Rings, Determinants, the Smith Normal Form, and Canonical Forms for Similarity of Matrices.

Supplementary notes for Mathematics 701, Spring 2008.

Ralph Howard
Department of Mathematics
University of South Carolina
Columbia, S.C. 29208, USA
howard@math.sc.edu

Contents

1. Rings.

1.1. **The definition of a ring.** We have been working with fields, which are the natural generalization of familiar objects like the real, rational and complex numbers where it is possible to add, subtract, multiply and divide. However there are other natural objects such as the integers and polynomials over a field where we can add, subtract, and multiply, but where it not possible to divide. Such objects are called rings. Here is the official definition:

Definition 1.1. A ***commutative ring*** $(R, +, \cdot)$ is a set R with two binary operations $+$ and $\cdot$ (we usually just write $x \cdot y = xy$) so that

(1) The operations $+$ and $\cdot$ are both commutative and associative:

$$x+y = y+x, \quad x+(y+z) = (x+y)+z, \qquad xy = yx, \quad x(yz) = (xy)z.$$

(2) Multiplication distributes over addition:

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

(3) There is a unique element $0 \in R$ so that for all $x \in R$

$$x + 0 = 0 + x = x.$$

This element will be called the ***zero*** of R.

(4) There is a unique element $1 \in F$ so that for all $x \in R$

$$x \cdot 1 = 1 \cdot x = x.$$

This element is called the ***identity*** of R.

(5) $0 \neq 1$. (This implies R has at least two elements.)

(6) For any $x \in R$ there is a unique $-x \in R$ so that

$$x + (-x) = 0.$$

(This element is called the ***negative*** or ***additive inverse*** of x. And from now on we write $x + (-y)$ as $x - y$.) □

We will usually just refer to "the commutative ring R" rather than "the commutative ring $(R, +, \cdot)$". Also we will often be lazy and refer to R as just a "ring" rather than a "commutative ring"[1]. As in the case of fields we can view the positive integer n as an element of ring R by setting

$$n := \underbrace{1 + 1 + \cdots + 1}_{n \text{ terms}}$$

Then for negative n we can set $n := -(-n)$ where $-n$ is defined by the last equation. That is $5 = 1+1+1+1+1$ and $-5 = -(1+1+1+1+1)$.

1.1.1. *Inverses, units and associates.* While in a general ring it is not possible to divide by arbitrary nonzero elements (that is to say that arbitrary nonzero elements do not have inverses as division is defined in terms of multiplication by the inverse), it may happen that there are some elements that do have inverses and we can divide by these elements. We give a name to these elements.

Definition 1.2. Let R be a commutative ring. Then an element $a \in R$ is a ***unit*** or ***has an inverse*** iff there exists a $b \in R$ such that $ab = 1$. In this case we write $b = a^{-1}$ and call b the ***multiplicitive inverse*** (or just the ***inverse***) of a. □

Thus when talking about elements of a commutative ring saying that a is a unit just means a has an inverse. Note that inverses, if they exist, are unique. For if b and b' are inverses of a then $ab = ab' = 1$ which implies that $b' = b'1 = b'(ab) = (b'a)b = 1b = b$. Thus the notation a^{-1} is well defined. It is traditional, and useful, to give a name to elements a, b of a ring that differ by multiplication by a unit.

Definition 1.3. If a, b are elements of the commutative ring R then a and b are ***associates*** iff there is a unit $u \in R$ so that $b = ua$. □

[1]For those of you how can not wait to know: A non-commutative ring satisfies all of the above except that multiplication is no longer assumed commutative (that is it can hold that $xy \neq yx$ for some $x, y \in R$) and we have to add that both the left and right distributive laws $x(y + z) = xy + xz$ and $(y + z)x = yx + zx$ hold. A natural example of a non-commutative ring is the set of square $n \times n$ matrices over a field with the usual addition and multiplication.

Problem 1. Show that being associates is an equivalence relation on R. That is if $a \sim b$ is defined to mean that a and b are associates then show

(1) $a \sim a$ for all $a \in R$,
(2) that $a \sim b$ implies $b \sim a$, and
(3) $a \sim b$ and $b \sim c$ implies $a \sim c$. □

1.2. **Examples of rings.**

1.2.1. *The Integers.* The integers $\mathbf{Z}$ are as usual the numbers $0, \pm 1, \pm 2, \pm 3, \ldots$ with the addition and multiplication we all know and love. This is a good example to keep in mind when thinking about rings. In $\mathbf{Z}$ the only units (elements with inverses) are 1 and -1.

1.2.2. *The Ring of Polynomials over a Field.* Let $\mathbf{F}$ be a field and let $\mathbf{F}[x]$ be the set of all polynomials

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

where $a_0, \ldots, a_n \in \mathbf{F}$ and $n = 0, 1, 2, \ldots$. These are added, subtracted, and multiplied in the usual manner. This is the example that will be most important to us, so we review a little about polynomials. First if $p(x)$ is not the zero polynomial and $p(x)$ is as above with $a_n \neq 0$ then n is the ***degree*** of $p(x)$ and this will be denoted by $n = \deg p(x)$. The the nonzero constant polynomials a have degree 0 and we do not assign any degree to the zero polynomial. If $p(x)$ and $q(x)$ are nonzero polynomials then we have

$$\deg(p(x)q(x)) = \deg(p(x)) + \deg(q(x)).$$

Also if given $p(x)$ and $f(x)$ with $p(x)$ not the zero polynomial we can divide[2] $p(x)$ into $f(x)$. That is there are unique polynomials $q(x)$ (the ***the quotient***) and $r(x)$ (the ***the reminder***) so that

$$f(x) = q(x)p(x) + r(x) \quad \text{where} \quad \begin{cases} \deg r(x) < \deg p(x) \text{ or} \\ r(x) \text{ is the zero polynomial.} \end{cases}$$

This is the ***division algorithm*** for polynomials. If $p(x) = x - a$ for some $a \in \mathbf{F}$ this becomes

$$f(x) = q(x)(x - a) + r \quad \text{where } r \in \mathbf{F}.$$

By letting $x = a$ in this equation we get the fundamental:

[2]Here we are using the word "divide" in a sense other than "multiplying by the inverse". Rather we mean "find the quotient and remainder". I will continue to use the word "divide" in both these senses and trust it is clear from context which meaning is being used.

Proposition 1.4 (Remainder Theorem)**.** *If $x-a$ is divided into $f(x)$ then the remainder is $r=f(a)$. If particular $f(a)=0$ if and only if $x-a$ divides $f(x)$. That is $f(a)=0$ iff $f(x)=(x-a)q(x)$ for some polynomial $q(x)$ with* $\deg q(x)=\deg f(x)-1$. $\square$

I am assuming that you know how to add, subtract and multiply polynomials, and that given $f(x)$ and $p(x)$ with $p(x)$ not the zero polynomial that you can divide $p(x)$ into $f(x)$ and find the quotient $q(x)$ and remainder $r(x)$.

Problem 2. Show that the units in $R:=\mathbf{F}[x]$ are the nonzero constant polynomials. $\square$

The following shows that in our standard examples of rings, the integers $\mathbf{Z}$ and the polynomials over a field $\mathbf{F}[x]$, that if two elements are associates then they are very closely related.

Proposition 1.5. *In the ring of integers $\mathbf{Z}$ two elements a and b are associate if and only if $b=\pm a$. In the ring $\mathbf{F}[x]$ of polynomials over a field two polynomials $f(x)$ and $g(x)$ are associate if and only if there is a constant $c\neq 0$ so that $g(x)=cf(x)$.*

Problem 3. Prove this. $\square$

1.2.3. *The Integers Modulo n.* This is an example you will have likely seen before. But it is worth reviewing as it shows that rings can be quite different than the basic example of the integers and the polynomials over a field. Basically this is a generalization of the example of finite fields. Let $n>1$ be an integer and let $\mathbf{Z}/n$ be the integers reduced modulo n. That is we consider two integers x and y to be "equal" (really congruent modulo n) if and only if they have the same remainder when divided by n in which case we write $x\equiv y \mod n$. Therefore $x\equiv y \mod n$ if and only if $x-y$ is evenly divisible by x. It is easy to check that

$$\begin{gathered}x_1\equiv y_1 \mod n \quad\text{and}\quad x_2\equiv y_2 \mod n \quad\text{implies}\\ x_1+y_2\equiv x_1+y_2 \mod n \quad\text{and}\quad x_1x_2\equiv y_1y_2 \mod n.\end{gathered}$$

Then $\mathbf{Z}/n$ is the set of congruence classes modulo n. It only takes a little work to see that with the "obvious" choice of addition and multiplication that $\mathbf{Z}/n$ satisfies all the conditions of a commutative ring. Show this yourself as an exercise.) Here is the case $n=6$ in detail. The possible remainders when a number is divided by 6 are 0, 1, 2, 3, 4, 5. Thus we can use for the elements of $\mathbf{Z}/6$ the set $\{0,1,2,3,4,5\}$. Addition works like this. $3+4=1$ in $\mathbf{Z}/6$ as the remainder of $4+3$ when divided by 6 is 1. Likewise $2\cdot 4=2$ in $\mathbf{Z}/6$ as the remainder of

$2 \cdot 4$ when divided by 6 is 2. Here are the addition and multiplication tables for $\mathbf{Z}/6$

$+$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	4	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

$\cdot$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

This is an example of a ring with ***zero divisors***, that is nonzero elements a and b so that $ab = 0$. For example in $\mathbf{Z}/6$ we have $3 \cdot 4 = 0$. This is different from what we have seen in fields where $ab = 0$ implies $a = 0$ or $b = 0$. We also see from the multiplication table that the units in $\mathbf{Z}/6$ are 1 and 5. In general the units of $\mathbf{Z}/n$ are the correspond to the numbers x that are relatively prime to n.

1.3. **Ideals and quotient rings.** We have formed quotients of vector spaces by subspaces, now we want to form quotients of rings. When forming a quotient ring R/I the natural object I to quotient out by is not a subring, but an ideal.

Definition 1.6. Let R be a commutative ring. Then a nonempty subset $I \subset R$ is an ***ideal*** if and only if it is closed under addition and multiplication by elements of R. That is

$$a, b \in I \quad \text{implies} \quad a + b \in I$$

(this is closure under addition) and

$$a \in I,\ r \in R \quad \text{implies} \quad ar \in I$$

(this is closure under multiplication by elements of R). $\square$

1.3.1. *Principal ideas and generating ideals by elements of the ring.* There are two trivial examples of ideals in any R. The set $I = \{0\}$ is an ideal as is $I = R$. While it is possible to give large numbers of other examples of ideals in various rings for this course the most important example (cf. Theorem 2.7) is given by the following example:

Proposition 1.7. *Let R be a commutative ring and let $a \in R$. Let $\langle a \rangle$ be the set of all multiples of a by elements of R. That is*

$$\langle a \rangle := \{ra : r \in R\}.$$

Then $I := \langle a \rangle$ is an ideal in R.

Problem 4. Prove this. $\square$

Definition 1.8. If R is a commutative ring and $a \in R$, then $\langle a \rangle$ as defined in the last theorem is the ***principal ideal*** defined generated by a. $\square$

Proposition and Definition 1.9. *If R is a commutative ring and $a_1, a_2, \ldots, a_k \in R$, then set*

$$\langle a_1, a_2, \ldots, a_k \rangle = \{r_1 a_1 + r_2 a_2 + \cdots + r_k a_k : r_1, r_2, \ldots, r_k \in R\}.$$

*Then $\langle a_1, , \ldots, a_k \rangle$ is an ideal in R called the **ideal generated by** $a_1, \ldots, a_k$. (Thus the idea generated by $a_1, a_2, \ldots, a_k$ is the set of linear combinations of $a_1, a_2, \ldots, a_k$ with coefficients r_i from R.)*

Problem 5. Prove this. $\square$

1.3.2. *The quotient of a ring by an ideal.* Given a ring R and an ideal I in R then we will form a quotient ring R/I, which is defined in almost exactly the same way that we defined quotient vector spaces. You might want to review the problem set on quotients of a vector space by a subspace.

Let R be a ring and I and ideal in R. Define an equivalence relation $\equiv \mod I$ on R by

$$a \equiv b \mod I \quad \text{if and only if} \quad b - a \in I.$$

Problem 6. Show that this is an equivalence relation. This means you need to show that $a \equiv a \mod I$ for all $a \in R$, that $a \equiv b \mod I$ implies $b \equiv a \mod I$, and $a \equiv b \mod I$ and $b \equiv c \mod I$ implies $a \equiv b \mod I$. (If you want to make this look more like the notation we used in dealing quotients of vector spaces and write $a \sim b$ instead of $a \equiv b \mod I$ that is fine with me.) $\square$

Denote by $[a]$ the equivalence class of $a \in R$ under the equivalence relation $\sim_I$. That is

$$[a] := \{b \in R : b \equiv a \mod I\} = \{b \in R : b - a \in I\}.$$

Problem 7. Show $[a] = a + I$ where $a + I = \{a + r : r \in I\}$. $\square$

Let R/I be the set of all equivalence classes of $\sim_I$. That is

$$R/I := \{[a] : a \in R\} = \{a + I : a \in R\}.$$

The equivalence class $[a] = a + I$ is the ***coset of*** a ***in*** R. The following relates this to a case you are familiar with.

Problem 8. Let $R = \mathbf{Z}$ be the ring of integers and for $n \geq 2$ let I be the ideal $\langle n \rangle = \{an : a \in \mathbf{Z}\}$. Then show, with the notation of Section 1.2.3, that for $a, b \in \mathbf{Z}$

$$a \equiv b \mod n \qquad \text{if and only if} \qquad a \equiv b \mod I.$$

□

Exactly analogous to forming the ring $\mathbf{Z}/n$ or forming the quotient of a vector space V/W by a subspace we define a sum and multiplication of elements of elements of R/I by

$$[a] + [b] = [a + b], \quad \text{and} \quad [a][b] = [ab].$$

Problem 9. Show this is well defined. This means you need to show

$$[a] = [a'] \text{ and } [b] = [b'] \quad \text{implies} \quad [a + b] = [a' + b'] \text{ and } [ab] = [a'b'].$$

□

Theorem 1.10. *Assume that $I \neq R$. Then with this product R/I is a ring. The zero element of R/I is $[0]$ and the multiplicative identity of R/I is $[1]$.*

Proof. We first show that addition is commutative and associative in R/I. This will follow from the corresponding facts for addition in R.

$$\begin{aligned}[a] + ([b] + [c]) &= [a] + ([b + c]) = [a + (b + c)] \\ &= [(a + b) + c] = [a + b] + [c] = ([a] + [b]) + [c]\end{aligned}$$

and

$$[a] + [b] = [a + b] = [b + a] = [b] + [a].$$

The same calculation works for multiplication

$$[a]([b][c]) = [a]([bc]) = [a(bc)] = [(ab)c] = [ab][c] = ([a][b])[c]$$

and

$$[a][b] = [ab] = [ba] = [b][a].$$

So both addition and multiplication are associative in R/I.

For any $[a] \in R/I$ we have

$$[a] + [0] = [a + 0] = [a] = [0 + a] = [0] + [a]$$

and therefore $[0]$ the zero element of R/I. Likewise

$$[a][1] = [a1] = [a] = [1a] = [1][a]$$

so that $[1]$ is the multiplicative identity of R/I. Finally all that is left is to show that every $[a]$ has an additive inverse. To no one's surprise this is $[-a]$. To see this note

$$[a] + [-a] = [a - a] = [0] = [-a + a] = [-a] + [a].$$

Thus $-[a] = [-a]$. Finally there is the distributive law. Again this just follows from the distributive law in R:

$$[a]([b]+[c]) = [a][b+c] = [a(b+c)] = [ab+ac] = [ab]+[ac] = [a][b]+[a][c].$$

We still have not used that $I \neq R$ and still have not shown that $[0] \neq [1]$. But $[1] = [0]$ if and only if $1 \in I$ so we need to show that $1 \notin I$. Assume, toward a contradiction, that $1 \in I$. Then for any $a \in R$ we have $a = a1 \in I$ as I is closed under multiplication by elements from R. But then $R \subseteq I \subseteq R$ contradicting that $I \neq R$. This completes the proof. □

If R is a commutative ring and I and ideal in R then it is important to realize that if $a \in I$ then $[a] = [0]$ in R/I. This is obvious from the definition of R/I, but still should be kept in mind when working with quotient rings. Here is an example both a quotient ring and of why keeping this in mind is useful.

Let $R = \mathbf{R}[x]$ be the polynomials with coefficients in the real numbers $\mathbf{R}$. Let $q(x) = x^2+1$ and let $I = \langle q(x)\rangle$ be the ideal of all multiples of $q(x) = x^2 + 1$. That is

$$I = \{(x^2 + 1)f(x) : f(x) \in \mathbf{R}[x]\}.$$

Clearly $x^2+1 = 1(x^2+1) \in I$. Therefore in the ring $R/I = \mathbf{R}[x]/\langle x^2+1\rangle$ we have that $[x^2 + 1] = [0]$. Therefore

$$[0] = [x^2 + 1] = [x^2] + [1] = [x]^2 + [1].$$

that is $[x]^2 = -[1]$. Thus $-[1]$ has a square root in R/I. With a little work you can show that R/I is just the complex numbers dressed up a bit. (See Problem 21, p. 18)

1.4. **A condition for one ideal to contain another.** The results here are elementary, but a little notationally messy. They will not be used until Section 4.7 so there is no harm for a reader not yet very comfortable with the notion of an ideal in a ring to skip this section until it is needed.

Let R be a commutative ring. Let $\{a_1, a_2, \ldots, a_m\} \subset R$ and $\{b_1, b_2, \ldots, b_n\} \subset R$ be two non-empty of elements from R. We wish to understand when the two ideals (See Proposition and Definition 1.9)

$$\langle a_1, a_2, \ldots, a_m\rangle, \qquad \langle b_1, b_2, \ldots, b_n\rangle$$

are equal, or more generally when one contains the other.

Definition 1.11. We say that each element of $\{a_1, a_2, \ldots, a_m\}$ is a ***linear combination of elements of*** $\{b_1, b_2, \ldots, b_n\}$ iff there are elements $r_{ij} \in R$ with $1 \le i \le m$ and $1 \le j \le n$ and so that

$$a_i = \sum_{j=1}^{n} r_{ij} b_j \quad \text{for} \quad 1 \le i \le m.$$

□

Proposition 1.12. *Let* $\{a_1, a_2, \dots, a_m\} \subset R$ *and* $\{b_1, b_2, \dots, b_n\} \subset R$ *be non-empty. Then*

$$\langle a_1, a_2, \dots, a_m \rangle \subseteq \langle b_1, b_2, \dots, b_n \rangle$$

if and only if each element of $\{a_1, a_2, \dots, a_m\}$ *is a linear combination of elements of* $\{b_1, b_2, \dots, b_n\}$.

Proof. Fist assume that $\langle a_1, a_2, \dots, a_m \rangle \subseteq \langle b_1, b_2, \dots, b_n \rangle$. Then as $\langle a_1, a_2, \dots, a_m \rangle$ contains each element a_i we have that $a_i \in \langle b_1, b_2, \dots, b_n \rangle$. Therefore, but the definition of $\langle b_1, b_2, \dots, b_n \rangle$, there are elements $r_{i1}, r_{i2}, \dots, r_{in}$ such that

$$a_i = r_{i1} b_1 + r_{i2} b_2 + \cdots + r_{in} b_n = \sum_{j=1}^{n} r_{ij} b_j.$$

This shows that each element of $\{a_1, a_2, \dots, a_m\}$ is a linear combination of elements of $\{b_1, b_2, \dots, b_n\}$.

