/n$.
    - Proof: Suppose by way of contradiction that $\sqrt{2}$ were rational so that $\sqrt{2} = m/n$ for some integers m and n, which (by negating if needed) we may assume are positive.
    - Squaring both sides and clearing denominators yields the equivalent equation $2n^2 = m^2$.
    - Now consider the prime factorizations of both sides: say $m = 2^{m_2} 3^{m_3} \cdots$ and $n = 2^{n_2} 3^{n_3} \cdots$.
    - We obtain the equality $2^{2m_2+1} 3^{m_3} \cdots = 2^{2n_2} 3^{2n_3} \cdots$, and so by the uniqueness of prime factorizations, all of the corresponding exponents must be equal.

- In particular, we see that $2m_2 + 1 = 2n_2$, so that $2(n_2 - m_2) = 1$. But this is impossible, because 2 does not divide 1.
- Therefore, it could not have been true that $\sqrt{2} = m/n$, so $\sqrt{2}$ must be irrational as claimed.

- There are very many results and open problems relating to prime numbers, some of which are as follows:

  The Prime Number Theorem: Euclid's result, while extremely elegant, does not tell us much about the actual primes themselves: for example, it does not say anything about how common the primes are. Are most numbers prime? Or are most numbers composite? More rigorously, if we let $\pi(n)$ be the number of primes in the interval $[1, n]$, we would like to know how fast $\pi(n)$ increases as n increases: does it grow like n, or $\sqrt{n}$, or something else?
  The answer is given by the Prime Number Theorem proven independently by Hadamard and de la Vallée Poissin in 1896: $\pi(n) \sim \dfrac{n}{\ln(n)}$, meaning that as $n \to \infty$, the limit $\lim\limits_{n\to\infty} \dfrac{\pi(n)}{n/\ln(n)} = 1$.

  Twin Primes: Another question is: how close do primes get? It is obvious that 2 is the only even prime, so aside from 2 and 3, any pair of primes has to differ by at least 2: such pairs are called twin primes. One can write down a long list of twin primes: (3,5), (5,7), (11,13), (17,19), (29,31), (41,43), (59,61), and so forth. Are there infinitely many? The answer is not known, although twin primes are expected to be quite rare. However, as of August 2014, the Polymath project has sharpened results of Maynard and Zhang to show that there exist infinitely many pairs of primes (p_1, p_2) such that $|p_2 - p_1| \leq 246$.

  Goldbach's Conjecture: One can observe that $2 + 2 = 4$, $3 + 3 = 6$, $3 + 5 = 8$, $3 + 7 = 10$, $5 + 7 = 12$, $3 + 11 = 14$, $3 + 13 = 16$, $5 + 13 = 18$, $7 + 13 = 20$, and so forth. It appears that every even number greater than 2 can be written as the sum of two primes. It is not known whether this pattern continues, although it has been numerically verified for every even integer less than $4 \cdot 10^{18}$. In 2013, a proof that every odd integer greater than 5 can be written as a sum of three primes was announced by Helfgott. (This result is weaker than Goldbach's conjecture, but it is of the same type.)

1.4 Rings and Other Number Systems

- One of our goals in number theory is to examine properties of the integers that generalize to other number systems. To do this in a reasonable way, we will phrase our results using the language of commutative rings.
  - A fuller discussion of the theory of rings belongs to abstract algebra, but the language of rings provides the best setting in which to compare different number systems.
  - We will only give a brief overview of ring arithmetic here, leaving a more extensive discussion of generalizations to a later chapter.

1.4.1 Definition and Examples

- Definition[3]: A commutative ring with 1 is a set R having two (closed) binary operations $+$ and $\cdot$ that satisfy the eight axioms [R1]-[R8]:

  [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.

[3] We will remark here that there are also notions of a noncommutative ring (in which the axiom [R7] on the commutativity of multiplication is dropped) and a ring without 1 (in which the axiom [R8] on the existence of a multiplicative identity is dropped). We will not use these more general types of rings, but they are important objects of study in abstract algebra. One fundamental example of a noncommutative ring, familiar from linear algebra, is the ring of $n \times n$ matrices under matrix addition and multiplication.

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

[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 commutative ring with identity further satisfies the axiom [F], it is called a field.

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

  - Remark: Fields may be a familiar object from linear algebra, as fields are a central underlying component of a vector space.

- Here are some examples (and non-examples) of rings[4]:
- Example: The integers $\mathbb{Z}$ are a commutative ring with identity.
- 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.
- (Non-)Example: The set of even integers forms a commutative ring without 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 element $n \in R$ such that $n \cdot 2 = 2$.
- 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.
  - 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 generalize the two examples above: if D is any integer, the set of complex numbers of the form $a+b\sqrt{D}$ for $a, b \in \mathbb{Z}$ forms a ring, denoted $\mathbb{Z}[\sqrt{D}]$.
  - Associated to this ring is a particularly important map called the norm map, which is defined as follows: $N(a+b\sqrt{D}) = (a+b\sqrt{D})\cdot(a-b\sqrt{D}) = a^2 - Db^2$. Observe that this function takes values in the integers, and that it is also multiplicative: $N(rs) = N(r)N(s)$ for any $r, s \in R$.
- There is one other important class of rings we will discuss, namely, polynomial rings:
- Definition: Let F be a field and x be an indeterminate. A polynomial in x with coefficients in F 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 $a_i \in F$. The set of all such polynomials is denoted by $F[x]$ and has the structure of a commutative ring with 1 under the familiar addition and multiplication of polynomials:

[4] 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.

