CHAPTER 4

Semigroups

We very often encounter binary operations in mathematics, and nearly all of these are associative: addition, multiplication, composition etc. In this chapter we introduce a sufficiently abstract notion to deal with all such operations.

4.1. Definition of a Semigroup

4.1.1 Definition A *semigroup* is an ordered pair $(S, *)$ such that S is a non-empty set, and $*$ is an associative binary operation on S.

Note that S must be non-empty. We usually abuse notation and just talk of "the semigroup S" instead of "the semigroup $(S, *)$". This can be dangerous, because the same set may have more than one binary operation on it.

Working in such generality takes some getting used to: any statement we prove about semigroups will be true for addition of real numbers, addition of matrices, multiplication of matrices, composition of functions etc!

4.1.2 Example

- ★ $\mathbb{R}$ is a semigroup under the binary operation $+$, since $+$ is associative. $\mathbb{R}$ is also a semigroup under multiplication. These two semigroups are not the same, since the binary operation is different.
- ★ $\mathbb{R}$ is not a semigroup under subtraction.
- ★ $\mathbb{R}^n$ is a semigroup under $+$. More generally, any vector space V is a semigroup under vector addition $+$.
- ★ $\mathbb{R}^3$ has another binary operation, the cross product $\times$. However this is not associative, so $(\mathbb{R}^3, \times)$ is not a semigroup.
- ★ $\mathbb{Z}/n\mathbb{Z}$ is a semigroup under addition, and also a semigroup under multiplication.
- ★ The set of $n \times n$ real matrices $M_n(\mathbb{R})$ is a semigroup under addition, and also a (different) semigroup under multiplication.
- ★ The set $\mathscr{F}$ of all functions from $\mathbb{R}$ to $\mathbb{R}$ is a semigroup under composition. More generally, if A is any set, the set of all functions $A \to A$ is a semigroup under composition. The set of functions from A to B is not a semigroup under composition, since if $f, g\colon A \to B$ we cannot compose f and g.
- ★ Let A be any set. Let $\mathscr{P}(A)$ consist of all the subsets of A. (This set is often called the *power set* of A.) Then $(\mathscr{P}(A), \cup)$ is a semigroup, where $\cup$ is the union operation. $(\mathscr{P}(A), \cap)$ is another semigroup.

4.1.3 Example Let $\mathscr{F}$ be the set of all functions $f\colon \mathbb{R} \to \mathbb{R}$. We would like to define the sum of two functions $f, g\colon \mathbb{R} \to \mathbb{R}$. We define this to be the function $f + g$ satisfying

$$(f + g)(x) = f(x) + g(x).$$

This is called *pointwise* addition. The displayed equation is not just an obviously true statement. It is the *definition* of addition of functions, using only knowledge of how to add *numbers.* This distinction is unfortunately lost in the notation, which uses the same symbol + to mean two entirely different things: addition of numbers, and addition of functions.

Claim: + is associative.

Proof We have to prove $f + (g + h) = (f + g) + h$. This is an equality of functions, so we have to show the two sides agree for every input. But

$$\begin{aligned}
(f + (g + h))(x) &= f(x) + (g + h)(x) && \text{Def sum of } f \text{ and } (g + h) \\
&= f(x) + \Big(g(x) + h(x)\Big) && \text{Def sum } (g + h) \\
&= \Big(f(x) + g(x)\Big) + h(x) && \text{Associativity + in } \mathbb{R} \\
&= (f + g)(x) + h(x) && \text{Def sum } (f + g) \\
&= ((f + g) + h)(x) && \text{Def sum of } (f + g) \text{ and } h. \quad \Box
\end{aligned}$$

This proves associativity of addition. Thus $\mathscr{F}$ forms a semigroup under addition.

In a similar way we can define the pointwise product of f and g by

$$(f \cdot g)(x) = f(x) \cdot g(x).$$

Then $\cdot$ is also associative. The proof is similar.

4.1.4 Convention We often use juxtaposition to indicate the semigroup binary operation. Thus we write ab instead of $a * b$. We also write a^2 for $a * a$, and a^3 for $a * a * a$ and so on.