Conversely if each element of $\{a_1, a_2, \dots, a_m\}$ is a linear combination of elements of $\{b_1, b_2, \dots, b_n\}$. That is $a_i = \sum_{j=1}^{n} r_{ij} b_j$ with $r_{ij} \in R$. Let $x \in \langle a_1, a_2, \dots, a_m \rangle$. By definition of $\langle a_1, a_2, \dots, a_m \rangle$ this implies there are $c_1, c_2, \dots, c_m \in R$ with

$$x = c_1 a_1 + c_2 a_2 + \cdots + c_m a_m = \sum_{i=1}^{m} c_i a_i.$$

Then we expand the a_i's in terms of the b_j's and interchanging the order of summation we have

$$x = \sum_{i=1}^{m} c_i a_i = \sum_{i=1}^{m} c_i \sum_{j=1}^{n} r_{ij} b_j = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} c_i r_{ij} \right) b_j = \sum_{j=1}^{n} s_j b_j$$

where

$$s_j = \sum_{i=1}^{m} c_i r_{ij}.$$

But then by the definition of $\langle b_1, b_2, \dots, b_n \rangle$ this implies $x \in \langle b_1, b_2, \dots, b_n \rangle$. As x was any element of $\langle a_1, a_2, \dots, a_m \rangle$ this implies $\langle a_1, a_2, \dots, a_m \rangle \subseteq \langle b_1, b_2, \dots, b_n \rangle$ and completes the proof. □

Corollary 1.13. *Let* $\{a_1, a_2, \dots, a_m\} \subset R$ *and* $\{b_1, b_2, \dots, b_n\} \subset R$ *be non-empty. Then*

$$\langle a_1, a_2, \dots, a_m \rangle = \langle b_1, b_2, \dots, b_n \rangle$$

if and only if each element of $\{a_1, a_2, \dots, a_m\}$ *is a linear combination of elements of* $\{b_1, b_2, \dots, b_n\}$ *and each element of* $\{b_1, b_2, \dots, b_n\}$ *is a linear combination of* $\{a_1, a_2, \dots, a_m\}$.

Problem 10. Prove this. □

2. Euclidean Domains.

2.1. **The definition of Euclidean domain.** As we said above for us the most important examples of rings are the ring of integers and the ring of polynomials over a field. We now make a definition that captures many of the basic properties these two examples have in common.

Definition 2.1. A commutative ring R is a ***Euclidean domain*** iff

(1) R has no zero divisors[3]. That is if $a \neq 0$ and $b \neq 0$ then $ab \neq 0$. (Or in the contrapositive form $ab = 0$ implies $a = 0$ or $b = 0$.)
(2) There is a function $\delta\colon (R \setminus \{0\}) \to \{0, 1, 2, 3, \dots\}$ (that is δ maps nonzero elements of R to nonnegative integers) so that
   (a) If $a, b \in R$ are both nonzero then $\delta(a) \leq \delta(ab)$.
   (b) The ***division algorithm*** holds in the sense that if $a, b \in R$ and $a \neq 0$ then we can divide a into b to get a ***quotient*** q and a ***reminder*** r so that

$$b = aq + r \quad \text{where } \delta(r) < \delta(a) \text{ or } r = 0$$ □

2.2. **The Basic Examples of Euclidean Domains.** Our two basic examples of Euclidean domains are the integers $\mathbf{Z}$ with $\delta(a) = |a|$, the absolute value of a and $\mathbf{F}[x]$, the ring of polynomials over a field $\mathbf{F}$ with $\delta(p(x)) = \deg p(x)$. We record this as theorems:

Theorem 2.2. *The integers* $\mathbf{Z}$ *with* $\delta(a) := |a|$ *is a Euclidean domain.*

Theorem 2.3. *The ring of polynomials* $\mathbf{F}[x]$ *over a field* $\mathbf{F}$ *with* $\delta(p(x)) = \deg p(x)$ *is a Euclidean domain.*

Proofs. These follow from the usual division algorithms in $\mathbf{Z}$ and $\mathbf{F}[x]$. □

Remark 2.4. The example of the integers shows that the quotient q and remainder r need not be unique. For example in $R = \mathbf{Z}$ let $a = 4$ and $b = 26$. Then we can write

$$26 = 4 \cdot 6 + 2 = 4q_1 + r_1 \quad \text{and} \quad 26 = 4 \cdot 7 + (-2) = 4q_2 + r_2.$$

In number theory sometimes the extra requirement that $r \geq 0$ is made and then the quotient and remainder are unique. □

[3]In general a commutative ring R with no zero divisors is called an ***integral domain*** or just a ***domain***

2.3. **Primes and factorization in Euclidean domains.** We now start to develop the basics of "number theory" in Euclidean domains. By this is meant that we will show that it is possible to define things like "primes" and "greatest common divisors" and show that they behave just as in the case of the integers. Many of the basic facts about Euclidean domains are proven by starting with subset S of the Euclidean domain in question and then choosing an element a in S that minimizes $\delta(a)$. While it is more or less obvious that it is always possible to do this we record (without proof) the result that makes it all work.

Theorem 2.5 (Axiom of Induction)**.** *Let* $\mathbf{N} := \{0, 1, 2, 3, \dots\}$ *be the natural numbers (which is the same thing as the nonnegative integers). Then any nonempty subset* S *of* $\mathbf{N}$ *has a smallest element.* □

2.3.1. *Divisors, irreducibles, primes, and great common divisors.* We start with some elementary definitions:

Definition 2.6. Let R be a commutative ring. Let $a, b \in R$.

(1) Then a is a ***divisor*** of b, (or a ***divides*** b, or a is a ***factor*** of b) iff there is $c \in R$ so that $b = ca$. This is written as $a \mid b$.
(2) b is a ***multiple*** of a iff a divides b. That is iff there is $c \in R$ so that $b = ac$.
(3) The element $b \neq 0$ is a ***prime***[4], also called an ***irreducible***, iff b is not a unit and if $a \mid b$ then either a is a unit, or $a = ub$ for some unit $u \in R$.
(4) The element c of R is a ***greatest common divisor*** of a and b iff $c \mid a$, $c \mid b$ and if $d \in R$ is any other element of R that divides both a and b then $d \mid c$. (Note that greatest common divisors are not unique. For example in the integers $\mathbf{Z}$ there both 4 and -4 are greatest common divisors of 12 and 20, while in the polynomial ring $\mathbf{R}[x]$ the element the $c(x - 1)$ is a greatest common divisor of $x^2 - 1$ and $x^2 - 3x + 2$ for any $c \neq 0$.)
(5) The elements a and b are ***relatively prime*** iff 1 is a greatest common divisor of a and b. Or what is the same thing the only elements that divide both a and b are units. □

2.3.2. *Ideals in Euclidean domains.* There are commutative rings where some pairs of elements do not have any greatest common divisors. We now show that this is not the case in Euclidean domains.

[4] I have to be honest and remark that this is not the usual definition of a prime in a general ring, but is the usual definition of an irreducible. Usually a prime is defined by the property of Theorem 2.10. In our case (Euclidean domains) the two definitions turn out to be the same.

Theorem 2.7. *Let R be a Euclidean domain. Then every ideal in R is principle. That is if I is an ideal in R then there is an $a \in R$ such that $I = \langle a \rangle$. Moreover if $\{0\} \neq I = \langle a \rangle = \langle b \rangle$ then $a = ub$ for some unit u.*

Problem 11. Prove this along the following lines:

(1) By the Axiom of induction, Theorem 2.5, the set $S := \{\delta(r) : r \in I, r \neq 0\}$ has a smallest element. Let a be a nonzero element of I that minimizes $\delta(r)$ over nonzero elements of I. Then for any $b \in I$ show that there is a $q \in R$ with $b = aq$ by showing that if $b = aq + r$ with $r = 0$ or $\delta(r) < \delta(a)$ (such q and r exist by the definition of Euclidean domain) than in fact $r = 0$ so that $b = qa$.
(2) With a as in the last step show $I = \langle a \rangle$, and thus conclude I is principle.
(3) If $\langle a \rangle = \langle b \rangle$ then $a \in \langle b \rangle$ so there is a c_1 so that $a = c_1 b$. Likewise $b \in \langle a \rangle$ implies there is a $c_2 \in R$ such that $b = c_2 a$. Putting these together implies $a = c_1 c_2 a$. Show this implies $c_1 c_2 = 1$ so that c_1 and c_2 are units. *Hint:* Use that $a(1 - c_1 c_2) = 0$ and that in a Euclidean domain there are no zero divisors. $\square$

Theorem 2.8. *Let R be a Euclidean domain and let a and b be nonzero elements of R. Then a and b have at least one greatest common divisor. More over if c and d are both greatest common divisors of a and b then $d = cu$ for some unit $u \in R$. Finally if c is any greatest common divisor of a and b then there are elements $x, y \in R$ so that*

$$c = ax + by.$$

Problem 12. Prove this as follows:

(1) Let $I := \{ax + by : x, y \in R\}$. Show that I is an ideal of R.
(2) Because I is an ideal by the last theorem the ideal I is principal so $I = \langle c \rangle$ for some $c \in R$. Show that c is a greatest common divisor of a and b and that $c = ax + by$ for some $x, y \in R$. *Hint:* That $c = ax + by$ for some $x, y \in R$ follows from the definition of I. From this show c is a greatest common divisor of a and b.
(3) If c and d are both greatest common divisors of a and b then by definition $c \mid d$ and $d \mid c$. Use this to show $d = uc$ for some unit u. $\square$

Theorem 2.9. *Let R be a Euclidean domain and let $a, b \in R$ be relatively prime. Then there exist $x, y \in R$ so that*

$$ax + by = 1.$$

Problem 13. Prove this as a corollary of the last theorem. □

Theorem 2.10. *Let R be a Euclidean domain and let $a, b, p \in R$ with p prime. Assume that $p \mid ab$. Then $p \mid a$ or $p \mid b$. That is if a prime divides a product, then it divides one of the factors.*

Problem 14. Prove this by showing that if p does not divide a then it must divide b. Do this by showing the following:

(1) As p is prime and we are assuming p does not divide a then a and p are relatively prime.
(2) There are x and y in R so that $ax + py = 1$.
(3) As $p \mid ab$ there is a $c \in R$ with $ab = cp$. Now multiply both sides of $ax + py = 1$ by b to get $abx + pby = b$ and use $ab = cp$ to conclude p divides b. □

Corollary 2.11. *If p is a prime in the Euclidean domain R and p divides a product $a_1 a_2 \cdots a_n$ then p divides at least one of a_1, a_2, …, a_n.*

Proof. This follows from the last proposition by a straightforward induction. □

2.3.3. *Units and associates in Euclidean domains.*

Lemma 2.12. *Let R be a Euclidean domain. Then a nonzero element a of R is a unit iff $\delta(a) = \delta(1)$.*

Problem 15. Prove this. *Hint:* First note that if $0 \neq r \in R$ then $\delta(1) \leq \delta(1r) = \delta(r)$. Now use the division algorithm to write $1 = aq + r$ where either $\delta(r) < \delta(a) = \delta(1)$ or $r = 0$. □

Proposition 2.13. *Let R be a Euclidean domain and a and b nonzero elements of R. If $\delta(ab) = \delta(a)$ then b is a unit (and so a and ab are associates).*

Problem 16. Prove this. *Hint:* Use the division algorithm to divide ab into a. That is there are q and $r \in R$ so that $a = (ab)q + r$ so that either $r = 0$ or $\delta(r) < \delta(a)$. Then write $r = a(1 - bq)$ and use that if x and y are nonzero $\delta(x) \leq \delta(xy)$ to show $(1 - bq) = 0$. From this show b is a unit. □

2.3.4. *The Fundamental Theorem of Arithmetic in Euclidean domains.*

Theorem 2.14 (Fundamental Theorem of Arithmetic)**.** *Let a be a non-zero element of a Euclidean domain that is not a unit. Then a is a product $a = p_1 p_2 \cdots p_n$ of primes $p_1, p_2, \ldots, p_n$. Moreover we have the following uniqueness. If $a = q_1 q_2 \cdots q_m$ is another expression of a as a* *product of primes, then $m = n$ and after a reordering of $q_1, q_2, \ldots, q_n$ there are units $u_1, u_2, \ldots, u_n$ so that $q_i = u_i p_i$ for $i = 1, \ldots, n$.*

Problem 17. Prove this by induction on $\delta(a)$ in the following steps.

(1) As a is not a unit the last lemma implies $\delta(a) > \delta(1)$. Let $k := \min\{\delta(r) : r \in R, \delta(r) > \delta(1)\}$. Show that if $\delta(a) = k$ then a is a prime. (This is the base of the induction.)

(2) Assume that $\delta(a) = n$ and that it has been shown that for any $b \neq 0$ with $\delta(b) < n$ that either b is a unit or b is a product of primes. Then show that a is a product of primes. *Hint:* If a is prime then we are done. Thus it can be assumed that a is not prime. In this case $a = bc$ where b and c are not units. a is a product $a = bc$ with both b and c not units. By the last proposition this implies $\delta(b) < \delta(a)$ and $\delta(c) < \delta(a)$. So by the induction hypothesis both b and c are products of primes. This shows $a = bc$ is a product of primes.

(3) Now show uniqueness in the sense of the statement of the theorem. Assume $a = p_1 p_2 \cdots p_n = q_1 q_2 \cdots q_m$ where all the p_i's and q_j's are prime. Then as p_1 divides the product $q_1 q_2 \cdots q_m$ by Corollary 2.11 this means that p_1 divides at least one of $q_1, q_2, \ldots, q_m$. By reordering we can assume that p_1 divides q_1. As both p_1 and q_1 are primes this implies $q_1 = u_1 p_1$ for some unit u_1. Continue in this fashion to complete the proof. □

2.3.5. *Some related results about Euclidean domains.*

2.3.5.1. The greatest common divisor of more than two elements. We will need the generalization of the greatest common divisor of a pair $a, b \in R$ for the greatest common divisor of a finite set $a_1, \ldots, a_k$. This is straightforward to do

Definition 2.15. Let R be commutative ring and $a_1, \ldots, a_k \in R$.

(1) The element c of R is a ***greatest common divisor*** of $a_1, \ldots, a_k$ iff c divides all of the elements $a_1, \ldots, a_k$ and if d is any other element of R that divides all of $a_1, \ldots, a_k$, then $d \mid c$.

(2) The elements $a_1, \ldots, a_k$ are ***relatively prime*** iff 1 is a greatest common divisor of $a_1, \ldots, a_k$. □

Note that have $a_1, \ldots, a_k$ relatively prime does not imply that they are pairwise elementary relatively prime. For example when the ring is $R = \mathbf{Z}$ the integers, the $6 = 2 \cdot 3$, $10 = 2 \cdot 5$ and $15 = 3 \cdot 5$ are relatively prime, but no pair of them is.

Theorem 2.16. *Let R be a Euclidean domain and let $a_1, \ldots, a_k$ be nonzero elements of R. Then $a_1, \ldots, a_k$ have at least one greatest common divisor. More over if c and d are both greatest common divisors of*

$a_1, \dots, a_k$ then $d = cu$ for some unit $u \in R$. Finally if c is any greatest common divisor of $a_1, \dots, a_k$ then there are elements $x_1, \dots, x_k \in R$ so that

$$c = a_1x_1 + a_2x_2 + \cdots + a_kx_k.$$

Moreover the greatest common divisor c is the generator of the ideal $\langle a_1, a_2, \dots, a_k\rangle$ of R.

Problem 18. Prove this as follows:

(1) Let $I := \langle a_1, a_2, \dots, a_k\rangle = \{a_1x_1 + a_2x_2 + \cdots a_kx_k : x_1, \dots, x_k \in R\}$. Then show that I is an ideal of R.
(2) Because I is an ideal by Theorem 2.7 the ideal I is principal so $I = \langle c\rangle$ for some $c \in R$. Show that c is a greatest common divisor of $a_1, a_2, \dots, a_k$ and that $c = a_1x_1 + a_2x_2 + \cdots + a_kx_k$ for some $x_1, x_2, \dots, x_k \in R$. *Hint:* That $c = a_1x_1 + a_2x_2 + \cdots + a_kx_k$ for some $x_1, x_2, \dots, x_k \in R$ follows from the definition of I. From this show c is a greatest common divisor of $a_1, \dots, a_k$.
(3) If c and d are both greatest common divisors of $a_1, \dots, a_k$ then by definition $c \mid d$ and $d \mid c$. Use this to show $d = uc$ for some unit u and that the principal ideals $\langle c\rangle$ and $\langle d\rangle$ are equal. □

Theorem 2.17. *Let R be a Euclidean domain and let $a_1, \dots, a_k \in R$ be relatively prime. Then there exist $x_1, \dots, x_k \in R$ so that*

$$a_1x_1 + a_2x_2 + \cdots + a_kx_k = 1.$$

Problem 19. Prove this as a corollary of the last theorem. □

2.3.5.2. Euclidean Domains modulo a prime are fields. We finish this section with a method for constructing fields.

Theorem 2.18. *Let R be a Euclidean domain and let $p \in R$ be a prime. Then the quotient ring $R/\langle p\rangle$ is a field. (As usual $\langle p\rangle = \{ap : a \in R\}$ is the ideal of all multiples of p.)*

Problem 20. Prove this along the following lines: As $R/\langle p\rangle$ is a ring to show that it is a field we only need to show that each $[a] \in R/\langle p\rangle$ with $[a] \neq [0]$ has a multiplicative inverse. So let $[a] \neq [0]$ and show that $[a]$ has a multiplicative inverse along the following lines.

(1) First show that p and a are relatively prime. *Hint:* As $[a] \neq [0]$ in $R/\langle p\rangle$ we see that a is not a multiple of p. But p is prime so this implies that 1 is a greatest common divisor of p and a.
(2) Show there are $x, y \in R$ so that $ax + py = 1$.
(3) Show for this x that $[a][x] = [1]$ so that $[x]$ is the multiplicative inverse of $[a]$ in $R/\langle p\rangle$. *Hint:* From $ax + py = 1$ we have $[ax + py] = [1]$. But $py \in \langle p\rangle$ so $[py] = [0]$. □

Problem 21. As an application of Theorem (2.18) let $R = \mathbf{R}[x]$ (where $\mathbf{R}$ is the field of real numbers). Then let $p(x) = x^2 + 1$. This is irreducible in R and thus prime. Let $F = \mathbf{R}[x]/\langle p(x)\rangle$. Then show that F is a copy of the complex numbers $\mathbf{C}$ by showing the following

(1) If $[x]$ is the coset of x in $F = \mathbf{R}[x]/\langle p(x)\rangle$, then $[x]^2 = -1$. *Hint:* As $p(x) = x^2 + 1 \in \langle p(x)\rangle$ we have $[x^2+1] = 0$ in F. But then $[x]^2 + 1 = [x^2+1] = 0$.

(2) Show that every element of F is of the form $a + b[x]$ with $a, b \in \mathbf{R}$. *Hint:* Let $[f(x)] \in F$. Then divide x^2+1 into $f(x)$ to get $f(x) = q(x)(x^2+1) + r(x)$ where $r(x) = 0$ or $\deg r(x) < 2 = \deg(x^2+1)$. Thus $r(x)$ is of the form $r(x) = a + bx$. Therefore

$$\begin{aligned} [f(x)] &= [q(x)(x^2+1) + r(x)] \\ &= [q(x)(x^2+1)] + [a + bx] \\ &= 0 + a + b[x] \\ &= a + b[x]. \end{aligned}$$

as $q(x)(x^2+1) \in \langle p(x)\rangle$ and so $[q(x)(x^2+1)] = 0$.