- Addition is defined "termwise":

  $(a_nx^n+a_{n-1}x^{n-1}+\cdots+a_1x+a_0)+(b_nx^n+b_{n-1}x^{n-1}+\cdots+b_1x+b_0) = (a_n+b_n)x^n+(a_{n-1}+b_{n-1})x^{n-1}+\cdots+(a_0+b_0)$

  and 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 $a_0b_j + a_1b_{j-1} + \cdots + a_jb_0 = \sum_{k=0}^{j} a_kb_{j-k}$.
- These operations are all defined, since F is a field, and it is reasonably straightforward to see that each of the axioms [R1]-[R8] hold: thus, $F[x]$ is a commutative ring with 1.
- The term "indeterminate" is deliberately undefined in the definition above. A more concrete[5] (but vastly less intuitive) definition of polynomials can be given using Cartesian products, 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_nx^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.

- The degree of a polynomial is quite fundamental so we will record a few basic properties now:
- Proposition (Degrees in Polynomial Rings): If p and q are any polynomials in a polynomial ring $F[x]$, then $\deg(p+q) \leq \max(\deg p, \deg q)$, and $\deg(pq) = \deg p + \deg q$.
  - 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 k or higher in p or q, then there are no terms of degree k or higher in $p+q$ either.
  - For pq, observe that if p has leading term a_nx^n and q has leading term b_mx^m, then the leading term of pq is $a_nb_mx^{m+n}$ since $a_nb_m \neq 0$ because F is a field and a_n and b_m are nonzero.

1.4.2 Arithmetic in Rings, Units

- Our immediate goal in discussing rings is to study properties of arithmetic in $\mathbb{Z}$ 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:

[5] Specifically: inside the Cartesian product $\prod_{\mathbb{Z}_{\geq 0}} F = (r_0, r_1, r_2, \dots)$ 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 $r_0 + r_1x + r_2x^2 + \cdots + r_nx^n$. We can then define the operations of polynomial addition and multiplication solely in terms of these sequences.

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.

- It is 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 commutative ring R with $1 \neq 0$, we say that an element a is a unit if there exists a $b \in R$ such that $ab = 1$.
  - The set of units in R is denoted $R^\times$.
  - Example: In $\mathbb{Z}$, there are only two units, namely 1 and -1.
  - Example: In a field, every nonzero element is a unit. In fact, the field axiom [F] says that a commutative ring with $1 \neq 0$ is a field if and only if every nonzero element is a unit.
  - Example: In the ring of polynomials $\mathbb{R}[x]$, the units are the nonzero constant polynomials. To see this, observe that all of the nonzero constants are units, and that a nonconstant polynomial $p(x)$ cannot satisfy $p(x) \cdot q(x) = 1$ by degree considerations: if $q(x)$ is zero then so is the product, and otherwise, the degree of the product is at least the degree of $p(x)$.

- Example: In the ring of Gaussian integers $\mathbb{Z}[i]$, the units are 1, -1, i, and $-i$. This can be seen as follows: suppose that $a+bi$ is a unit, with $(a+bi)(c+di)=1$. Applying the norm map $N(a+bi)=a^2+b^2$ to both sides yields $(a^2+b^2)(c^2+d^2)=1$. But both of these quantities are nonnegative integers, so it must be the case that $a^2+b^2=c^2+d^2=1$. But the only integral solutions to this equation are $(a,b)=(\pm 1,0)$ and $(0,\pm 1)$, meaning that the only units are ± 1 and $\pm i$.
- 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 $(\sqrt{2}+1)\cdot(\sqrt{2}-1)=1$. This observation implies that any power of $\sqrt{2}+1$ is a unit in this ring, since its inverse is the corresponding power of $\sqrt{2}-1$. Therefore, $\mathbb{Z}[\sqrt{2}]$ has infinitely many units, unlike $\mathbb{Z}[i]$ which has only 4 units. Note that $\mathbb{Z}[\sqrt{2}]$ is not a field, however, because $\sqrt{2}$ is not a unit.