Similarly, the binary operation on a semigroup S may often be called the *product* on S. This does not necessarily mean that the binary operation *is* multiplication, only that it is analogous to multiplication.

There is an exception: if the binary operation is commonly denoted +, then we stick with that notation.

4.1.5 Theorem Let S be a semigroup, let $a \in S$ and let m, $n \in \mathbb{N}$. Then

(a) $a^m a^n = a^{m+n}$,
(b) $(a^m)^n = a^{mn}$.

Proof The proof is the same as the proof for exponentiation of integers: $a^m a^n$ consists of m copies of a followed by n copies of a, which is $m + n$ copies altogether. Etc. $\Box$

Warning:

$$(ab)^n \neq a^n b^n$$

in general. Indeed $(ab)^n = (ab)(ab) \cdots (ab)$ but we may not have $ab = ba$ so we cannot collect all the a's on one side and all the b's on the other, since we are not assuming that our binary operation is commutative. For example $(AB)^2 \neq A^2 B^2$ for matrices.

4.2. Identities

4.2.1 Definition Let $*$ be a binary operation on a set A. We say that an element $e \in A$ is an *identity* for $*$ if $a * e = a = e * a$ for every $a \in A$.

Note that an identity must satisfy *both* conditions $a * e = a$ and $e * a = a$.

4.2.2 Theorem Let A be a set, and let $*$ be a binary operation on A. If an identity exists for $*$, it is unique.

Proof Assume that e and e' are both identities. Then $e * e' = e'$ since e is an identity. But $e * e' = e$ also, since e' is an identity. Hence $e = e'$. $\Box$

Thus we talk of *the identity* not an identity. From now on, we shall usually use 1_A or 1 to denote the identity in A. The exception is where the operation is $+$, when we use 0_A to denote the identity.

4.2.3 Example Let A be a set and let $\mathscr{F}$ be the set of all functions $A \to A$ under composition. The identity is the identity function 1_A, so this notation agrees with our earlier definition of 1_A, Example 3.2.2.

4.2.4 Example

- ★ 1 is the identity for $\cdot$ (multiplication) in the semigroup $(\mathbb{R}, \cdot)$.
- ★ 0 is the identity for $+$ in $(\mathbb{R},+)$.
- ★ The set of natural numbers $\mathbb{N} = \{1, 2, \ldots\}$ is a semigroup under $+$. However $0 \notin \mathbb{N}$, so there is no identity in $(\mathbb{N}, +)$.
- ★ There is no identity for subtraction $-$ on $\mathbb{R}$. *Proof* If $a - e = a = e - a$ then $e = 0$. But then $e - a = 0 - a = -a \neq a$ in general.
- ★ 0 is the identity in $(\mathbb{Z}/n\mathbb{Z}, +)$, and 1 is the identity in $(\mathbb{Z}/n\mathbb{Z}, \cdot)$.
- ★ Consider multiplication of $n \times n$ matrices with real entries. The $n \times n$ identity matrix I is the identity, since $I \cdot A = A = A \cdot I$ for every matrix A. Hence the name "identity matrix".

46 Exercise Is there an identity in $(\mathscr{P}(A), \cup)$? Same question for $(\mathscr{P}(A), \cap)$.

4.3. Inverses

4.3.1 Definition Let S be a semigroup with identity 1, and let a be an element of S. We say that b is an *inverse* for a if

$$ab = 1 = ba.$$

There are two conditions to check: $ab = 1$ and $ba = 1$.

We have already seen the following theorem twice before: for numbers mod n (theorem 2.6.7) and for functions (theorem 3.4.2). You have already seen the same result for matrices. Now we can prove the general result behind all these special cases.