(3) Thus elements of F are of the form $a + b[x]$ where $[x]^2 = -1$. That is F is a copy of $\mathbf{C}$. □

2.4. Ring homomorphisms and the first isomorphism theorem.

Definition 2.19. Let R_1 and R_2 be rings. Then a map $\varphi\colon R_1 \to R_2$ is a ***homomorphism*** iff

$$\varphi(a+b) = \varphi(a) + \varphi(b), \qquad \varphi(ab) = \varphi(a)\varphi(b).$$

and $\varphi(1) = \varphi(1)$ (this means that φ maps the identity of R_1 to the identity of R_2.) An isomorphism of rings that is bijective (that is one to one and onto) is a bi ring isomorphism. □

Proposition 2.20. *Let $\varphi\colon R_1 \to R_2$ be a homomorphism of rings. Then the **kernel**,*

$$\ker(\varphi) := \{r \in R_1 : \varphi(r) = 0\}$$

*is a and ideal of R_1 and the **image***

$$\operatorname{Image}(\varphi) = \varphi[R_1] := \{\varphi(r) : r \in R_1\}$$

is a subring of R_2.

Proposition 2.21. *Prove this.* □

Theorem 2.22 (Frist homomorphism theorem for rings)**.** *Let $\varphi\colon R_1 \to R_2$ be a homomorphism of rings. Then the image, $\varphi[R_1]$, is isomorphic to the quotient ring $R_1/\ker(\varphi)$. Thus if φ is surjective, then R_2 is isomorphic to $R_1/\ker(\varphi)$.*

Problem 22. Prove this. *Hint:* Define $\tilde{\varphi} : R_1/\ker(\varphi) \to \varphi[R_2]$ by $\tilde{\varphi}(r + \ker(\varphi)) = \varphi(r)$. Show this is well defined, a homomorphism, and bijective. □

2.5. **Sums and products of ideals and theorems motivated by the Chinese Remainder Theorem.** This section is not required for the rest of these notes and can skipped. It is included because both because I like the results, and because the arguments show how some ideas (such as a pair of numbers being being relatively prime) from basic number theory generalize to more general rings (where we consider ideals, rather than elements, being relatively prime). And doing some well motivated calculations with ideals and quotient rings should make them more concrete.

Recall the Chinese Remainder Theorem of elementary number theory is

Theorem 2.23. *Let $m_1, \ldots, m_n$ be pairwise relatively prime positive integers. Then for any $a_1, \ldots, a_n \in \mathbf{Z}$ the set of simultaneous congruences*

$$\begin{aligned} x &\equiv a_1 \mod m_1 \\ x &\equiv a_2 \mod m_2 \\ &\vdots \qquad\quad \vdots \\ x &\equiv a_n \mod m_n \end{aligned}$$

has a solution. It is unique up to congruence modulo the product $m = m_1 m_2 \cdots m_n$. □

Problem 23. Use this to show that if $m_1, \ldots, m_n \in \mathbf{Z}$ are pairwise relatively prime and $m = m_1 m_2 \cdots m_n$ then $\mathbf{Z}_m$ is isomorphic to the product $\mathbf{Z}_{m_1} \times \cdots \times \mathbf{Z}_{m_n}$. (cf. Problem 2.23.) □

We now generalize the Chinese Remainder Theorem and some other concepts from elementary number theory to ideals in rings.

Definition 2.24. If I, J are ideals in the ring R the ***sum*** of I and J is

$$I + J := \{a + b : a \in I, b \in J\}.$$

and their ***product*** is

$$IJ := \text{set of all finite sums of propucts } ab \text{ with } a \in I \text{ and } b \in J.$$

(Explicitly elements of IJ are of the form $\sum_{j=1}^{n} a_j b_j$ for some $n \geq 1$ with $a_j \in I$ and $b_j \in J$.)

Proposition 2.25. *The sum, $I+J$, the product, IJ, and the intersection, $I \cap J$, of ideals I and J are ideals.*

Problem 24. Prove this. $\square$

Proposition 2.26. *Let R be an Euclidean domain (you will lose nothing by letting $R = \mathbf{Z}$; the important point being that all ideals are principal). Let $I = \langle a \rangle$ and $J = \langle b \rangle$ be the principal ideals generated by the elements a and b of R. Then*

$$\begin{aligned} IJ &= \langle ab \rangle \\ I \cap J &= \langle \gcd(a,b) \rangle \\ I + J &= \langle \operatorname{lcm}(a,b) \rangle. \end{aligned}$$

*Here $\gcd(a,b)$ is any greatest common divisor of a and b (cf. Def 2.6 and Thm 2.8) and $\operatorname{lcm}(a,b)$ is any **least common multiple** of a and b (that is any element c of R that is a common multiple of a and b (i.e. $a \mid c$, $b \mid c$) and if d is any other common multiple of a and b, then $c \mid d$. It is easy to see that least common multiples are unique up to multiplication by units.)*

Problem 25. Prove this. This includes showing in a Euclidean domain that least common multiples exist (*Hint:* Let $c = ab/\gcd(a,b)$ and show this is a least common multiple of a and b). $\square$

The concepts of sums and products of ideals can be generalized to more than two ideals.

Definition 2.27. Let $I_1, \ldots, I_n$ be ideals in the ring R. Then their ***sum*** is

$$I_1 + \cdots + I_n := \{a_1 + \cdots + a_n : a_j \in I_j \text{ for } 1 \le j \le n\}$$

and their ***product*** is

$$I_1 I_2 \cdots I_n := \text{set of all finite sums of products } a_1 a_2 \cdots a_n \text{ with } a_j \in I_j.$$

The power I^n of an ideal is $I^n = I_1 I_2 \cdots I_n$ where $I_j = I$ for all j. $\square$

If I and J are ideals then it is easy to see that $IJ \subseteq I \cap J$. The inequality can be strict. For example in the ring of integers, if $I = \langle 6 \rangle$ and $J = \langle 10 \rangle$ then $IJ = \langle 60 \rangle$ and $I \cap J = \langle 30 \rangle$. But if the ideals are relatively prime in the sense that $I + J = R$, then equality will hold. It is worth remarking that $I + J = R$ is equivalent being able to write 1 as a sum $1 = i + j$ with $i \in I$ and $j \in J$. In many of the proofs that follow, this is how the condition $I + J = R$ is used.

Problem 26. Let I and J be ideals of the commutative ring R with $I + J = R$. Then the following hold:

(1) $I \cap J = IJ$. *Hint:* As $I + J = R$ there are $i \in I$ and $j \in J$ with $i+j = 1$. Let $a \in I \cap J$. Then $a = 1a = (i+j)a = ia + ja$.
(2) For any positive integers m, n we have $I^m + J^n = R$ and thus $I^m \cap J^n = I^m J^n$. *Hint:* It is enough to show $1 \in I^m + J^n$. Let $1 = i + j$ with $i \in I$ and $j \in J$. Then by the binomial theorem, which is easily seen to hold in any commutative ring, $1 = 1^{m+n} = \sum_{k=0}^{m+n} \binom{n+m}{k} i^k j^{m+n-k}$ and for each k either $k \geq m$ or $n + m - k \geq n$. □

Problem 27. Let $I_1, I_2, \ldots, I_n$ be ideals in the commutative ring R such that

$$I_i + I_j = R \qquad \text{when} \qquad i \neq j.$$

Then

(1) The equality

$$I_1 I_2 \cdots I_{n-1} + I_n = R$$

holds. *Hint:* For each $k \in \{1, \ldots, n-1\}$ write $1 = i_k + a_k$ with $i_k \in I_k$ and $a_k \in I_k$. Then $1 = (i_1 + a_1) \cdots (i_{n-1} + a_{n-1}) = i_1 i_2 \cdots i_{n-1} + a$ with $a \in I_n$.
(2) The equality

$$I_1 \cap I_2 \cap \cdots \cap I_n = I_1 I_2 \cdots I_n.$$

holds. *Hint:* Use induction on n and Problem 26. □

The following is the generalization of the Chinese Remainder Theorem to general rings.

Proposition 2.28. *Let $I_1, I_2, \ldots, I_n$ be ideals in the commutative ring R that are pairwise relatively prime in the sense that*

$$I_i + I_j = R \qquad \textit{when} \qquad i \neq j.$$

Then

$$I_1 \cap I_2 \cap \cdots \cap I_n = I_1 I_2 \cdots I_n$$

and

$$R/\left(I_1 \cap I_2 \cap \cdots \cap I_n\right) = \bigoplus_{j=1}^{n} R/I_j.$$

(Or to make the analogy with the usual Chinese remainder theorem more precise, the equality $R/\left(I_1 \cap I_2 \cap \cdots \cap I_n\right) = \bigoplus_{j=1}^{n} R/I_j$ is equivalent to saying that for any $r_1, \ldots, r_n \in R$ there is $x \in R$ with $x \equiv r_j \mod I_j$ for $j = 1, \ldots, n$ and this x is unique up to congruence mod $I_1 I_2 \cdots I_n$.*)* □

Problem 28. Prove this. *Hint:* Some of this is covered by the last problem. Define a ring homomorphism by $\varphi\colon R \to \bigoplus_{j=1}^{n} R/I_j$ by $\varphi(r) = (r + I_1, r + I_2, \ldots, r + I_n)$. The kernel of φ is $I_1 \cap I_2 \cap \cdots \cap I_n = I_1 I_2 \cdots I_n$. By the first homomorphism theorem for modules Image(φ) is isomorphic to $R/\ker(\varphi)$. So it is enough to show that φ is surjective. For each $k \in \{1, \ldots, n\}$ let $J_k = I_1 I_2 \cdots I_{k-1} I_{k+1} \cdots I_n$ (that is the product $\prod_{j \neq k} I_j$). By Problem 27 $I_k + J_k = R$. Let $1 = i_k + j_k$ with $i_k \in I_k$ and $j_k \in J_k$. For $1 \leq k \leq n$ let $r_k \in R$ set $r = r_1 j_1 + \cdots + r_n j_n$ and show $\varphi(r) = (r_1 + I_1, r_2 + I_2, \ldots, r_n + I_n)$. $\square$

3. Matrices over a Ring.

In this section R will be any ring, but in the long run we will mostly need the results in the case that R is an Euclidean domain.

3.1. **Basic properties of matrix multiplication.**

A matrix with entries on a ring is defined just as in the case of fields.

Notation 3.1. If R is a ring let $M_{m\times n}(R)$ be the m by n matrices whose elements are in R. (This is m rows and n columns). Thus an element $A \in M_{m\times n}(R)$ is of the form

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

with $a_{ij} \in R$. $\square$

3.1.1. *Definition of addition and multiplication of matrices.* If $A \in M_{m\times n}(R)$ and $r \in R$ then A can be multiplied by a "scalar" $r \in R$ as rA is the matrix

$$rA := \begin{bmatrix} ra_{11} & ra_{12} & \cdots & ra_{1n} \\ ra_{21} & ra_{22} & \cdots & ra_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ ra_{m1} & ra_{m2} & \cdots & ra_{mn} \end{bmatrix}.$$

Likewise if $A, B \in M_{m\times n}(R)$ with A as above and

$$B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix}$$

then $A+B$ is the matrix with elements $(A+B)_{ij} = a_{ij} + b_{ij}$. If $A \in M_{m\times n}(R)$ and $B \in M_{n\times p}$, say

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \qquad B = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{np} \end{bmatrix},$$

then the product matrix is defined in the usual manner. That is the product AB is the m by p matrix with elements

$$(AB)_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}.$$

3.1.2. *The basic algebraic properties of matrix multiplication and addition.* The usual properties of matrix addition and multiplication hold with the usual proofs. We record this as:

Proposition 3.2. *Let R be a ring. Then*

(1) *For $r, s \in R$ and $A \in M_{m\times n}(R)$ the distributive law*
$$(r+s)A = rA + sA$$
holds.

(2) *For $r \in R$, and $A, B \in M_{m\times n}(R)$ the distributive law*
$$r(A+B) = rA + rB$$
holds.

(3) *If $A, B, C \in M_{m\times n}(R)$ then*
$$(A+B)+C = A+(B+C).$$

(4) *If $r, s \in R$ and $A \in M_{m\times n}(R)$ then*
$$r(sA) = (rs)A.$$

(5) *If $r \in R$, $A \in M_{m\times n}(R)$, and $B \in M_{m\times p}(R)$ then*
$$r(AB) = (rA)B.$$

(6) *If $A, B \in M_{m\times n}(R)$ and $C \in M_{n\times p}(R)$ then*
$$(A+B)C = AC + BC.$$

(7) *If $A \in M_{m\times n}(R)$ and $B, C \in M_{n\times p}(R)$ then*
$$A(B+C) = AB + AC.$$

(8) *If $A \in M_{m\times n}(R)$, $B \in M_{n\times p}(R)$, and $C \in M_{p\times q}(R)$ then*
$$(AB)C = A(BC).$$

(9) *If* $A \in M_{m\times n}(R)$ *and* $B \in M_{n\times p}(R)$ *then the transposes* $A^t \in M_{n\times m}(R)$ *and* $B \in M_{p\times n}(R)$ *satisfy the standard "reverse of order" under multiplication:*

$$(AB)^t = B^t A^t.$$

Proof. Basically these are all somewhat boring chases through the definitions. We do a couple just to give the idea. For example if $A = [a_{ij}]$, $B = [b_{ij}]$ then denoting the entries of $r(A+B)$ as $(r(A+B))_{ij}$ and the entries of $rA + rB$ as $(rA + rB)_{ij}$.

$$(r(A+B))_{ij} = r(a_{ij} + b_{ij}) = ra_{ij} + rb_{ij} = (rA + rB)_{ij}.$$

Thus shows $r(A+B)$ and $rA+rB$ have the same entries and therefore $r(A+B) = rA + rB$. This shows 2 holds.

To see that (8) holds let $A = [a_{ij}] \in M_{m\times n}(R)$, $B = [b_{jk}] \in M_{m\times p}(R)$, and $C = [c_{kl}] \in M_{p,q}(R)$. Then we write out the entries of $(AB)C$ (changing the order of summation at one point) to get

$$\begin{aligned}
((AB)C)_{il} &= \sum_{k=1}^{p} (AB)_{ik} c_{kl} = \sum_{k=1}^{p} \sum_{j=1}^{n} a_{ij} b_{jk} c_{kl} \\
&= \sum_{j=1}^{n} a_{ij} \sum_{k=1}^{p} b_{jk} c_{kl} = \sum_{j=1}^{n} a_{ij} (BC)_{jl} \\
&= (A(BC))_{il}.
\end{aligned}$$

This shows $(AB)C$ and $A(BC)$ have the same entries and so (8) is proven. The other parts of the proposition are left to the reader. □

Problem 29. Prove the rest of the last proposition. □

In the future we will make use of the properties given in Proposition 3.2 without explicitly quoting the Proposition.

3.1.3. *The identity matrix, diagonal matrices, and the Kronecker delta.* The n by n ***identity matrix*** I_n in $M_{n\times n}(R)$ is the diagonal matrix with all diagonal elements equal to $1 \in R$ and all off diagonal elements equal to 0:

$$I_n = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}.$$

We will follow a standard convention and denote the entries of I_n by δ_{ij} and call this the ***Kronecker delta***. Explicitly

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j; \\ 0, & \text{if } i \neq j. \end{cases}$$

Then if $A \in M_{m\times n}(R)$ is as above then we compute the entries of $I_m A$.

$$\begin{aligned} (I_m A)_{ik} &= \sum_{j=1}^{m} \delta_{ij} a_{jk} && \text{(all but one term in the sum is zero)} \\ &= a_{ik} = A_{ik} && \text{(the surviving term).} \end{aligned}$$

Therefore $I_m A$ and A have the same entries, whence $I_m A = A$. A similar calculation shows $AI_n = A$. Whence

$$I_m A = AI_n \quad \text{for all} \quad A \in M_{m\times n}(R).$$

So the identity matrices are identities with respect to matrix multiplication.

More generally for $c_1, c_2, \ldots, c_n \in R$ then we can define the ***diagonal matrix*** $\operatorname{diag}(c_1, c_2, \ldots, c_n) \in M_{n\times n}$ with $c_1, \ldots, c_n$ down the main diagonal and zeros elsewhere. That is

$$\operatorname{diag}(c_1, c_2, \ldots, c_n) = \begin{bmatrix} c_1 & 0 & 0 & \cdots & 0 \\ 0 & c_2 & 0 & \cdots & 0 \\ 0 & 0 & c_3 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & c_n \end{bmatrix}.$$

In terms of δ_{ij} the components of $\operatorname{diag}(c_1, c_2, \ldots, c_n)$ are given by

$$\operatorname{diag}(c_1, c_2, \ldots, c_n) = \delta_{ij} c_i = \delta_{ij} c_j.$$

Thus if $D = \operatorname{diag}(c_1, c_2, \ldots, c_n)$ and $A = \begin{bmatrix} a_{ij} \end{bmatrix}$ is an $n \times p$ over R then

$$\begin{aligned} (DA)_{ik} &= \sum_{j=1}^{m} c_i \delta_{ij} a_{jk} && \text{(all but one term in the sum is zero)} \\ &= c_i a_{ik} = c_i A_{ik} && \text{(the surviving term).} \end{aligned}$$

and if $B = \begin{bmatrix} b_{ij} \end{bmatrix}$ is $m \times n$ then an almost identical calculation gives

$$(BD)_{ik} = b_{ik} c_k = B_{ik} c_k.$$

These facts can be stated in a particularly nice form:

Problem 30. Let $D = \operatorname{diag}(c_1, c_2, \ldots, c_n)$. Let A be a matrix with n rows (and any number of columns)

$$A = \begin{bmatrix} A^1 \\ A^2 \\ \vdots \\ A^n \end{bmatrix}.$$

Then show multiplying A on the left by D multiplies the rows of by $c_1, c_2, \ldots, c_n$. That is

$$DA = \begin{bmatrix} c_1 A^1 \\ c_2 A^2 \\ \vdots \\ c_n A^n \end{bmatrix}.$$

Likewise show that if B has n columns (and any number of rows)

$$B = \begin{bmatrix} B_1, B_2, \ldots, B_n \end{bmatrix}$$

then multiplying B on the right by D multiplies the columns by $c_1, \ldots, c_n$. That is

$$BD = \begin{bmatrix} c_1 B_1, c_2 B_2, \ldots, c_n B_n \end{bmatrix}. \qquad \square$$

3.1.4. *Block matrix multiplication.* It is often easier see what is in going in matrix multiplication if the matrices are partitioned into smaller matrices (called blocks). As first example of this we give

Proposition 3.3. *Let R be a commutative ring and $A \in M_{m\times n}(R)$, $B \in M_{n\times p}(R)$. Let*

$$A = \begin{bmatrix} A^1 \\ A^2 \\ \vdots \\ A^m \end{bmatrix}$$

where $A^1, A^2, \ldots, A^m$ are the rows of A and let

$$B = \begin{bmatrix} B_1, B_2, \ldots, B_p \end{bmatrix}$$

where $B_1, B_2, \ldots, B_p$ are the columns of B. Then

$$AB = \begin{bmatrix} A^1 B \\ A^2 B \\ \vdots \\ A^m B \end{bmatrix} = \begin{bmatrix} AB_1, AB_2, \ldots, AB_p \end{bmatrix}.$$

That is it is possible to multiple B (on the left) by A column at a time and if is possible to multiple A (on the right) by B row at a time.