- We can generalize the observations about units in $\mathbb{Z}[i]$ and $\mathbb{Z}[\sqrt{2}]$.
- Proposition (Units in $\mathbb{Z}[\sqrt{D}]$): For a fixed D, an element r in the ring $\mathbb{Z}[\sqrt{D}]$ is a unit if and only if $N(r)=\pm 1$.
  - Proof: Suppose $r=a+b\sqrt{D}$.
  - If $N(r)=\pm 1$, then we see that $r\cdot(a-b\sqrt{D})=\pm 1$, so (by multiplying by -1 if necessary) we see that $\pm(a-b\sqrt{D})$ is a multiplicative inverse for r.
  - Conversely, suppose r is a unit and $rs=1$. Taking norms yields $N(rs)=N(r)N(s)=1$. Since $N(r)$ and $N(s)$ are both integers, we see that $N(r)$ must either be 1 or -1.
- Example: Determine whether $\sqrt{3}$, $1+\sqrt{3}$, and $2+\sqrt{3}$ are units in $\mathbb{Z}[\sqrt{3}]$.
  - We have $N(\sqrt{3})=\sqrt{3}\cdot\sqrt{3}=3$, $N(1+\sqrt{3})=(1+\sqrt{3})(1-\sqrt{3})=-2$, and $N(2+\sqrt{3})=(2+\sqrt{3})(2-\sqrt{3})=1$. So $\sqrt{3}$ and $1+\sqrt{3}$ are $\boxed{\text{not units}}$, while $2+\sqrt{3}$ $\boxed{\text{is a unit}}$.
- Here are a few basic properties of units in general rings:
- Proposition (Units): Let R be a ring with $1\neq 0$. Inside R, the multiplicative inverse of a unit is unique, the product of two units is a unit, and the multiplicative inverse of a unit is also a unit.
  - Proof: For uniqueness, if a is a unit with $ab=1$ and also $ac=1$, then $b=b(ac)=(ba)c=c$.
  - For the second statement, if a is a unit with $ab=1=ba$, then by definition b is also a unit.
  - For the last statement, 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.
  - Remark (for those who like group theory): Together with the observation that 1 is a unit, the second and third statements 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 with 1 and $a,b\in R$, we say that a divides b, written $a|b$, if there exists some $k\in R$ such that $b=ak$.
  - Example: In $\mathbb{Z}[i]$, the element $2+i$ divides 5, because $5=(2+i)(2-i)$.
  - Example: In $\mathbb{R}[x]$, the polynomial $3x+6$ divides the polynomial x^2-x-6, because $x^2-x-6=(3x+6)\cdot(\frac{1}{3}x-1)$.
  - 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!
- Some of the properties of divisibility inside $\mathbb{Z}$ carry over to general commutative rings R:
- Proposition (Basic Divisibility in Rings): If R is a commutative ring with 1, then for any $a,b,c,x,y\in R$, the following hold:

1. If $a|b$, then $a|bc$ for any c in R.

2. If $a|b$ and $b|c$, then $a|c$.
3. If $a|b$ and $a|c$, then $a|(xb+yc)$ for any x and y in R.
    - Proof: Each of these follows immediately from the definition of divisibility, in the same way as in $\mathbb{Z}$.

- However, many other properties of divisibility do not carry over to general rings. To illustrate, we give an example of an element that has two different factorizations into irreducible terms.
- Definition: If R is a commutative ring with 1, a nonzero element $a \in R$ is irreducible if it is not a unit and, for any "factorization" $a = bc$ 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).
    - Example: The element 5 is reducible in $\mathbb{Z}[i]$, since we can write $5 = (2+i)(2-i)$ and neither $2+i$ nor $2-i$ is a unit in $\mathbb{Z}[i]$.
- Example: Show that 6 possesses two different factorizations as products of irreducible elements in $\mathbb{Z}[\sqrt{-5}]$.
    - Observe that $6 = (1+\sqrt{-5})(1-\sqrt{-5}) = 2 \cdot 3$.
    - We claim that each of $1 \pm \sqrt{-5}$, 2, and 3 is irreducible in $\mathbb{Z}[\sqrt{-5}]$.
    - To see this, observe that $N(1 \pm \sqrt{-5}) = 6$, $N(2) = 4$, and $N(3) = 9$.
    - If we had a nontrivial factorization $3 = ab$, then taking norms would give $9 = N(3) = N(a)N(b)$. Since neither a nor b is a unit by hypothesis, both $N(a)$ and $N(b)$ must be greater than 1.
    - The only possibility is then $N(a) = N(b) = 3$, but there are no elements of norm 3 in $\mathbb{Z}[\sqrt{-5}]$ since $N(c+d\sqrt{-5}) = 3$ would require an integer solution to the equation $c^2 + 5d^2 = 3$, and there clearly are no such solutions.
    - In the same way we can see that there are no elements of norm 2 in $\mathbb{Z}[\sqrt{-5}]$. Hence 2 and $1 \pm \sqrt{-5}$ must also be irreducible.
    - But then we can see that neither of the factorizations $(1+\sqrt{-5})(1-\sqrt{-5})$ and $2 \cdot 3$ can be expanded further, and they also cannot be transformed into one another by multiplying by units (since the only units in $\mathbb{Z}[\sqrt{-5}]$ are ± 1).
    - This means that they are, in fact, completely different factorizations of the same number.
- We will continue our examination of these properties in other rings in a later chapter.

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