4.3.2 Theorem Let S be a semigroup with identity. Let $a \in S$. If a has an inverse, the inverse is unique.

Proof Suppose a has two inverses, b and c. Then

$$\begin{aligned} b &= 1 \cdot b && \text{Definition 1} \\ &= (ca)b && c \text{ is an inverse of } a \\ &= c(ab) && \text{Associativity} \\ &= c \cdot 1 && b \text{ is an inverse of } a \\ &= c && \text{Definition 1} \end{aligned}$$

□

4.3.3 Definition We write a^{-1} for the unique inverse of a.

4.3.4 Example

- ★ In $(\mathbb{R}, \cdot)$ the identity is 1. Every element a has an inverse $a^{-1} = 1/a$ except for 0. 0 does not have an inverse, since $0 \cdot b = 1$ is impossible.
- ★ In $(\mathbb{R}, +)$ the identity is 0. Every element a has inverse $-a$, since $a + (-a) = 0 = (-a) + a$.
- ★ Similarly, in $(\mathbb{Z}/n\mathbb{Z}, +)$ the identity is 0 and every a has inverse $-a$.
- ★ In $(M_n(\mathbb{R}), \cdot)$ the identity is the $n \times n$ identity matrix, and the invertible elements are the invertible matrices (those matrices with non-zero determinant).

- ★ The set of functions $\mathbb{R} \to \mathbb{R}$ forms a semigroup under pointwise addition (see Example 4.1.3). The identity is the constant function 0. Every element is invertible. The inverse of f is $-f$.
- ★ The set of functions $\mathbb{R} \to \mathbb{R}$ under composition is also a semigroup. The identity is the function $1_{\mathbb{R}}$. According to theorem 3.4.5 the invertible functions are exactly the bijections.

Thus we see why the inverse of a function and the inverse of a matrix both use the same notation and terminology: both *are* examples of inverses, although they are in different semigroups.

We proved half of the next result for the semigroup $(\mathbb{Z}/n, \cdot)$ in theorem 2.6.7. You are also familiar with this result for matrices. We now prove it in full generality.

4.3.5 Theorem Let S be a semigroup with identity, and suppose a, $b \in A$.

(a) If a is invertible then so is a^{-1} and $\left(a^{-1}\right)^{-1} = a$.
(b) If a and b are invertible, so is ab, and $(ab)^{-1} = b^{-1}a^{-1}$.

Proof

(a) We must show that the inverse of a^{-1} is a. That is, we must show that $aa^{-1} = 1 = a^{-1}a$. But this is true by the definition of a^{-1}.

(b) $(ab)(b^{-1}a^{-1}) = a(bb^{-1})a^{-1} = a1a^{-1} = aa^{-1} = 1$. Similarly $(b^{-1}a^{-1})(ab) = 1$. Thus the inverse of ab exists and is equal to $b^{-1}a^{-1}$. □

4.3.6 Definition Let S be a semigroup with identity 1. We define[1]

$$a^0 = 1.$$

for every a. Suppose $a \in S$ is invertible. Let $n \in \mathbb{N}$. We define

$$a^{-n} = (a^{-1})^n.$$

So if a is invertible, then a^m is defined for all integers m. Theorem 4.1.5 extends:

4.3.7 Theorem Let S be a semigroup with identity, and let $a \in S$ be invertible. Then for every m, $n \in \mathbb{Z}$

(a) $a^m a^n = a^{m+n}$,
(b) $(a^m)^n = a^{mn}$,
(c) $(a^n)^{-1} = (a^{-1})^n$.

Proof

(a) We check all of the possible cases. If m, $n > 0$ the result is theorem 4.1.5. If m or $n = 0$ the result is obvious. Suppose $m > 0$ but $n < 0$. Then $a^m a^n = a^m(a^{-1})^{|n|}$. If $m \geq |n|$ we may cancel out all the a^{-1} terms one by one, leaving $a^{m-|n|} = a^{m+n}$. If $m < |n|$ we may cancel all the a terms, leaving $(a^{-1})^{|n|-m}$. By definition, this is $a^{m-|n|} = a^{m+n}$. Similarly if $m < 0$ but $n > 0$. Finally if m, $n < 0$ let $b = a^{-1}$. Then $a^m a^n = b^{|m|}b^{|n|} = b^{|m|+|n|} = a^{-(|m|+|n|)} = a^{m+n}$.

(b) is similar. For (c), $a^n a^{-n} = a^0 = 1$ using (a) and $a^{-n}a^n = 1$ similarly, so the inverse of a^n is a^{-n}. □

Warning: Remember that $(ab)^n \neq a^n b^n$ unless we know that multiplication is commutative.

[1]Except if S is written additively, we usually do not define 0^0.