Problem 31. Prove this. *Hint:* On way to approach this (which may be a bit on the formal side for some people) is as follows. Let the kth column of B be

$$B_k = \begin{bmatrix} b_{1k} \\ b_{2k} \\ \vdots \\ b_{pk} \end{bmatrix}$$

Then

$$AB_k = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix}$$

where

$$c_i = \sum_{j=1}^{m} a_{ij} b_{jk}.$$

But then $c_1, c_2, \ldots, c_n$ are the elements of the kth column of AB. A similar calculation works for the rows of AB. $\square$

Let $A \in M_{m\times n}(R)$ and $B \in M_{n\times p}(R)$. Let

$$m = m_1 + m_2, \qquad n = n_1 + n_2, \qquad p = p_1 + p_2.$$

where m_i, n_i, p_i are positive integers. Then write A and B as

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} \tag{3.1}$$

where A_{ij} and B_{ij} are matrices of the following sizes

$$\begin{array}{lll lll} A_{11} & \text{is} & m_1 \times n_1 & B_{11} & \text{is} & n_1 \times p_1 \\ A_{12} & \text{is} & m_1 \times n_2 & B_{12} & \text{is} & n_1 \times p_2 \\ A_{21} & \text{is} & m_2 \times n_1 & B_{21} & \text{is} & n_2 \times p_1 \\ A_{22} & \text{is} & m_2 \times n_2 & B_{22} & \text{is} & n_2 \times p_2. \end{array}$$

(Which could be more succinct expressed by saying that A_{ij} is $m_i \times n_j$ and B_{ij} is $n_i \times p_j$.) The matrices A_{ij} and B_{ij} are often called ***blocks*** or ***partitions*** of A and B. For example if

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{bmatrix}$$

and $m_1 = 3$, $m_2 = 1$, $n_1 = 2$, and $n_2 = 3$ then we split A in to blocks A_{ij} with

$$A_{11} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \end{bmatrix}, \qquad A_{21} = \begin{bmatrix} a_{14} & a_{15} \end{bmatrix},$$
$$A_{21} = \begin{bmatrix} a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix}, \qquad A_{22} = \begin{bmatrix} a_{24} & a_{25} \\ a_{34} & a_{35} \\ a_{44} & a_{45} \end{bmatrix}.$$

Proposition 3.4. *Let $A \in M_{m\times n}(R)$ and $B \in M_{n\times p}(R)$ and let A and B be partitioned as in (3.1). Then the product AB can be computed block at a time:*

$$AB = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix} = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}.$$

Problem 32. Prove this. (Depending on your temperament, it may or may not be worth writing out a formal proof. Just thinking hard about how you multiply matrices should be enough to convince you it is true.) □

This generalizes to larger numbers of blocks.

Problem 33. Let $A \in M_{m\times n}(R)$ and $B \in M_{n\times p}(R)$ and assume that A and B are partitioned as

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1s} \\ A_{21} & A_{22} & \cdots & A_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ A_{r1} & A_{m2} & \cdots & A_{rs} \end{bmatrix}, \qquad B = \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1t} \\ B_{21} & B_{22} & \cdots & B_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ B_{s1} & B_{s2} & \cdots & B_{st} \end{bmatrix}.$$

Then, provided the size of the blocks is such the products involved are defined,

$$\begin{aligned} AB &= \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1s} \\ A_{21} & A_{22} & \cdots & A_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ A_{r1} & A_{m2} & \cdots & A_{rs} \end{bmatrix} \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1t} \\ B_{21} & B_{22} & \cdots & B_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ B_{s1} & B_{s2} & \cdots & B_{st} \end{bmatrix} \\ &= \begin{bmatrix} C_{11} & C_{12} & \cdots & C_{1t} \\ C_{21} & C_{22} & \cdots & C_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ C_{r1} & C_{r2} & \cdots & C_{rt} \end{bmatrix} \end{aligned}$$

where each C_{ik} is the sum of matrix products

$$C_{ik} = \sum_{j=1}^{s} A_{ij}B_{jk}.$$

Prove this. (Again thinking hard about how you multiple matrices may be as productive as writing out a detailed proof.) □

3.2. **Inverses of matrices.** As in the case of matrices over a field inverses of matrices of square matrices with elements in a ring are important. The theory is just enough more complicated to be fun.

3.2.1. *The definition and basic properties of inverses.* The definition of being invertible is just as one would expect from the case of fields.

Definition 3.5. Let R be a commutative ring and let $A \in M_{n\times n}(R)$. Then B is the ***inverse*** of A iff

$$AB = BA = I_n.$$

(Note this is symmetric in A and B so that A is inverse of B.) When A has in inverse we say that A is ***invertible***.

If A has an inverse it is unique. For if B_1 and B_2 are inverses of A then

$$B_1 = B_1 I_n = B_1(AB_2) = (B_1A)B_2 = I_nB_2 = B_2.$$

Because of the uniqueness we can write the inverse of A as A^{-1}. Note that the symmetry of A and B in the definition of inverse implies that if A is invertible then so is $B = A^{-1}$ and $B^{-1} = A$. That is

$$(A^{-1})^{-1} = A.$$

Before giving examples of invertible matrices we record some elementary properties of invertible matrices and inverses.

Proposition 3.6. *Let R be a commutative ring.*

(1) *If $A, B \in M_{n\times n}(R)$ and both A and B are invertible then so is the product AB and it has inverse*

$$(AB)^{-1} = B^{-1}A^{-1}.$$

(2) *If $A \in M_{n\times n}(R)$ is invertible, then for $k = 0, 1, 2, \ldots$ then A^k is invertible and*

$$(A^k)^{-1} = (A^{-1})^k.$$

From now on we write A^{-k} for $(A^k)^{-1} = (A^{-1})^k$. (Note this includes the case of $A^0 = I_n$.)

(3) *Generalizing both these cases we have that if $A_1A_2, \ldots, A_k \in M_{n\times n}(R)$ are all invertible then so is the product $A_1A_2\cdots A_k$ and*

$$(A_1A_2\cdots A_k)^{-1} = A_k^{-1}A_{k-1}^{-1}\cdots A_1^{-1}.$$

Proof. If A, B are both invertible then set $C = B^{-1}A^{-1}$ and compute

$$(AB)C = ABB^{-1}A^{-1} = AI_nA^{-1} = AA^{-1} = I_n$$

and

$$C(AB) = B^{-1}A^{-1}AB = B^{-1}I_nB = B^{-1}B = I_n.$$

Thus C is the inverse of AB as required. The other two parts of the proposition follow by repeated use of the first part (or by induction if you like being a bit more formal). $\square$

3.2.2. *Inverses of* 2×2 *matrices.* We now give some examples of invertible matrices. First if $A := \begin{bmatrix} a_1 & 0 \\ 0 & a_2 \end{bmatrix} \in M_{2\times 2}(R)$ is a 2×2 diagonal matrix and both a_1 and a_2 are units (that is have inverses in R) then A^{-1} exists and is given by $A^{-1} = \begin{bmatrix} a_1^{-1} & 0 \\ 0 & a_2^{-1} \end{bmatrix}$. But if either of a_1 or a_2 are not units then A will not have an inverse in $M_{2\times 2}(R)$. As a concrete example let $R = \mathbf{Z}$ be the integers and let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

Then if A^{-1} existed it would have to be given by

$$A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

but the entries of this are not all integers so A has no inverse in $M_{2\times 2}(\mathbf{Z})$. More generally it is not hard to understand when a 2×2 matrix has an inverse. (The following is a special case of Theorem 4.21, p. 57 below.)

Theorem 3.7. *Let R be a commutative ring and let $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \in M_{2\times 2}(R)$. Then A has in inverse in $M_{n\times n}(R)$ if and only if* $\det(A) = (ad - bc)$ *is a unit. In this case the inverse is given by*

$$A^{-1} = (ad - bc)^{-1} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Proof. Set $B = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$ and compute

$$AB = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} = \begin{bmatrix} ad - bc & 0 \\ 0 & ad - bc \end{bmatrix} = (ad - bc)I_2 \tag{3.2}$$

and

$$BA = \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} ad - bc & 0 \\ 0 & ad - bd \end{bmatrix} = (ad - bc)I_2 \tag{3.3}$$

Therefore if $(ad - bc)$ is a unit, then $(ad - bc)^{-1} \in R$ and so $(ad - bc)^{-1}B \in M_{2\times 2}(R)$. Thus multiplying (3.2) and (3.3) by $(ad - bc)^{-1}$ gives that $((ad - bc)^{-1}B)A = A((ad - bc)^{-1}B) = I_2$ and thus $(ad - bc)^{-1}B$ is the inverse of A.

Conversely if A^{-1} exists then we use that the determinant of a product is the product of the determinants (a fact we will prove later. See 4.16) to conclude

$$1 = \det(A^{-1}A) = \det(A^{-1})\det(A)$$

but this implies that $\det(A)$ is a unit in R with inverse $\det(A^{-1})$. □

3.2.3. *Inverses of diagonal matrices.* Another easy case class of matrices to understand form the point of view of inverses is the diagonal matrices.

Theorem 3.8. *Let R be a commutative ring, then a diagonal matrix $D = \operatorname{diag}(a_1, a_2, \ldots, a_n) \in M_{n\times n}(R)$ is invertible if and only if all the diagonal elements $a_1, a_2, \ldots, a_n$ are units in R.*

Proof. One direction is clear. If all the elements $a_1, a_2, \ldots, a_n$ are units in R then the inverse of D exists and is given by

$$D^{-1} = \operatorname{diag}(a_1^{-1}, a_2^{-1}, \ldots, a_n^{-1}).$$

Conversely assume that D has an inverse. As D is diagonal its elements are of the form $D_{ij} = a_i\delta_{ij}$ where δ_{ij} the Kronecker delta. Let $B = [b_{ij}] \in M_{n\times n}(R)$ be the inverse of D. Then $BD = I_n$. As the entries of I_n are δ_{ij} the equation $I_n = BD$ is equivalent to

$$\delta_{ik} = \sum_{j=1}^{n} b_{ij}D_{jk} = \sum_{j=1}^{n} b_{ij}a_j\delta_{jk} = b_{ik}a_k.$$

Letting $k = i$ in this leads to $1 = \delta_{ii} = b_{ii}a_i$. Therefore a_i has an inverse in R: $a_i^{-1} = b_{ii}$. Thus all the diagonal elements $a_1, a_2, \ldots, a_n$ of D are units. □

3.2.4. *Nilpotent matrices and inverses of triangular matrices.*

Definition 3.9. A matrix $N \in M_{n\times n}$ is ***nilpotent*** iff there is an $m \geq 1$ so that $N^m = 0$. If m is the smallest positive integer for which $N^m = 0$ we call m the ***index of nilpotency*** of N.

Remark 3.10. The rest of the material on finding inverses of matrices is a (hopefully interesting) aside and is not essential to the rest of these notes and you can skip to directly to Section 4.1 on Page 36. (However the definition of nilpotent is important and you should make a point of knowing it.) □

Proposition 3.11. *If R is a commutative ring and $N \in M_{n\times n}(R)$ is nilpotent with nilpotency index m, then $I - N$ is invertible with inverse*

$$(I-N)^{-1} = I + N + N^2 + \cdots + N^{m-1}.$$

By replacing N by $-N$ we see that $I + N$ is also invertible and has inverse

$$(I+N)^{-1} = I - N + N^2 - N^3 + \cdots (-1)^{m-1}N^{m-1}.$$

Problem 34. Prove this. *Hint:* Set $B = I + N + N^2 + \cdots + N^{m-1}$ and compute directly that $(I-N)B = B(I-N) = I$ □

Remark 3.12. Recall from calculus that if $a \in \mathbf{R}$ has $|a| < 1$ then the inverse $1/(1-a)$ can be computed by the geometric series

$$\frac{1}{1-a} = 1 + a + a^2 + a^3 + \cdots = \sum_{k=0}^{\infty} a^k.$$

The formula above for $(I-N)^{-1}$ can be "derived" from this by just letting $a = N$ in the series for $1/(1-a)$ an using that $N^k = 0$ for $k \geq m$. □

We now give examples of nilpotent matrices. Recall that a matrix $A \in M_{n\times n}(R)$ is ***upper triangular*** iff all the elements of A below the main diagonal are zero. That is if A is of the form

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & a_{33} & \cdots & a_{3\,n-1} & a_{3\,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & 0 & \cdots & 0 & a_{nn} \end{bmatrix}.$$

More formally

$$A = [a_{ij}] \text{ is upper triangular} \quad \iff \quad a_{ij} = 0 \text{ for } i > j.$$

Also recall that a matrix B is ***strictly upper triangular*** iff all the elements of B on or below the main diagonal of B are zero. (Thus being strictly upper triangular differs from just being upper triangular by the extra requirement of having the diagonal elements vanish). So if B is strictly upper triangular it is of the form

$$B = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1\,n-1} & a_{1\,n} \\ 0 & 0 & a_{23} & \cdots & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & 0 & \cdots & a_{3\,n-1} & a_{3\,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{n-1\,n} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}.$$

Again we can be formal:

$$B = [b_{ij}] \text{ is strictly upper triangular} \quad \iff \quad b_{ij} = 0 \text{ for } i \geq j.$$

We define ***lower triangular*** and ***strictly lower triangular*** matrices in an analogous manner.

We show, as an application of block matrix multiplication, that a strictly upper triangular matrix is nilpotent.

Proposition 3.13. *Let R be a commutative ring and let $A \in M_{n\times n}(R)$ be either strictly upper triangular or strictly lower triangular. Then A is nilpotent. In fact $A^n = 0$.*

Proof. We will do the proof for strictly upper triangular matrices, the proof for strictly lower triangular matrices being just about identical. The proof is by induction on n. When $n = 2$ a strictly upper triangular matrix $A \in M_{2\times 2}(R)$ is of the form $A = \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix}$ for some $a \in R$. But then

$$A^2 = \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 0 & a \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

This is the base case for the induction. Now assume that the result holds for all $n \times n$ strictly upper triangular matrices and let A be a strictly upper triangular $(n+1) \times (n+1)$ matrix. We write A as a block matrix

$$A = \begin{bmatrix} B & v \\ 0 & 0 \end{bmatrix}$$

where B is $n \times n$, v is $n \times 1$, the first 0 in the bottom is $1 \times n$ and the second 0 is 1×1. As A is strictly upper triangular the same will be true for B. As B is $n \times n$ we have by the induction hypothesis that

$B^n = 0$. Now compute:

$$\begin{aligned}
A^2 &= \begin{bmatrix} B & v \\ 0 & 0 \end{bmatrix} \begin{bmatrix} B & v \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} B^2 & Bv \\ 0 & 0 \end{bmatrix}, \\
A^3 = AA^2 &= \begin{bmatrix} B & v \\ 0 & 0 \end{bmatrix} \begin{bmatrix} B^2 & Bv \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} B^3 & B^2 v \\ 0 & 0 \end{bmatrix} \\
A^4 = AA^3 &= \begin{bmatrix} B & v \\ 0 & 0 \end{bmatrix} \begin{bmatrix} B^3 & B^2 v \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} B^4 & B^3 v \\ 0 & 0 \end{bmatrix} \\
\vdots \qquad & \vdots \\
A^{n+1} &= \begin{bmatrix} B^{n+1} & B^n v \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.
\end{aligned}$$

This closes the induction and completes the proof. $\square$

We can now give another example of invertible matrices.

Theorem 3.14. *Let R be a commutative ring and let $A \in M_{n\times n}(R)$ be upper triangular and assume that all the diagonal elements a_{ii} of A are units. Then A is invertible. (Likewise a lower triangular matrix that has units along its diagonal is invertible.)*

Remark 3.15. The proof below is probably not the "best" proof, but it illustrates ideas that are useful elsewhere. The standard proof is to just back solve in the usual manner. In doing this one only needs to divide by the diagonal elements and so the calculations works just as it does over a field. A 3×3 example should make this clear. Let

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & 0 & a_{33} \end{bmatrix}$$

To find the inverse of A we form the matrix $[A\, I_3]$ and row reduce. This is

$$[A\, I_3] = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 1 & 0 & 0 \\ 0 & a_{22} & a_{23} & 0 & 1 & 0 \\ 0 & 0 & a_{33} & 0 & 0 & 1 \end{bmatrix}.$$

Row reducing this to echelon form only involves division by the elements a_{11}, a_{22}, and a_{33} and as we are assuming that this are units the elements a_{11}^{-1}, a_{22}^{-1}, and a_{33}^{-1} exist. If you do the calculation you should get

$$A^{-1} := \begin{bmatrix} \dfrac{1}{a_{1\,1}} & -\dfrac{a_{1\,2}}{a_{1\,1}\,a_{2\,2}} & \dfrac{a_{1\,2}\,a_{2\,3} - a_{1\,3}\,a_{2\,2}}{a_{1\,1}\,a_{2\,2}\,a_{3\,3}} \\ 0 & \dfrac{1}{a_{2\,2}} & -\dfrac{a_{2\,3}}{a_{2\,2}\,a_{3\,3}} \\ 0 & 0 & \dfrac{1}{a_{3\,3}} \end{bmatrix}$$

The same pattern holds in higher dimensions. □

Proof. Let A be upper triangular and let $D = \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn})$ be the diagonal part of A, that is the diagonal matrix that has the same entries down the diagonal as A. We now factor A into a product $D(I_n + N)$ where N is upper triangular and thus nilpotent. The idea is that $A = D(D^{-1}A)$ and a multiplication by on the left by the diagonal matrix D^{-1} multiplies the rows by $a_{11}^{-1}, a_{22}^{-1}, \ldots, a_{nn}^{-1}$ the matrix $D^{-1}A$ will have 1's down the main diagonal We can therefore write $D^{-1}A$ as the sum of the identity I_n and a strictly upper triangular matrix. Explicitly:

$$\begin{aligned} A &= \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1\,n-1} & a_{1\,n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2\,n-1} & a_{2\,n} \\ 0 & 0 & a_{33} & \cdots & a_{3\,n-1} & a_{3\,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n-1\,n-1} & a_{n-1\,n} \\ 0 & 0 & 0 & \cdots & 0 & a_{nn} \end{bmatrix} \\ &= D \begin{bmatrix} 1 & a_{12} & a_{13}/a_{11} & \cdots & a_{1\,n-1}/a_{11} & a_{1\,n}/a_{11} \\ 0 & 1 & a_{23}/a_{22} & \cdots & a_{2\,n-1}/a_{22} & a_{2\,n}/a_{22} \\ 0 & 0 & 1 & \cdots & a_{3\,n-1}/a_{33} & a_{3\,n}/a_{33} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & a_{n-1\,n}/a_{n-1\,n-1} \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \\ &= D \left(\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \right. \end{aligned}$$

$$+\begin{bmatrix} 0 & a_{12} & a_{13}/a_{11} & \cdots & a_{1\,n-1}/a_{11} & a_{1\,n}/a_{11} \\ 0 & 0 & a_{23}/a_{22} & \cdots & a_{2\,n-1}/a_{22} & a_{2\,n}/a_{22} \\ 0 & 0 & 0 & \cdots & a_{3\,n-1} & a_{3\,n}/a_{33} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{n-1\,n}/a_{n-1\,n-1} \\ 0 & 0 & 0 & \cdots & 0 & 0 \end{bmatrix}\Bigg)$$
$$= D(I_n + N)$$

where the matrix N is clearly strictly upper triangular. The diagonal matrix D is invertible by Theorem 3.8 and $I_n + N$ is invertible by Proposition 3.13 and Proposition 3.11. Thus the product is invertible. In fact we have (using Proposition 3.11)

$$A^{-1} = (I_n + N)^{-1} D^{-1} = (I - N + N^2 - N^3 + \cdots (-1)^{n-1} N^{n-1}) D^{-1}.$$

This completes the proof. □

4. Determinants

4.1. **Alternating n linear functions on $M_{n\times n}(R)$.** We will derive the basic properties of determinants of matrices by showing that they are the unique n-linear alternating functions defined on $M_{n\times n}(R)$ that take the value 1 on the identity matrice. As I am assuming that you have seen determinants is some form or another before, this presentation will be rather brief and many of the details will be left to you. We start by defining the terms just used.

Let R^n be the set of length n column vectors with elements in the ring R. Then an element $A \in M_{n\times n}(R)$ can be thought as $A = [A_1, A_2, \ldots, A_n]$ where $A_1, A_2, \ldots, A_n$ are the columns of A so that each $A_j \in R^n$.

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

Then $A = [A_1, A_2, \ldots, A_n]$ where

$$A_1 = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix}, \ A_2 = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix}, \ldots, \ A_j = \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix}, \ldots, \ A_n = \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{nn} \end{bmatrix}.$$

The following isolates one of the basic properties of determinants, that they are linear functions of each of their columns.

Definition 4.1. A function $f\colon M_{n\times n}(R)\to R$ is n ***linear over*** R iff it is a linear function of each of its columns if the other $n-1$ columns are kept fixed. For the first column this means that if $A_1', A_1'', A_2, \dots, A_n \in M_{n\times n}(R)$ and $c', c'' \in R$, then

$$\begin{aligned} f(c'A_1' + c''A_1'',&A_2, A_3, \dots, A_n) \\ &= c'f(A_1', A_2, A_3, \dots, A_n) + c''f(A_1'', A_2, A_3, \dots, A_n). \end{aligned}$$

For the second column this means that if $A_1, A_2', A_2'', A_3, \dots, A_n \in M_{n\times n}(R)$ and $c', c'' \in R$, then

$$\begin{aligned} f(A_1, c'A_2'+&c''A_2'', A_3, , \dots, A_n) \\ &= c'f(A_1, A_2', A_3, \dots, A_n) + c''f(A_1, A_2'', A_3, \dots, A_n). \end{aligned}$$

And so on for the rest of the columns. $\square$

One way to think of this definition is that a function $f\colon M_{n\times n}(R)\to R$ is one that can be expanded down any of its columns. Instead of trying to make this precise we just give a couple of examples. First consider the 2×2 case. That is

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \left[\begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}\right] = \left[a_{11}\begin{bmatrix} 1 \\ 0 \end{bmatrix} + a_{21}\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}\right].$$

So if $f\colon M_{2\times 2}(R)\to R$ is 2 linear over R then

$$\begin{aligned} f(A) &= f\left(\left[a_{11}\begin{bmatrix} 1 \\ 0 \end{bmatrix} + a_{21}\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}\right]\right) \\ &= a_{11}f\left(\left[\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}\right]\right) + a_{21}f\left(\left[\begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}\right]\right) \\ &= a_{11}f\left(\begin{bmatrix} 1 & a_{12} \\ 0 & a_{22} \end{bmatrix}\right) + a_{21}f\left(\begin{bmatrix} 0 & a_{12} \\ 1 & a_{22} \end{bmatrix}\right) \end{aligned}$$

Likewise

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \left[\begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix}, a_{12}\begin{bmatrix} 1 \\ 0 \end{bmatrix} + a_{22}\begin{bmatrix} 0 \\ 1 \end{bmatrix}\right]$$

implies that

$$f(A) = a_{12}f\left(\begin{bmatrix} a_{11} & 1 \\ a_{21} & 0 \end{bmatrix}\right) + a_{22}f\left(\begin{bmatrix} a_{11} & 0 \\ a_{21} & 1 \end{bmatrix}\right).$$

For $n=3$, let $A\in M_{3\times 3}(R)$ be given by

$$A := \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}.$$

Using that

$$\begin{bmatrix} a_{11} \\ a_{21} \\ a_{31} \end{bmatrix} = a_{11} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + a_{21} \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + a_{31} \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

we find that if $f\colon M_{3\times 3}(R) \to R$ is 3 linear over R then

$$\begin{aligned} f(A) &= f\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \right) \\ &= a_{11} f\left(\begin{bmatrix} 1 & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} \right) + a_{21} f\left(\begin{bmatrix} 0 & a_{12} & a_{13} \\ 1 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{bmatrix} \right) \\ &\quad + a_{31} f\left(\begin{bmatrix} 0 & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 1 & a_{32} & a_{33} \end{bmatrix} \right) \end{aligned}$$

with corresponding formulas for expanding down the second or third columns.

We now isolate another of the determinant's essential properties.

Definition 4.2. Let $f\colon M_{n\times n}(R) \to R$ be n linear over R. Then f is ***alternating*** iff whenever two columns of A are equal then $f(A) = 0$. That is if $A = [A_1, A_2, \dots, A_n]$ and $A_j = A_k$ for some $j \neq k$ then $f(A) = 0$. $\square$

This implies another familiar property of determinants.

Proposition 4.3. *Let $f\colon M_{n\times n}(R) \to R$ be n linear over R and alternating. Then for $A \in M_{n\times n}(R)$ interchanging two columns of A changes the sign of $f(A)$. Explicitly for the first two columns of A this means that*

$$f([A_2, A_1, A_3, A_4, \dots, A_n]) = -f([A_1, A_2, A_3, A_4, \dots, A_n]).$$

More generally we have

$$f([\dots, A_k, \dots, A_j, \dots]) = -f([\dots, A_i, \dots, A_k, \dots])$$

where $[\dots, A_k, \dots, A_j, \dots]$ and $[\dots, A_j, \dots, A_k, \dots]$ only differ by having the j-th and k-th columns interchanged.

Proof. We first look at the case of the first two columns. Let $A = [A_1, A_2, A_3, \dots, A_n]$. Consider the matrix $[A_1 + A_2, A_1 + A_2, A_3, \dots, A_n]$ which as its first two columns $A_1 + A_2$ and the rest of its columns the same as the corresponding columns of A. Then as two columns are equal we have $f([A_1 + A_2, A_1 + A_2, A_3, \dots, A_n]) = 0$. Likewise

$f([A_1, A_1, A_3, \dots, A_n]) = 0$ and $f([A_2, A_2, A_3, \dots, A_n])$. Using these facts and that f is n linear over R we find

$$\begin{aligned}
0 =& f([A_1 + A_2, A_1 + A_2, A_3, \dots, A_n]) \\
=& f([A_1, A_1 + A_2, A_3, \dots, A_n]) + f([A_2, A_1 + A_2, A_3, \dots, A_n]) \\
=& f([A_1, A_1, A_3, \dots, A_n]) + f([A_1, A_2, A_3, \dots, A_n]) \\
& \quad + f([A_2, A_1, A_3, \dots, A_n]) + f([A_2, A_2, A_3, \dots, A_n]) \\
=& 0 + f([A_1, A_2, A_3, \dots, A_n]) + f([A_2, A_1, A_3, \dots, A_n]) + 0 \\
=& f([A_1, A_2, A_3, \dots, A_n]) + f([A_2, A_1, A_3, \dots, A_n])
\end{aligned}$$

This implies

$$f([A_2, A_1, A_3, \dots, A_n]) = -f([A_1, A_2, A_3, \dots, A_n])$$

as required.

The case for general columns is the same, just messier notationally. For those of you who are gluttons for punishment here it is. Let $A = [\dots, A_j, \dots, A_k, \dots]$. Then all three of the matrices $[\dots, A_j + A_k, \dots, A_j + A_k, \dots]$, $[\dots, A_j, \dots, A_j, \dots]$, and $[\dots, A_k, \dots, A_k, \dots]$ have repeated columns and therefore

$$\begin{aligned}
f([\dots, A_j + A_k, \dots, A_j + A_k, \dots]) &= f([\dots, A_j, \dots, A_j, \dots]) \\
&= f([\dots, A_k, \dots, A_k, \dots]) = 0.
\end{aligned}$$

Again using this and that f is n linear over R we have

$$\begin{aligned}
0 =& f([\dots, A_j + A_k, \dots, A_j + A_k, \dots]) \\
=& f([\dots, A_j, \dots, A_j + A_k, \dots]) + f([\dots, A_k, \dots, A_j + A_k, \dots]) \\
=& f([\dots, A_j, \dots, A_j, \dots]) + f([\dots, A_j, \dots, A_k, \dots]) \\
& \quad + f([\dots, A_k, \dots, A_j, \dots]) + f([\dots, A_k, \dots, A_k, \dots]) \\
=& 0 + f([\dots, A_j, \dots, A_k, \dots]) + f([\dots, A_k, \dots, A_j, \dots]) + 0 \\
=& f([\dots, A_j, \dots, A_k, \dots]) + f([\dots, A_k, \dots, A_j, \dots])
\end{aligned}$$

which implies

$$f([\dots, A_k, \dots, A_j, \dots]) = -f([\dots, A_j, \dots, A_k, \dots])$$

and completes the proof. $\square$

4.1.1. *Uniqueness of alternating n linear functions on $M_{n\times n}(R)$ for $n = 2, 3$.* We now find all alternating $f\colon M_{n\times n}(R) \to R$ that are n linear over R for some small values of n. Toward this end let $e_1, e_2, \dots, e_n$ be the standard basis of R^n. That is

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \ e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \ e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \dots, \ e_{n-1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix}, \ e_n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}.$$

Let's look at the case of $n = 2$. Let $f\colon M_{2\times 2}(R) \to R$ be alternating and 2 linear over R. Let $A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \in M_{2\times 2}(R)$. Then we can write $A = [A_1, A_2]$ where the columns of A are

$$A_1 = \begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} = a_{11}e_1 + a_{21}e_2, \quad A_2 = \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix} = a_{12}e_1 + a_{22}e_2.$$

Therefore, using $f(e_1, e_1) = f(e_2, e_2) = 0$ and $f(e_2, e_1) = -f(e_1, e_2)$, we find

$$\begin{aligned} f(A) &= f([A_1, A_2]) = f(a_{11}e_1 + a_{21}e_2, a_{12}e_1 + a_{22}e_2) \\ &= a_{11}f(e_1, a_{12}e_1 + a_{22}e_2) + a_{21}f(e_2, a_{12}e_1 + a_{22}e_2) \\ &= a_{11}a_{12}f(e_1, e_1) + a_{11}a_{22}f(e_1, e_2) \\ &\qquad + a_{21}a_{12}f(e_2, e_1) + a_{21}a_{22}f(e_2, e_2) \\ &= a_{11}a_{22}f(e_1, e_2) + a_{21}a_{12}f(e_2, e_1) \\ &= a_{11}a_{22}f(e_1, e_2) - a_{21}a_{12}f(e_1, e_2) \\ &= (a_{11}a_{22} - a_{21}a_{12})f(e_1, e_2). \end{aligned}$$

Now note that $[e_1, e_2] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2$. Thus our calculation of $f(A)$ can be summarized as

Proposition 4.4. *Let $f\colon M_{2\times 2}(R) \to R$ be 2 linear and alternating. Then*

$$f(A) = (a_{11}a_{22} - a_{21}a_{12})f(I_2) = f(I_2)\det(A). \tag{4.1}$$

□

Let's try the same thing when $n = 3$. If

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

so that the columns of $A = [A_1, A_2, A_3]$ are

$$\begin{aligned} A_1 &= a_{11}e_1 + a_{21}e_2 + a_{31}e_3, \\ A_2 &= a_{12}e_1 + a_{22}e_2 + a_{32}e_3, \\ A_3 &= a_{13}e_1 + a_{23}e_2 + a_{33}e_3. \end{aligned}$$

Now we can expand $f(A)$ as we did in the $n = 2$ case. In doing this expansion we can drop all terms such as $f(e_1, e_1, e_3)$ or $f(e_2, e_1, e_2)$ that have a repeated factor as these will vanish as f is alternating. The result will be that there are only 6 terms that survive

$$\begin{aligned} f(A) &= f(a_{11}e_1 + a_{21}e_2 + a_{31}e_3, a_{12}e_1 + a_{22}e_2 + a_{32}e_3, a_{13}e_1 + a_{23}e_2 + a_{33}e_3) \\ &= a_{11}a_{22}a_{33}f(e_1, e_2, e_3) + a_{21}a_{32}a_{13}f(e_2, e_3, e_1) + a_{31}a_{12}a_{23}f(e_3, e_1, e_2) \\ &\quad + a_{21}a_{12}a_{33}f(e_2, e_1, e_3)a_{11}a_{32}a_{23}f(e_1, e_3, e_2) + a_{31}a_{22}a_{13}f(e_3, e_2, e_1) \end{aligned} \tag{4.2}$$

We now use the altenating property to simplify farther.

$$\begin{aligned} f(e_2, e_3, e_1) &= -f(e_1, e_3, e_2) = f(e_1, e_2, e_3) \\ f(e_3, e_1, e_2) &= -f(e_2, e_1, e_3) = f(e_1, e_2, e_3) \\ f(e_2, e_1, e_3) &= -f(e_1, e_2, e_3) \\ f(e_1, e_3, e_2) &= -f(e_1, e_2, e_3) \\ f(e_3, e_2, e_1) &= -f(e_1, e_2, e_3). \end{aligned}$$

Using these in the expansion (4.2) gives

$$\begin{aligned} f(A) &= \big(a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} \\ &\qquad\qquad - a_{21}a_{12}a_{33} - a_{11}a_{32}a_{23} - a_{31}a_{22}a_{13}\big)f(e_1, e_2, e_3) \\ &= \det(A)f(e_1, e_2, e_3) \end{aligned}$$

But again

$$[e_1, e_2, e_3] = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = I_3.$$

And so this calculation can also be summarized as

Proposition 4.5. *Let $f\colon M_{3\times 3}(R) \to R$ be 3 linear and alternating. Then*

$$f(A) = f(I_3)\det(A). \tag{4.3}$$

□

4.1.2. *Application of the uniqueness result.* We now show that for $A, B \in M_{3\times 3}(R)$ that $\det(BA) = \det(B)\det(A)$. Toward this end fix $B \in M_{3\times 3}(R)$ and define $f_B\colon M_{3\times 3}(R) \to R$ by

$$f_B(A) = \det(BA).$$

Writing A in terms of its columns $A = [A_1, A_2, A_3]$ the product BA then has columns $BA = [BA_1, BA_2, BA_3]$. Thus $f_B(A)$ can be written as

$$f_B(A) = f_B(A_1, A_2, A_3) = \det(BA_1, BA_2, BA_3).$$

We know that det is a linear function of each of its columns. Thus for $c', c'' \in \mathbf{F}$ and $A_1', A_1'' \in \mathbf{R}^3$ we have

$$\begin{aligned} f_B(c'A_1' + c''A_1'', A_2, A_3) &= \det(B(c'A_1' + c''A_1''), BA_2, BA_3) \\ &= \det(c'BA_1' + c''BA_1'', BA_2, BA_3) \\ &= c' \det(BA_1'+, BA_2, BA_3) \\ &\qquad + c'' \det(BA_1'', BA_2, BA_3) \\ &= c' f_B(A_1', A_2, A_3) + c'' f_B(A_1'', A_2, A_3). \end{aligned}$$

Thus f_B is a linear function its first column. Similar calculations show that it is linear as a function of the second and third columns. Thus f_B is 3 linear. If two columns of A are equal, say $A_2 = A_3$, then $BA_2 = BA_3$ and so

$$f_B(A) = \det(BA_1, BA_2, BA_2) = 0$$

as $\det = 0$ on matrices with two equal columns. Thus f_B is alternating. Thus we can use equation (4.3) to conclude that

$$\begin{aligned} \det(BA) &= f_B(A) = f_B(I_3)\det(A) \\ &= \det(BI_3)\det(A) \\ &= \det(B)\det(A) \end{aligned}$$

as required. Once we have the n dimensional version of Proposition 4.5 we will be able to use this argument to show that $\det(AB) = \det(A)\det(B)$ for $A, B \in M_{n\times n}(R)$ for any $n \geq 1$ and any commutative ring R.

4.2. **Existence of determinants.** Before going on we need to prove that there always exists a nonzero alternating n linear function $f\colon M_{n\times n}(R) \to R$. For $n = 2$ this is easy. We define the usual determinant for 2×2 matrices.

$$\det{}_2\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right) = a_{11}a_{22} - a_{21}a_{12}.$$

Then it is not hard to check that f is alternating, 2 linear, and that $\det_2(I_2) = 1$.

Problem 35. Verify these properties of $\det_2$. $\square$

Before giving our general existence result we need some notation. If $A \in M_{n\times n}(R)$ then let $A[ij] \in M_{(n-1)\times(n-1)}(R)$ be the $(n-1)\times(n-1)$ matrix obtained by crossing on the i-th row and the j-th column. This $(n-1)\times(n-1)$ is called the ij ***minor*** of A. If

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{4.4}$$

then, using the notation $\cancel{a}_{kl}$ for indicating that we are deleting the element a_{kl}, we have:

$$A[11] = \begin{bmatrix} \cancel{a}_{11} & \cancel{a}_{12} & \cancel{a}_{13} \\ \cancel{a}_{21} & a_{22} & a_{23} \\ \cancel{a}_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix},$$

$$A[32] = \begin{bmatrix} a_{11} & \cancel{a}_{12} & a_{13} \\ a_{21} & \cancel{a}_{22} & a_{23} \\ \cancel{a}_{31} & \cancel{a}_{32} & \cancel{a}_{33} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{bmatrix}$$

and if

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

then

$$A[23] = \begin{bmatrix} a_{11} & a_{12} & \cancel{a}_{13} & a_{14} \\ \cancel{a}_{21} & \cancel{a}_{22} & \cancel{a}_{23} & \cancel{a}_{24} \\ a_{31} & a_{32} & \cancel{a}_{33} & a_{34} \\ a_{41} & a_{42} & \cancel{a}_{43} & a_{44} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{14} \\ a_{31} & a_{32} & a_{34} \\ a_{41} & a_{42} & a_{44} \end{bmatrix}.$$

If $f\colon M_{n\times n}(R) \to R$ is n linear and alternating then for $1 \le i \le n+1$ define a function $D_i f\colon M_{(n+1)\times(n+1)}(R) \to R$ by

$$D_i f(A) = \sum_{j=1}^{n+1} (-1)^{i+j} a_{ij} f(A[ij]). \tag{4.5}$$

This is not as off the wall as you might first think. If $D_i f$ is the usual determinant then this is nothing more than expanding $D_i f(A)$ along the i-th row. For example when $n = 2$, so that $D_i f$ is defined on 3×3 matrices,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

by

$$\begin{aligned}
D_1 f(A) &= a_{11} f(A[11]) - a_{12} f(A[12]) + a_{13} f(A[13]) \\
&= a_{11} f\left(\begin{bmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{bmatrix}\right) - a_{12} f\left(\begin{bmatrix} a_{21} & a_{23} \\ a_{31} & a_{33} \end{bmatrix}\right) \\
&\qquad + a_{13} f\left(\begin{bmatrix} a_{21} & a_{22} \\ a_{31} & a_{32} \end{bmatrix}\right) \\
D_2 f(A) &= -a_{21} f(A[21]) + a_{22} f(A[22]) - a_{23} f(A[23]) \\
&= -a_{21} f\left(\begin{bmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{bmatrix}\right) + a_{22} f\left(\begin{bmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{bmatrix}\right) \\
&\qquad - a_{23} f\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{bmatrix}\right) \\
D_3 f(A) &= a_{31} f(A[31]) - a_{32} f(A[32]) + a_{33} f(A[33]) \\
&= a_{31} f\left(\begin{bmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{bmatrix}\right) - a_{32} f\left(\begin{bmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{bmatrix}\right) \\
&\qquad + a_{33} f\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right)
\end{aligned}$$

which are the usual rules for expanding determinants along the first second and third rows.

Proposition 4.6. *Let $f\colon M_{n\times n}(R) \to R$ be n linear over R and alternating. Then each of the functions $D_i f\colon M_{(n+1)\times(n+1)}(R) \to R$ defined by (4.5) above is $(n+1)$ linear over R and alternating. Also*

$$D_i f(I_{n+1}) = f(I_n).$$

Proof. The function $D_i f(A)$ is a sum of terms

$$(-1)^{i+j} a_{ij} f(A[ij]).$$

Consider this term as a function of the k-th column. If $j \neq k$ then a_{ij} does not depend on the k-th column and $f(A[ij])$ depends linearly on the k-th column we see that the term depends linearly on the k-th column of A. If $j = k$ then $f(A[ik])$ dose not depend on the k-th column of A, but a_{ik} does depend linearly on the k-th column. Thus our term depends linearly on the k-th column in this case also. But as the sum of linear functions is linear we see that $D_i f$ depends linearly on the k-th column. Thus $D_i f$ is $(n+1)$ linear over R.

Problem 36. Write out the details of this argument when $n = 2$ and $n = 3$. $\square$

If the column A_k and A_l of A are equal with $k \neq l$ then for $j \notin \{k, l\}$ the sub-matrix $A[ij]$ will have two equal columns and as f is alternating this implies $f(A[ij]) = 0$. Therefore in the definition (4.5) all but two terms vanish so that

$$\begin{aligned} D_i f((A) &= (-1)^{i+k} a_{ik} f(A[ik]) + (-1)^{i+l} a_{il} f(A[il])) \\ &= a_{ik}(-1)^i \big((-1)^k f(A[ik]) + (-1)^l f(A[il])\big). \end{aligned} \tag{4.6}$$

(We used that $a_{ik} = a_{il}$ as $A_k = A_l$.) The matrices $A[ik]$ and $A[il]$ have the same columns, but not in the same order. We can assume that $k < l$. It takes $l - k - 1$ interchanges of columns to make $A[il]$ the same as $A[ik]$. Therefore as f is alternating this implies that $f(A[ik]) = (-1)^{l-k-1} f(A[il])$. Using this in (4.6) gives

$$\begin{aligned} D_i f(A) &= a_{ik}(-1)^i \big((-1)^k (-1)^{l-k-1} f(A[il]) + (-1)^l f(A[il])\big) \\ &= a_{ik}(-1)^i \big((-1)^{l-1} f(A[il]) + (-1)^l f(A[il])\big) \\ &= a_{ik}(-1)^{i+l} \big(-f(A[il]) + f(A[il])\big) \\ &= 0. \end{aligned}$$

Thus $D_i f$ is alternating.

Problem 37. Verify the claims about $A[ik]$ and $A[il]$ having the same columns and the number of interchanges needed to put the columns of $A[il]$ in the same order as those of $A[ik]$. $\square$

To finish we compute $D_i f(I_{n+1})$. The only element in the i-th row of I_{n+1} that is not zero if the 1 which occurs in the $i\,i$-th place. Also $I_{n+1}[ii] = I_n$. Therefore in the definition (4.5) of $D_i f$ we have that

$$D_i f(I_{n+1}) = (-1)^{i+i} 1 f(I_{n+1}[ii]) = f(I_n).$$

This completes the proof. $\square$

Definition 4.7. For each $n \geq 1$ define a function $\det_n \colon M_{n\times n}(R) \to R$ by recursion. $\det_1([a_{11}]) = a_{11}$ and once $\det_n$ is defined let $\det_{n+1} = D_1 \det_n$. This is our official definition of the ***determinant***. $\square$

You can use this to check that for small values of n this gives the familiar formulas:

$$\det_2\left(\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}\right) = a_{11}a_{22} - a_{21}a_{12}$$

and

$$\det_3\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}\right) = a_{11}a_{22}a_{33} + a_{21}a_{32}a_{13} + a_{31}a_{12}a_{23} \\ - a_{21}a_{12}a_{33} - a_{11}a_{32}a_{23} - a_{31}a_{22}a_{13}.$$

Already $n = 4$ is not so small and we[5] get

$$\begin{aligned} &\det_4\left(\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}\right) \\ &= + a_{11}a_{22}a_{33}a_{44} - a_{11}a_{22}a_{34}a_{43} - a_{11}a_{32}a_{23}a_{44} \\ &\quad + a_{11}a_{32}a_{24}a_{43} + a_{11}a_{42}a_{23}a_{34} - a_{11}a_{42}a_{24}a_{33} \\ &\quad - a_{21}a_{12}a_{33}a_{44} + a_{21}a_{12}a_{34}a_{43} + a_{21}a_{32}a_{13}a_{44} \\ &\quad - a_{21}a_{32}a_{14}a_{43} - a_{21}a_{42}a_{13}a_{34} + a_{21}a_{42}a_{14}a_{33} \\ &\quad + a_{31}a_{12}a_{23}a_{44} - a_{31}a_{12}a_{24}a_{43} - a_{31}a_{22}a_{13}a_{44} \\ &\quad + a_{31}a_{22}a_{14}a_{43} + a_{31}a_{42}a_{13}a_{24} - a_{31}a_{42}a_{14}a_{23} \\ &\quad - a_{41}a_{12}a_{23}a_{34} + a_{41}a_{12}a_{24}a_{33} + a_{41}a_{22}a_{13}a_{34} \\ &\quad - a_{41}a_{22}a_{14}a_{33} - a_{41}a_{32}a_{13}a_{24} + a_{41}a_{32}a_{14}a_{23}. \end{aligned} \tag{4.7}$$

This is clearly too much of a mess to be of any direct use. If $\det_5(A)$ is expanded the result has 120 terms and $\det_n(A)$ has $n!$ terms.

We record that $\det_n$ does have the basic properties we expect.

Theorem 4.8. *The function* $\det_n\colon M_{n\times n}(R) \to R$ *is alternating and* n *linear over* R. *Its value on the identity matrix is*

$$\det_n(I_n) = 1.$$

Proof. The proof is by induction on n. For small values of n, say $n = 1$ and $n = 2$, this is easy to check directly. Thus the base of the induction holds. Now assume that $\det_n$ is alternating, n linear over R and satisfies $\det_n(I_n) = 1$. Then by Proposition 4.6 the function $\det_{n+1} = D_1 \det_n$ is alternating, $(n+1)$ linear over R and satisfies $\det_{n+1}(I_{n+1}) = \det_n(I_n) = 1$. This closes the induction and completes the proof. $\square$

[5] In this case "we" was the computer package Maple which will not only do the calculation but will output it as LaTeX code that can be cut and pasted into a document.

4.2.1. *Cramer's rule.* Consider a system of n equations in n unknowns $x_1, \ldots, x_n$,

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\
\vdots \qquad\qquad\qquad\quad \vdots\quad &\quad \vdots \\
a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n
\end{aligned} \tag{4.8}
$$

where $a_{ij}, b_i \in R$. We can use the existence of the determinant to give a rule for solving this system. By setting

$$
A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}
$$

The system (4.8) can be written as

$$
Ax = b.
$$

Or letting $A_1, \ldots, A_n$ be the columns of A, so that $A = [A_1, A_2, \ldots, A_n]$, this can be rewritten as

$$
x_1A_1 + x_2A_2 + \ldots x_nA_n = b. \tag{4.9}
$$

We look at the case of $n = 3$. Then this is

$$
x_1A_1 + x_2A_2 + x_3A_3 = b.
$$

Now if this holds we expand $\det_3(b, A_2, A_3)$ as follows:

$$
\begin{aligned}
\det{}_3(b, A_2, A_3) &= \det{}_3(x_1A_1 + x_2A_2 + x_3A_3, A_2, A_3) \\
&= x_1 \det{}_3(A_1, A_2, A_3) + x_2 \det{}_3(A_2, A_2, A_3) \\
&\qquad + x_3 \det{}_3(A_3, A_2, A_3) \\
&= x_1 \det(A)
\end{aligned}
$$

where we have used that $\det_3(A_1, A_2, A_3) = \det_3(A)$ and that $\det_3(A_2, A_2, A_3) = \det_3(A_3, A_2, A_3) = 0$ as a the determinant of a matrix with a repeated column vanishes. We can likewise expand

$$
\begin{aligned}
\det{}_3(A_1, b, A_3) &= \det(A_1, x_1A_1 + x_2A_2 + x_3A_3, A_3) \\
&= x_1 \det{}_3(A_1, A_1, A_3) + x_2 \det(A_1, A_2, A_3) \\
&\qquad + x_3 \det{}_3(A_1, A_3, A_3) \\
&= x_2 \det(A)
\end{aligned}
$$

and

$$\begin{aligned}\det_3(A_1, A_2, b) = &\det_3(A_1, A_2, x_1A_1 + x_2A_2 + x_3A_3)\\ =&x_1 \det_3(A_1, A_2, A_1) + x_2 \det_3(A_1, A_2, A_2)\\ &\quad + x_3 \det_3(A_1, A_2, A_3)\\ =&x_3 \det_3(A)\end{aligned}$$

Summarizing

$$\begin{aligned}\det_3(A)x_1 &= \det_3(b, A_2, A_3)\\ \det_3(A)x_2 &= \det_3(A_1, b, A_3)\\ \det_3(A)x_3 &= \det_3(A_1, A_2, b).\end{aligned}$$

In the case that R is a field and $\det_3(A) \neq 0$ then we can divide by $\det_3(A)$ and solve get formulas for x_1, x_2, x_3. This is the three dimensional version of Cramer's rule. The general case is

Theorem 4.9. *Let R be a commutative ring and assume that $x_1, \ldots, x_n$ is a solution to the system (4.8). Then*

$$\begin{aligned}\det_n(A)x_1 &= \det_n(b, A_2, A_3, \ldots, A_{n-1}, A_n)\\ \det_n(A)x_2 &= \det_n(A_1, b, A_3, \ldots, A_{n-1}, A_n)\\ \det_n(A)x_3 &= \det_n(A_1, A_2, b, \ldots, A_{n-1}, A_n)\\ &\vdots \qquad \vdots\\ \det_n(A)x_{n-1} &= \det_n(A_1, A_2, A_3, \ldots, b, A_n)\\ \det_n(A)x_n &= \det_n(A_1, A_2, A_3, \ldots, A_{n-1}, b).\end{aligned}$$

When R is a field and $\det_n(A) \neq 0$ then this gives formulas for $x_1, \ldots, x_n$.

Problem 38. Prove this along the lines of the three dimensional version given above. □

Problem 39. In the system (4.8) assume that $a_{ij}, b_i \in \mathbf{Z}$, the ring of integers. Then show that if $\det_n(A) \neq 0$ then (4.8) has a solution if and only if the numbers

$$\det_n(b, A_2, , \ldots, A_n),\ \det_n(A_1, b, \ldots, A_n)\ , \ldots,\ \det_n(A_1, A_2, \ldots, b)$$

are all divisible by $\det_n(A)$. □

4.3. Uniqueness of alternating n linear functions on $M_{n\times n}(R)$.

4.3.1. *The sign of a permutation.* Our next goal is to generalize the formulas (4.1) and (4.3) from $n = 2, 3$ to higher values of n. This unfortunately requires a bit more notation. Let S_n be the group of all permutations of the set $\{1, 2, \dots, n\}$. That is S_n is the set of all bijective functions $\sigma\colon \{1, 2, \dots, n\} \to \{1, 2, \dots, n\}$ with the group operation of function composition. If $e_1, e_2, \dots, e_n$ is the standard basis of R^n then the matrix $[e_1, e_2, \dots, e_n]$ is the identity matrix:

$$[e_1, e_2, \dots, e_n] = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} = I_n.$$

For $\sigma \in S_n$ we set $E(\sigma)$ to be the matrix

$$E(\sigma) = [e_{\sigma(1)}, e_{\sigma(2)}, e_{\sigma(3)}, \dots, \varepsilon_{\sigma(n)}].$$

Then $E(\sigma)$ is just $I_n = [e_1, e_2, \dots, e_n]$ with the columns in a different order.

Definition 4.10. For a permutation $\sigma \in S_n$ define

$$\operatorname{sgn}(\sigma) := \det_n(E(\sigma)).$$

□

As the matrix $E(\sigma)$ is just I_n with the columns in a different order we can reduce to I_n by repeated interchange of columns. This can be done as follows:

(1) If the first column of $E(\sigma)$ is equal to e_1 then do nothing and set $E'(\sigma) = E(\sigma)$. If the first column of $E(\sigma)$ is not e_1 then find the column of $E(\sigma)$ where e_1 appears and interchange this with the first column and let $E'(\sigma)$ be the result of this interchange. Then in either case we have that $E'(\sigma)$ has e_1 as its first column.
(2) If the second column of $E'(\sigma)$ is e_2 then do nothing and set $E''(\sigma) = E'(\sigma)$. If the second column of $E'(\sigma)$ is not equal to e_2 then find the column of $E'(\sigma)$ where e_2 appears and interchange this column with the second column of $E'(\sigma)$ and let $E''(\sigma)$ be the result of this interchange. Then in either case $E''(\sigma)$ has as its first two columns e_1 and e_2.
(3) If the third column of $E''(\sigma)$ is e_3 then do nothing and set $E'''(\sigma) = E''(\sigma)$. If the third column of $E''(\sigma)$ is not equal to e_3 then find the column of $E''(\sigma)$ where e_3 appears and interchange this column with the third column of $E''(\sigma)$ and let $E'''(\sigma)$ be

the result of this interchange. Then in either case $E''(\sigma)$ has as its first three columns e_1, e_2, and e_3.

(4) Continue in the manner and get a finite sequence
$$E(\sigma), E'(\sigma), \ldots, E^{(k)}(\sigma), \ldots, E^{(n)}(\sigma)$$
so that the first k columns of $E^{(k)}$ are $e_1, e_2, \ldots, e_k$ and at each step either $E^{(k)}(\sigma) = E^{(k-1)}(\sigma)$ or $E^{(k)}(\sigma)$ differs from $E^{(k-1)}(\sigma)$ by the interchange of two columns. The end result of this is that $E^{(n)} = [e_1, e_2, \ldots, e_n] = I_n$ and so I_n can be obtained from $E(\sigma)$ by $\leq n$ interchanges of columns.

As each interchange of a pair of columns of $E(\sigma)$ changes the sign of $\det_n(E(\sigma))$ (cf. Proposition 4.3) we have
$$\operatorname{sgn}(\sigma) = \begin{cases} +1, & \text{If } E(\sigma) \text{ can be reduced to } I_n \text{ with an even number of interchanges of columns,} \\ -1, & \text{If } E(\sigma) \text{ can be reduced to } I_n \text{ with an odd number of interchanges of columns.} \end{cases}$$
As the $\det_n(E(\sigma))$ has a definition that does not depend on interchanging columns this means given $\sigma \in S_n$ the number of interchanges to reduce $E(\sigma)$ to I_n is either always even or always odd. Given the many different ways and we could reduce $E(\sigma)$ to I_n by interchanging columns this is a rather remarkable fact. This observation has the following immediate application.

Lemma 4.11. *Let $f\colon M_{n\times n}(R) \to R$ be alternating and n linear over R. Then for any permutation $\sigma \in S_n$*
$$f([e_{\sigma(1)}, e_{\sigma(2)}, \ldots, e_{\sigma(n)}]) = \operatorname{sgn}(\sigma) f(I_n).$$

Proof. Recalling that $E(\sigma) = [e_{\sigma(1)}, e_{\sigma(2)}, \ldots, e_{\sigma(n)}]$ and that the interchange of two columns in $f([A_1, \ldots, A_n])$ changes the sign of $f([A_1, \ldots, A_n])$ we see that $f(E(\sigma)) = f([e_1, e_2, \ldots, e_n]) = f(I_n)$ if $E(\sigma)$ can be reduced to I_n by an even number of interchanges of columns and $f(E(\sigma)) = -f([e_1, e_2, \ldots, e_n]) = -f(I_n)$ if $E(\sigma)$ can be reduce to I_n by an odd number of interchanges of columns. That is $f(E(\sigma)) = \operatorname{sgn}(\sigma) f(I_n)$ as required. □

4.3.2. *Expansion as a sum over the symmetric group.* We now do the general case of the calculations that lead to (4.1) and (4.3). If $A = [a_{ij}] = [A_1, A_2, \ldots, A_n] \in M_{n\times n}(R)$ then we write the columns of A in terms of the standard basis:
$$A_1 = \sum_{i_1=1}^{n} a_{i_1 1} e_{i_1}, \quad A_2 = \sum_{i_2=1}^{n} a_{i_2 2} e_{i_2}, \ldots \quad A_n = \sum_{i_n=1}^{n} a_{i_n n} e_{i_n}.$$

Assume that $f\colon M_{n\times n}(R)\to R$ is n linear over R. Then we can expand $f(A)=f(A_1,A_2,\ldots,A_n)$ as

$$
\begin{aligned}
f(A) &= f\Big(\sum_{i_1=1}^n a_{i_11}e_{i_1}, \sum_{i_2=1}^n a_{i_22}e_{i_2}, \sum_{i_3=1}^n a_{i_33}e_{i_3}\ldots, \sum_{i_n=1}^n a_{i_nn}e_{i_n}\Big)\\
&= \sum_{i_1,i_2,i_3,\ldots,i_n=1}^n a_{i_11}a_{i_22}a_{i_33}\cdots a_{i_nn}f(e_{i_1},e_{i_2},e_{i_3},\ldots,e_{i_n})
\end{aligned}
$$

Now assume that besides being n linear over R that f is also alternating. Then, in any of the terms $f(e_{i_1},e_{i_2},e_{i_3},\ldots,e_{i_n})$, if $i_k=i_l$ for some $k\neq l$ then two columns of $[e_{i_1},e_{i_2},e_{i_3},\ldots,e_{i_n}]$ are the same and so $f(e_{i_1},e_{i_2},e_{i_3},\ldots,e_{i_n})=0$. Therefore the sum for $f(A)$ can be reduce to a sum over the terms where all of $i_1,i_2,i_3,\ldots,i_n$ are all distinct. That is the ordered n-tuple $(i_1,i_2,i_3,\ldots,i_n)$ is a permutation of $(1,2,3,\ldots,n)$. So we only have to sum over the tuples of the form $i_1=\sigma(1),i_2=\sigma(2),i_3=\sigma(3),\ldots,i_n=\sigma(n)$ for some permutation $\sigma\in S_n$. Thus for f alternating and n linear over R we get

$$
f(A)=\sum_{\sigma\in S_n} a_{\sigma(1)1}a_{\sigma(2)2}a_{\sigma(3)3}\cdots a_{\sigma(n)n}f\big(e_{\sigma(1)},e_{\sigma(2)},e_{\sigma(3)},\ldots,e_{\sigma(n)}\big)
$$

Now using Lemma 4.11 this simplifies farther to

$$
\begin{aligned}
f(A) &= \sum_{\sigma\in S_n} a_{\sigma(1)1}a_{\sigma(2)2}a_{\sigma(3)3}\cdots a_{\sigma(n)n}\operatorname{sgn}(\sigma)f(e_1,e_2,e_3,\ldots,e_n)\\
&= \Big(\sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)a_{\sigma(1)1}a_{\sigma(2)2}a_{\sigma(3)3}\cdots a_{\sigma(n)n}\Big)f(I_n) \qquad (4.10)
\end{aligned}
$$

This gives us another formula for $\det_n$.

Proposition 4.12. *The determinant of $A=[a_{ij}]\in M_{n\times n}(R)$ is given by*

$$
\begin{aligned}
\det_n(A) &= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)a_{\sigma(1)1}a_{\sigma(2)2}a_{\sigma(3)3}\cdots a_{\sigma(n)n}\\
&= \sum_{\sigma\in S_n}\operatorname{sgn}(\sigma)\prod_{i=1}^n a_{\sigma(i)i}
\end{aligned}
$$

Proof. We know (Theorem 4.8) that $\det_n$ is alternating, n linear over R and that $\det_n(I_n)=1$. Using this in (4.10) leads to the desired formulas for $\det_n(A)$. □

Remark 4.13. It is common to use the formula of the last proposition as the definition of the determinant. The problem with that from the point of view of the presentation here is that we defined $\operatorname{sgn}(\sigma)$ in terms of the determinant. However it is possible to give a definition of $\operatorname{sgn}(\sigma)$ that is independent of determinants and show that $\operatorname{sgn}(\sigma\tau) = \operatorname{sgn}(\sigma)\operatorname{sgn}(\tau)$ for all $\sigma, \tau \in S_n$. It is then not hard to show directly that $\det_n$ with this definition is n linear over R and alternating. While this sounds like less work, it is really about the same, as proving the facts about $\operatorname{sgn}(\sigma)$ requires an effort comparable to what we have done here. □

4.3.3. *The main uniqueness result.* We can now give a complete description of the alternating n linear functions $f\colon M_{n\times n}(R) \to R$.

Theorem 4.14. *Let R be a commutative ring and let $f\colon M_{n\times n}(R) \to R$ be an alternating function that is n linear over R. Then f is given in terms of the determinant as*

$$f(A) = \det_n(A)f(I_n).$$

Informally: Up to multiplication by elements of R, $\det_n$ is the unique n linear alternating function on $M_{n\times n}(R)$.

Proof. If $f\colon M_{n\times n}(R) \to R$ is an alternating function that is n linear over R, then combining the formula (4.10) with Proposition 4.12 yields the theorem. □

Remark 4.15. While this has taken a bit of work to get, the basic idea is quite easy and transparent. Review the calculations we did that lead up to (4.1) on Page 40 and (4.3) on Page 41 (which are the $n = 2$ and $n = 3$ versions of the result). The proof of Theorem 4.14 is just the same idea pushed through for larger values of n. That some real work should be involved in the general case can be seen by trying to do the "bare hands" proof in the cases of $n = 4$ or $n = 5$ (cf. (4.7)). □

4.4. **Applications of the uniqueness theorem and its proof.** It is a general meta-theorem in mathematics that uniqueness theorems allow one to prove properties of objects in ways that are often easier than a direct calculational proof. We now use Theorem 4.14 to give some non-computational proofs about the determinant.

4.4.1. *The product formula for determinants.* The first application is the important fact that the determinant is multiplicative (we have already done seen this for $n = 3$ in §4.1.2, page 42).

Theorem 4.16. *If $A, B \in M_{n\times n}(R)$ then* $\det_n(AB) = \det_n(A)\det_n(B)$.

Proof. We hold A fixed and define a function $f_A\colon M_{n\times n}(R) \to R$ by

$$f_A(B) = \det_n(AB).$$

If the columns of B are $B_1, B_2, \ldots, B_n$, so that $B = [B_1, B_2, \ldots, B_n]$, then block matrix multiplication implies that $AB = [AB_1, AB_2, \ldots, AB_n]$. Therefore we can rewrite f_A as

$$f_A(B) = \det_n(AB_1, AB_2, \ldots, AB_n).$$

As a function of B this is n linear over R. For example to see linearity in the first column let $c', c'' \in R$ and $B_1' B_1'' \in R^n$.

$$\begin{aligned}
f_A(c'B_1' + c''B_1'', &B_2, B_3, \ldots, B_n) \\
&= \det_n(A(c'B_1' + c''B_1''), AB_2, AB_3, \ldots, AB_n) \\
&= \det_n(c'AB_1' + c''AB_1'', AB_2, AB_3, \ldots, AB_n) \\
&= c' \det_n(AB_1', AB_2, B_3, \ldots, AB_n) \\
&\qquad + c'' \det_n(AB_1'', AB_2, AB_3, \ldots, AB_n) \\
&= c' f_A(B_1', B_2, B_3, \ldots, B_n) \\
&\qquad + c'' f_A(B_1'', B_2, B_3, \ldots, B_n)
\end{aligned}$$

So $f_A(B)$ is an R linear function of the first column of B. The same calculation shows that $f_A(B)$ is also a linear function of the other $n-1$ columns of B. Therefore $f_A \colon M_{n\times n}(R) \to R$ is n linear over R.

If two columns of B are the same, say $B_k = B_l$ with $k < l$, then as $AB = [AB_1, AB_2, \ldots, AB_k, \ldots, AB_l, \ldots, AB_n]$ and thus the k-th and l-th column of AB are also equal. Therefore, using that $\det_n$ is alternating, $f_A(B) = \det_n(AB) = 0$. This shows that f_A is alternating. We can now use Theorem 4.14 and conclude

$$\begin{aligned}
\det_n(AB) &= f_A(B) = \det_n(B) f_A(I_n) \\
&= \det_n(B)\det_n(AI_n) = \det_n(B)\det_n(A) \\
&= \det_n(A)\det_n(B)
\end{aligned}$$

as required. □

4.4.2. *Expanding determinants along rows and the determinant of the transpose.* Here is another application of the uniqueness theorem.

Theorem 4.17. *The determinant can expanded along any of its rows. That is for $A = [a_{ij}] \in M_{n\times n}(R)$*

$$\det_n(A) = \sum_{j=1}^{n+1} (-1)^{i+j} a_{ij} \det_{n-1}(A[ij]) \tag{4.11}$$

which is the formula for expansion along the i-th row.

Proof. Using the notation of equation (4.5) we wish to show that $\det_n = D_i \det_{n-1}$. But if set $f = D_i \det_{n-1}$, then Proposition 4.6 (applied to the function $\det_{n-1}$) implies that f is alternating, n linear and that $f(I_n) = \det_{n-1}(I_{n-1}) = 1$. Therefore by Theorem 4.14 we have $f(A) = \det_n(A)$. $\square$

We now show that the determinant of a matrix and its transpose are equal. If we use of Proposition 4.12 to compute we get a sum of products

$$\operatorname{sgn}(\sigma) a_{\sigma(1)1} a_{\sigma(2)2} a_{\sigma(3)3} \cdots a_{\sigma(n)n}.$$

If $(i,j) = (\sigma(j), j)$ then have $i = \sigma(j)$, or what is the same thing $j = \sigma^{-1}(i)$, so that $a_{ij} = a_{\sigma(j)j} = a_{\sigma_{i\sigma^{-1}(i)}}$. So we reorder the terms in the product so that the first index in a_{ij} is in increasing order. Then we have

$$\begin{aligned}\operatorname{sgn}(\sigma) &a_{\sigma(1)1} a_{\sigma(2)2} a_{\sigma(3)3} \cdots a_{\sigma(n)n} \\ &= \operatorname{sgn}(\sigma) a_{1\sigma^{-1}(1)} a_{2\sigma^{-1}(2)} a_{3\sigma^{-1}(3)} \cdots a_{n\sigma^{-1}(n)}.\end{aligned}$$

(This is a product of exactly the same terms, just in a different order.) But we also have

Problem 40. For all $\sigma \in S_n$ show $\operatorname{sgn}(\sigma^{-1}) = \operatorname{sgn}(\sigma)$. $\square$

and therefore

$$\begin{aligned}\operatorname{sgn}(\sigma) &a_{\sigma(1)1} a_{\sigma(2)2} a_{\sigma(3)3} \cdots a_{\sigma(n)n} \\ &= \operatorname{sgn}(\sigma^{-1}) a_{1\sigma^{-1}(1)} a_{2\sigma^{-1}(2)} a_{3\sigma^{-1}(3)} \cdots a_{n\sigma^{-1}(n)}.\end{aligned}$$

Using this in Proposition 4.12 and doing the change of variable $\tau = \sigma^{-1}$ in the sum gives for $A = [a_{ij}] \in M_{n\times n}(R)$ that

$$\begin{aligned}\det_n(A) &= \sum_{\sigma \in S_n} \operatorname{sgn}(\sigma^{-1}) a_{1\sigma^{-1}(1)} a_{2sigma^{-1}(2)} a_{3\sigma^{-1}(3)} \cdots a_{n\sigma^{-1}(n)} \\ &= \sum_{\tau \in S_n} \operatorname{sgn}(\tau) a_{1\tau(1)} a_{2\tau(2)} a_{3\tau(3)} \cdots a_{n\tau(n)} \\ &= \sum_{\tau \in S_n} \operatorname{sgn}(\tau) b_{\tau(1)1} b_{\tau(2)2} b_{\tau(3)3} \cdots b_{\tau(n)n} \\ &= \det_n(B)\end{aligned}$$

where $b_{ij} = a_{ji}$. That is $B = A^t$, the transpose of A. Thus we have proven:

Proposition 4.18. *For any $A \in M_{n\times n}(R)$ we have $\det_n(A^t) = \det_n(A)$. As taking the transpose interchanges rows and columns of A this implies that $\det_n(A)$ is also a alternating n linear function of the* rows *of A.* $\square$

Note that applying Theorem 4.17 to the transpose of $A = a_{ij}$ gives

$$\det_n(A) = \sum_{i=1}^{n+1} (-1)^{i+j} a_{ij} \det_{n-1}(A[ij]) \tag{4.12}$$

which is the formula for expanding A along a column.

Problem 41. Show that (4.12) can also be derived directly from the facts that $\det_n$ alternating and an n linear functions of its columns. $\square$

4.5. **The classical adjoint and inverses.** If R is a commutative ring and $A = [a_{ij}] \in M_{n\times n}(R)$ the ***classical adjoint*** is the matrix $\operatorname{adj}(A) \in M_{n\times n}(R)$ with elements

$$\operatorname{adj}(A)_{ij} = (-1)^{i+j} \det_{n-1}(A[ji]).$$

Note the interchange of order of i and j so that this is the transpose of the matrix $[(-1)^{i+j} \det_{n-1}(A[ij])]$. In less compact notation if

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

then

$$\operatorname{adj}(A) = \begin{bmatrix} +\det(A[11]) & -\det(A[21]) & +\det(A[31]) & -\det(A[41]) & \cdots \\ -\det(A[12]) & +\det(A[22]) & -\det(A[32]) & +\det(A[42]) & \cdots \\ +\det(A[13]) & -\det(A[23]) & +\det(A[33]) & -\det(A[43]) & \cdots \\ -\det(A[14]) & +\det(A[24]) & -\det(A[34]) & +\det(A[44]) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}$$

(where $\det = \det_{n-1}$).

This is important because of the following result.

Theorem 4.19. *Let R be a commutative ring. Then for any $A \in M_{n\times n}(R)$ we have*

$$\operatorname{adj}(A)A = A\operatorname{adj}(A) = \det_n(A)I_n.$$

Proof. Letting $A = [a_{ij}]$, the entries of $A\operatorname{adj}(A)$ are

$$\begin{aligned}(A\operatorname{adj}(A))_{ik} &= \sum_{j=1}^{n} a_{ij} \operatorname{adj}(A)_{jk} \\ &= \sum_{j=1}^{n} (-1)^{j+k} a_{ij} \det_{n-1}(A[kj]).\end{aligned}$$

Now if we let $k = i$ in this and use (4.11) (the expansion for $\det_n(A)$ along the i row) we get

$$(A\,\mathrm{adj}(A))_{ii} = \sum_{j=1}^{n}(-1)^{j+i}a_{ij}\det_{n-1}(A[ij]) = \det_n(A).$$

If $k \neq i$ then let $B = [b_{ij}]$ have all its rows the same as the rows of A, except that the k-th row is replaced by the i-the row of A (thus A and B only differ along the k-the row). Then B has two rows the same and so $\det_n(B) = 0$. (For the transpose B^t has two columns the same and so $\det_n(B) = \det_n(B^t) = 0$). Now for all j that $B[kj] = A[kj]$ as A and only differ in the k-th row and $A[kj]$ and $B[kj]$ only involve elements of A and B not on the k-row. Also from the definition of B we have $b_{kj} = a_{ij}$ (as the k-th row of B is the same as the i-row of A). Therefore we can compute $\det_n(B)$ by expanding along the k row

$$\begin{aligned}
0 = \det_n(B) &= \sum_{j=1}^{n}(-1)^{j+k}b_{kj}\det_{n-1}(B[kj]) \\
&= \sum_{j=1}^{n}(-1)^{j+k}a_{ij}\det_{n-1}(A[kj]) \\
&= (A\,\mathrm{adj}(A))_{ik}.
\end{aligned}$$

These calculations can be summarized as

$$(A\,\mathrm{adj}(A))_{ik} = \det_n(A)\delta_{ik}.$$

But this implies $A\,\mathrm{adj}(A) = \det_n(A)I_n$.

A similar computation (but working with columns rather than rows) implies that $\mathrm{adj}(A)A = \det_n(A)I_n$.

Problem 42. Write out the details that $\mathrm{adj}(A)A = \det_n(A)I_n$. □

This completes the proof. □

Remark 4.20. It is possible to shorten the last proof by proving directly that $A\,\mathrm{adj}(A) = \det_n(A)I_n$ implies that $\mathrm{adj}(A)A = \det_n(A)I_n$ by using that on matrices $(AB)^t = B^tA^t$. It is not hard to see that $\mathrm{adj}(A^t) = \mathrm{adj}(A)^t$. Replacing A by A^t in $A\,\mathrm{adj}(A) = \det_n(A)I_n$ gives that $A^t\,\mathrm{adj}(A^t) = \det_n(A^t)I_n = \det_n(A)I_n$. Taking transposes of this gives

$$\begin{aligned}
\det_n(A)I_n &= (\det_n(A)I_n)^t = (A^t\,\mathrm{adj}(A^t))^t \\
&= \mathrm{adj}(A^t)^t(A^t)^t = \mathrm{adj}((A^t)^t)(A^t)^t = \mathrm{adj}(A)A
\end{aligned}$$

as required. □

Recall that a unit a in a ring R is an element that has an inverse. The following gives a necessary and sufficient condition for matrix $A \in M_{n\times n}(R)$ to have an inverse in terms of the determinant $\det_n(A)$ being a unit.

Theorem 4.21. *Let R be a commutative ring. Then $A \in M_{n\times n}(R)$ has an inverse in $M_{n\times n}(R)$ if and only if $\det_n(A)$ is a unit in R. When the inverse does exist it is given by*

$$A^{-1} = \frac{1}{\det_n(A)} \operatorname{adj}(A). \tag{4.13}$$

(A slightly more symmetric statement of this theorem would be that A has an inverse in $M_{n\times n}(R)$ if and only if $\det_n(A)$ has an inverse in R.)

Remark 4.22. Recall that in a field $\mathbf{F}$ that all nonzero elements have inverses. Therefore for $A \in M_{n\times n}(\mathbf{F})$ this reduces to the statement that A^{-1} exists if and only if $\det_n(A) \neq 0$. □

Proof. First assume $\det_n(A) \in R$ is a unit in R. Then $(\det_n(A))^{-1} \in R$ and thus $(\det_n(A))^{-1} \operatorname{adj}(A) \in M_{n\times n}(R)$. Using Theorem 4.19 we then have

$$\begin{aligned} ((\det_n(A))^{-1} \operatorname{adj}(A))A &= A((\det_n(A))^{-1} \operatorname{adj}(A)) \\ &= \det_n(A)^{-1} \det_n(A) I_n = I_n. \end{aligned}$$

Thus the inverse of A exists and is given by (4.13).

Conversely assume that A has an inverse $A^{-1} \in M_{n\times n}(R)$. Then $AA^{-1} = I_n$ and so

$$1 = \det_n(I_n) = \det_n(AA^{-1}) = \det_n(A)\det_n(A^{-1})$$

But $\det_n(A)\det_n(A^{-1}) = 1$ implies that $\det_n(A)$ is a unit with inverse $(\det_n(A))^{-1} = \det_n(A^{-1})$. This completes the proof. □

The following is basically just a corollary of the last result, but it is important enough to be called a theorem.

Theorem 4.23. *Let R be a commutative ring and $A, B \in M_{n\times n}(R)$. Then $AB = I_n$ implies $BA = I_n$.*

Remark 4.24. Here it is important that A and B be square. For example if

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

then

$$AB = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2, \quad \text{but} \quad BA = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \neq I_3.$$

□

Proof. If $AB = I_n$ then $1 = \det_n(I_n) = \det_n(AB) = \det_n(A)\det_n(B)$. Therefore $\det_n(A)$ is a unit in R with inverse $\det_n(A)^{-1} = \det_n(B)$. But the last theorem implies that A^{-1} exists. Thus $B = I_n B = (A^{-1}A)B = A^{-1}(AB) = A^{-1}I_n = A^{-1}$. But if $B = A^{-1}$ then clearly $BA = I_n$. □

4.6. **The Cayley-Hamilton Theorem.** We now use Theorem 4.19 to prove what is likely the most celebrated theorem in linear algebra. First we extend the definition of characteristic polynomial to the case of matrices with elements in a ring.

Definition 4.25. Let R be a commutative ring and let $A \in M_{n\times n}(R)$. Then the ***characteristic polynomial*** of A, denoted by $\operatorname{char}_A(x)$, is

$$\operatorname{char}_A(x) = \det_n(xI_n - A).$$

□

Maybe a little needs to be said about this. If R is a commutative ring the ***ring of polynomials*** $R[x]$ ***over*** R is defined in the obvious way. That is elements $f(x) \in R[x]$ are of the form

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n$$

where $a_0, \ldots, a_n \in R$. These are added, subtracted, and multiplied in the usual manner. Therefore $R[x]$ is also commutative ring. If $A \in M_{n\times n}(R)$ then $xI_n - A \in M_{n\times n}(R[x])$. In the definition of $\operatorname{char}_A(x)$ the determinant $\det_n(xI_n - A)$ is computed in the ring $R[x]$.

Proposition 4.26. *If $A \in M_{n\times n}(R)$ then the characteristic polynomial $\operatorname{char}_A(x)$ is a monic polynomial of degree n (with coefficients in R).*

Proof. Letting $e_1, \ldots, e_n$ be the standard basis of R^n and $A_1, \ldots, A_n$ the columns of A we write

$$\begin{aligned} xI_n - A &= x[e_1, e_2, \ldots, e_n] - [A_1, A_2, \ldots, A_n] \\ &= [xe_1 - A_1, xe_2 - A_2, \ldots, xe_n - A_n]. \end{aligned}$$

Then expand $\det_n(xI_n - A) = \det_n(xe_1 - A_1, xe_2 - A_2, \ldots, xe_n - A_n)$ and group by powers of x. Each factor in the product is of first degree in x, so expanding a product of n factors will lead to a degree n expression. The coefficient of x^n is $\det_n(e_1, e_2, \ldots, e_n) = \det_n(I_n) = 1$ so this polynomial is monic. This basically completes the proof. But for the skeptics, or those not use to this type of calculation, here is more detail.

We first do this for $n = 3$ to see what is going on

$$\begin{aligned}
\operatorname{char}_A(x) &= \det_3(xe_1 - A_1, xe_2 - A_2, e_3 - A_3) \\
&= x^3 \det_3(e_1, e_2, e_3) \\
&\qquad - x^2\big(\det_3(A_1, e_2, e_3) + \det_3(e_1, A_2, e_3) + \det_3(e_1, e_2, A_3)\big) \\
&\qquad + x\big(\det_3(A_1, A_2, e_3) + \det_3(A_1, e_2, A_3) + \det_3(e_1, A_2, A_3)\big) \\
&\qquad - \det_3(A_1, A_2, A_3) \\
&= x^3 + a_2x^2 + a_1x + a_0
\end{aligned}$$

where

$$\begin{aligned}
a_2 &= -\big(\det_3(A_1, e_2, e_3) + \det_3(e_1, A_2, e_3) + \det_3(e_1, e_2, A_3)\big) \\
a_1 &= \det_3(A_1, A_2, e_3) + \det_3(A_1, e_2, A_3) + \det_3(e_1, A_2, A_3) \\
a_0 &= -\det_3(A_1, A_2, A_3) = -\det_3(A).
\end{aligned}$$

In the general case:

$$\begin{aligned}
\operatorname{char}_A(x) &= \det_n(xe_1 - A_1, xe_2 - A_2, \dots, xe_n - A_n) \\
&= x^n \det_n(e_1, e_2, , \dots, e_n) \\
&\qquad - x^{n-1} \sum_{j=1}^{n} \det_n(e_1, e_2, \dots, A_j, \dots, e_n) \\
&\qquad + x^{n-2} \sum_{1 \le j_1 < j_2 \le n} \det_n(e_1, e_2, \dots, A_{j_1}, \dots, A_{j_2}, \dots, e_n) \\
&\qquad\quad \vdots \\
&\qquad (-1)^n \det_n(A_1, A_2, \dots, A_n) \\
&= x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + (-1)^n a_0
\end{aligned}$$

where

$$a_{n-k} = (-1)^k \sum_{1 \le j_1 < j_2 < \cdots < j_k \le n} \det_n(\dots, A_{j_1}, \dots, A_{j_2}, \dots, A_{j_k}, \dots).$$

(The term in this sum the term corresponding to $j_1 < j_2 < \cdots < j_k$ has for its columns in the k places $j_1, j_2, \dots, j_k$ the corresponding columns of A and in all other places the corresponding columns of $I_n = [e_1, \dots, e_n]$.) This shows $\operatorname{char}_A(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_0$ which is a polynomial of the desired form. □

Now consider what happens when we use the matrix $xI - A$ in Theorem 4.19. We get

$$\begin{aligned} \operatorname{adj}(xI_n - A)(xI_n - A) &= (xI_n - A)\operatorname{adj}(xI_n - A) \\ &= \det_n(xI_n - A)I_n = \operatorname{char}_A(x)I_n. \end{aligned} \tag{4.14}$$

The matrix $\operatorname{adj}(xI_n - A)$ will be a polynomial in x with coefficients which are $n \times n$ matrices out of R. Write it as

$$\operatorname{adj}(xI_n - A) = x^k B_k + x^{k-1}B_{k-1} + \cdots + B_0$$

with $B_k \neq 0$. Then leading term of $\operatorname{adj}(xI_n - A)(xI_n - A)$ is $x^{k+1}B_k + \cdots$ so we have that $\operatorname{adj}(xI_n - A)(xI_n - A)$ is of degree $k+1$ but then $\operatorname{adj}(xI_n - A)(xI_n - A) = \operatorname{char}_A(x)I_n$ implies that $k+1 = n$ (as $\operatorname{char}_A(x)I_n$ has degree n). Thus $\operatorname{adj}(xI_n - A)$ has degree $n-1$. (This could also be seen using the definition of $\operatorname{adj}(xI_n - A)$ as a matrix whose elements are determinant of order $n-1$ and using an argument like that of the proof of Proposition 4.26.) If $n = 4$ and we let the characteristic polynomial of A be

$$\operatorname{char}_A(x) = x^4 + a_3x^3 + a_2x^2 + a_1x + a_0$$

and

$$\operatorname{adj}(xI_4 - A) = B_3x^3 + B_2x^2 + B_1x^1 + B_0.$$

Then

$$\begin{aligned} (xI_4 - A)\operatorname{adj}(xI_4 - A) &= (xI_4 - A)(B_3x^3 + B_2x^2 + B_1x^1 + B_0) \\ &= B_3x^4 + (B_2 - B_3A)x^3 + (B_1 - B_2A)x^2 + (B_0 - B_1A)x - B_0A \end{aligned}$$

But by (4.14)

$$(xI_4 - A)\operatorname{adj}(xI_4 - A) = \operatorname{char}_A(x)I_4 = (x^4 + a_3x^3 + a_2x^2 + a_1x + a_0)I_4.$$

Equating the coefficients of powers of x in the two expressions for $(xI_4 - A)\operatorname{adj}(xI_4 - A)$ gives

$$\begin{aligned} a_0I_4 &= -B_0A \\ a_1I_4 &= B_0 - B_1A \\ a_2I_4 &= B_1 - B_2A \\ a_3I_4 &= B_2 - B_3A \\ I_4 &= B_3. \end{aligned} \tag{4.15}$$

Multiply the second of these on the right by A, the third on the right by A^2, the fourth by A^3 and the last by A^4. The result is

$$\begin{aligned} a_0 I_4 &= -B_0 A \\ a_1 A &= B_0 A - B_1 A^2 \\ a_2 A^2 &= B_1 A^2 - B_2 A^3 \\ a_3 A^3 &= B_2 A^3 - B_3 A^4 \\ A^4 &= B_3 A^4. \end{aligned}$$

Now add these equations. On the right side the terms "telescope" (i.e. each term and its negative appear just once) so that after adding we get

$$A^4 + a_3 A^3 + a_2 A^2 + a_1 A + a_0 I_4 = 0.$$

The left side of this is just the characteristic polynomial, $\operatorname{char}_A(x)$, of A evaluated at $x = A$. That is

$$\operatorname{char}_A(A) = 0.$$

No special properties of $n = 4$ were used in this derivation so we have linear algebra's most famous result:

Theorem 4.27 (Cayley-Hamilton Theorem)**.** *Let R be a commutative ring, $A \in M_{n\times n}(R)$ and let $\operatorname{char}_A(x) = \det_n(xI_n - A)$ be the characteristic polynomial of A. Then A is a root of $\operatorname{char}_A(x)$. That is*

$$\operatorname{char}_A(A) = 0.$$

Problem 43. Prove this along the following lines: Write the characteristic polynomial as

$$\operatorname{char}_A(x) = x^n + a_{n-1}x^{n-1} + a_{n-2}x^{n-2} + \cdots + a_1 x + a_0$$

and write $\operatorname{adj}(xI_n - A)$ as

$$\operatorname{adj}(xI_n - A) = B_{n-1}x^{n-1} + B_{n-2}x^{n-2} + \cdots B_1 x + B_0.$$

Show then that equating coefficients of x in $(xI_n - A)\,\mathrm{adj}(xI_n - A) = \mathrm{char}_A(x)$ (cf. (4.14)) gives the equations

$$\begin{aligned}
a_0 I_n &= -B_0 A \\
a_1 I_n &= B_0 - B_1 A \\
a_2 I_n &= B_1 - B_2 A \\
\vdots &= \quad \vdots \\
a_{n-2} I_n &= B_{n-3} - B_{n-2} A \\
a_{n-1} I_n &= B_{n-2} - B_{n-1} A \\
I_n &= B_{n-1}.
\end{aligned}$$

Multiply these equations on the right by appropriate powers of A to get

$$\begin{aligned}
a_0 I_n &= -B_0 A \\
a_1 A &= B_0 A - B_1 A^2 \\
a_2 A^2 &= B_1 A^2 - B_2 A^3 \\
\vdots &= \quad \vdots \\
a_{n-2} A^{n-2} &= B_{n-3} A^{n-2} - B_{n-2} A_{n-1} \\
a_{n-1} A_{n-1} &= B_{n-2} A^{n-1} - B_{n-1} A^n \\
A^n &= B_{n-1} A^n.
\end{aligned}$$

Finally add these to get

$$A^n + a_{n-1}A^{n-1} + a_{n-2}A^{n-2} + \cdots + a_2 A^2 + a_1 A + a_0 I_n = 0.$$

as required. □

Problem 44. Assume that $A \in M_{n\times n}(R)$ and that $\det_n(A)$ is a unit in R. Then use the Cayley-Hamilton Theorem to show that the inverse A^{-1} is a polynomial in A. *Hint:* Let the characteristic polynomial be given by $\mathrm{char}_A(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0$. Then evaluation at $x = 0$ shows that $a_0 = \mathrm{char}_A(0) = \det_n(-A) = (-1)^n \det_n(A)$. The Cayley-Hamilton Theorem yields that

$$A^n + a_{n-1}A^{n-1} + a_{n-2}A^{n-2} + \cdots + a_1 A + a_0 I_n = 0$$

which can then be rewritten as

$$A(A^{n-1} + a_{n-1}A^{n-2} + \cdots + a_1 I_n) = -a_0 I_n = (-1)^n \det_n(A) I_n$$

□

Problem 45. In the system of equation (4.15) for B_0, B_1, B_2, B_3 in the $n = 4$ case we can back solve for the B_k's and get

$$\begin{aligned}
B_3 &= I_4\\
B_2 &= a_3 I_4 + B_3 A = a_3 I_4 + A\\
B_1 &= a_2 I_4 + B_2 A = a_2 I_4 + a_3 A + A^2\\
B_0 &= a_1 I_4 + B_1 A = a_1 I_4 + a_2 A + a_3 A^2 + A^3
\end{aligned}$$

Show that in the general case the formulas $B_{n-1} = I_n$ and

$$\begin{aligned}
B_k &= a_{k+1} I_n + a_{k+2} A + a_{k+3} A^2 + \cdots + a_{n-k-1} A^{n-k-2} + A^{n-k-1}\\
&= \sum_{j=0}^{n-k-1} a_{k+1+j} A^j
\end{aligned}$$

hold for $k = 0, \ldots, n-2$. □

4.7. **Sub-matrices and sub-determinants.** The results here are somewhat messy, but are needed in the uniqueness proof for the Smith normal form in Section 5.4. The reader will lose little by skipping thus section until needed. And the reader that is willing to take the uniqueness of the Smith Normal form on faith, can skip it altogether.

4.7.1. *The definition of sub-matrix and sub-determinant.* Let R be a commutative ring and $A \in M_{m\times n}$. We wish to define sub-matrices of A. Informally these are the results of crossing out some rows and columns of A and what is left is a sub-matrix. To be more precise let $1 \le k \le m$ and $1 \le \ell \le n$. Then for finite increasing sequences

$$\begin{aligned}
K &= (i_1, i_2, \ldots, i_k) \quad \text{with} \quad 1 \le i_1 < i_2 < \cdots < i_k \le m,\\
L &= (j_1, j_2, \ldots, j_\ell) \quad \text{with} \quad 1 \le j_1 < j_2 < \cdots < j_\ell \le n.
\end{aligned}$$

Then the ***sub-matrix*** $A_{K,L}$ is

$$A_{K,L} := \begin{bmatrix} a_{i_1 j_1} & a_{i_1 j_2} & \cdots & a_{i_1 j_\ell}\\ a_{i_2 j_1} & a_{i_2 j_2} & \cdots & a_{i_2 j_\ell}\\ \vdots & \vdots & \ddots & \vdots\\ a_{i_k j_1} & a_{i_k j_2} & \cdots & a_{i_k j_\ell} \end{bmatrix}.$$

Thus $A_{K,L}$ is the matrix that has elements a_{ij} with i in K and j in L. As a concrete example let

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \end{bmatrix}$$

and

$$K = (2,4), \quad L = (2,3,5).$$

Then

$$A_{K,L} = A_{(2,3),(2,3,5)} = \begin{bmatrix} a_{22} & a_{23} & a_{25} \\ a_{42} & a_{43} & a_{45} \end{bmatrix}.$$

We will write

$$|K| = k, \quad |L| = \ell$$

to be the number of elements in the lists K and L respectively. If $|K| = k$ and $|L| = \ell$ then $A_{K,L}$ is a $k \times \ell$ ***sub-matrix*** of A. If $|K| = |L|$ then $A_{I,L}$ will have the same number of rows and columns and so in this case $A_{I,L}$ is called a ***square sub-matrix*** of A.

For $|K| = |L|$ we can take the determinant of $A_{K,L}$ (this works even when the original matrix A is not square). Then the $K - L$th ***sub-determinant*** of A is $\det A_{K,L}$. When $A \in M_{n\times n}(R)$ is square, and $K = L$ then $A_{K,L} = A_{K,K}$ is called a ***principal sub-matrix*** of A and $\det A_{K,K}$ is a ***principal sub-determinant*** of A.

Problem 46. If $A \in M_{m\times n}(R)$ then show that the number of $k \times \ell$ sub-matrices of A is the product $\binom{m}{k}\binom{n}{\ell}$ of binomial coefficients. If $A \in M_{n\times n}(R)$ is square then show the number of $k \times k$ principal sub-matrices is $\binom{n}{k}$. □

We can use the idea of principal sub-determinants to give a formula for the coefficients of the characteristic polynomial of a matrix.

Proposition 4.28. *Let R be a commutative ring and $A \in M_{n\times n}(R)$ a square matrix over R. Let the characteristic polynomial of A be*

$$\operatorname{char}_A(x) = \det(xI - A) = x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0.$$

Then

$a_k = (-1)^{n-k} \times$ *the sum of the $k \times k$ principle sub-determinants of A.*

Problem 47. Prove this. *Hint:* This is contained more or less explicitly in the proof of Proposition 4.26. □

For example if

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 4 \\ 4 & 9 & 16 \end{bmatrix}$$

then Proposition 4.28 implies

$$\operatorname{char}_A(x) = x^3 + a_2 x^2 + a_1 x + a_0$$

where

$$a_0 = 1 \det(A) = -2$$

$$a_1 = \text{sum of } 1 \times 1 \text{ principalsub-determinants} = 1 + 3 + 16 = 20.$$

$$\begin{aligned} a_2 &= -\left(\det \begin{bmatrix} 1 & 1 \\ 2 & 3 \end{bmatrix} + \det \begin{bmatrix} 1 & 1 \\ 4 & 16 \end{bmatrix} + \det \begin{bmatrix} 3 & 4 \\ 9 & 16 \end{bmatrix} \right) \\ &= -(1 + 12 + 12) = -25. \end{aligned}$$

The following trivial result will be useful later (for example in showing that over a field that a square matrix and its transpose are always similar).

Proposition 4.29. *Let $A \in M_{m\times n}(R)$. Then for all K, L for which the sub-matrix $A_{K,L}$ is defined we have that the transpose is given by*

$$(A_{K,L})^t = (A^t)_{L,K}.$$

Thus as the determinant of a square matrix equals the determinant of the transpose of the matrix we have that $|K| = |L|$ implies

$$\det A_{K,L} = \det(A^t)_{L,K}.$$

Problem 48. Prove this. □

4.7.2. *The ideal of $k \times k$ sub-determinants of a matrix.* Let R be a commutative ring and $A \in M_{m\times n}(R)$. For reasons that will only become clear when we look at the uniqueness of the Smith normal form, we wish to look at the ideal generated by the set of all $k \times k$ sub-determinants of A. Recall (see Proposition and Definition 1.9) the definition of the ideal generated by a finite collection of elements of a ring.

Definition 4.30. Let $A \in M_{m\times n}(R)$. Then define

$$\mathcal{I}_k(A) := \text{ideal of } R \text{ generated by the } k \times k \text{ sub-determinants of } A.$$

where $1 \le k \le \min\{m, n\}$. □

As an example $R = \mathbf{Z}$ be be the ring of integers and let

$$A = \begin{bmatrix} 4 & 6 \\ 8 & 10 \\ 14 & 12 \end{bmatrix}.$$

Recall, Theorem 2.16, that in a principal ideal domain the ideal generated by a finite set, is just the principal ideal generated by the greatest common divisor of the elements. The 1×1 sub-determinants of A are just its elements. Thus

$$\mathcal{I}_1(A) = \langle 4, 6, 8, 10, 12, 14 \rangle = \langle 2 \rangle,$$

and

$$\begin{aligned} \mathcal{I}_2(A) &= \left\langle \det \begin{bmatrix} 4 & 6 \\ 8 & 10 \end{bmatrix}, \det \begin{bmatrix} 4 & 6 \\ 14 & 12 \end{bmatrix}, \det \begin{bmatrix} 8 & 10 \\ 14 & 12 \end{bmatrix} \right\rangle \\ &= \langle -8, -36, -44 \rangle \\ &= \langle 4 \rangle. \end{aligned}$$

The first result about the ideals $\mathcal{I}_k(A)$ is trivial.

Proposition 4.31. *If $A \in M_{m\times n}(R)$ and $1 \le k \le \min\{m, n\}$ then*

$$\mathcal{I}_k(A) = \mathcal{I}_k(A^t).$$

That is the ideals $\mathcal{I}_k$ are the same for a matrix and its transpose.

Problem 49. Prove thus. *Hint:* Proposition 4.29. □

We now wish to understand what happens to the ideals under matrix multiplication. First some notation. Let $1 \le k \le \min\{m, n\}$ and let K, L

$$\begin{aligned} K &= (i_1, i_2, \dots, i_k) \quad \text{with} \quad 1 \le i_1 < i_2 < \cdots < i_k \le m, \\ L &= (j_1, j_2, \dots, j_k) \quad \text{with} \quad 1 \le j_1 < j_2 < \cdots < j_k \le n. \end{aligned}$$

and, as above, let $A_{K,L}$ be the $k \times k$ sub-matrix

$$A_{K,L} := \begin{bmatrix} a_{i_1 j_1} & a_{i_1 j_2} & \cdots & a_{i_1 j_k} \\ a_{i_2 j_1} & a_{i_2 j_2} & \cdots & a_{i_2 j_k} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i_k j_1} & a_{i_k j_2} & \cdots & a_{i_k j_k} \end{bmatrix}.$$

For an element j_s of L let A_{K,j_s} be the sth column of $A_{K,L}$. That is

$$A_{K,j_s} = \begin{bmatrix} a_{i_1 j_s} \\ a_{i_2 j_s} \\ \vdots \\ a_{i_k j_s} \end{bmatrix}.$$

Then we can write $A_{K,L}$ in terms of its columns as

$$A_{K,L} = \left[A_{K,j_1}, A_{K,j_2}, \dots, A_{K,j_k} \right].$$

Let $P \in M_{n\times n}(R)$. Then if we write $A = \left[A_1, A_2, \dots, A_n\right]$ in terms of its columns and let $P = \left[p_{ij}\right]$ then the definition of matrix multiplication and some thought show that

$$AP = \left[\sum_{i=1}^{n} A_i p_{i1}, \sum_{i=1}^{n} A_i p_{i2}, \dots, \sum_{i=1}^{n} A_i p_{in}\right]$$

$$= \left[\sum_{i=1}^{n} p_{i1} A_i, \sum_{i=1}^{n} p_{i2} A_i, \dots, \sum_{i=1}^{n} p_{in} A_i\right]$$

Therefore the K, L square sub-matrix of AP is

$$(AP)_{K,L} = \left[\sum_{i=1}^{n} p_{ij_1} A_{K,i}, \sum_{i=1}^{n} p_{ij_2} A_{K,i}, \dots, \sum_{i=1}^{n} p_{ij_k} A_{K,i}\right]$$

Using that the determinant on $k\times k$ matrices is k linear on the columns and also an alternating function we can expand $\det((AP)_{K,L})$ in terms of the columns $A_{K,i}$ and use the alternating property to put the columns of the terms in increasing order of subscripts. The result of this is that

$$\det\left((AP)_{K,S}\right) = \sum_J p_J \det\left(A_{K,J}\right)$$

where the subscripts J range over all sequences $1 \le s_1 < s_2 < \dots s_k \le n$ and the ring elements p_J are all of the form $p_J = \pm \prod_{t=1}^{k} p_{is_t}$. This shows that any $k \times k$ sub-determinant $\det\left((AP)_{K,S}\right)$ of AP can be expressed as a linear combination of $k \times k$ sub-determinants of A. By Propostion 1.12 this implies that $\mathcal{I}_k(AP) \subseteq \mathcal{I}_k(A)$. We record this fact.

Lemma 4.32. *Let $A \in M_{m\times n}(R)$ and $P \in M_{n\times n}(R)$, Then the inclusion*

$$\mathcal{I}_k(AP) \subseteq \mathcal{I}_k(A)$$

holds for all k with $1 \le k \le \min\{m, n\}$. □

Theorem 4.33. *Let $A \in M_{m\times n}(R)$, $Q \in M_{m\times m}(R)$, and $P \in M_{n\times n}(R)$. Then*

$$\mathcal{I}_k(QAP) \subseteq \mathcal{I}_k(A)$$

holds for $1 \le k \le \min\{m, n\}$. If P and Q are invertible, then

$$\mathcal{I}_k(QAP) = \mathcal{I}_k(A)$$

Proof. We use Lemma 4.32 and that taking transposes does not change the ideas $\mathcal{I}_k$ (Proposition 4.31).

$$\begin{aligned}\mathcal{I}_k(QAP) &= \mathcal{I}_k((QA)P)\\ &\subseteq \mathcal{I}_k(QA) = \mathcal{I}_k((QA)^t) = \mathcal{I}_k(A^tQ^t)\\ &\subseteq \mathcal{I}_k(A^t) = \mathcal{I}_k(A).\end{aligned}$$

If P and Q are invertible, let $B = QAP$. Then $B = Q^{-1}AP^{-1}$ and so by what we have just shown

$$\mathcal{I}_k(A) \supseteq \mathcal{I}_k(QAP) = \mathcal{I}_k(B) \supseteq \mathcal{I}_k(Q^{-1}BP^{-1}) = \mathcal{I}_k(A).$$

Thus $\mathcal{I}_k(A) = \mathcal{I}_k(QAP)$. □

The following will not be used in what follows, but is of enough interest to be worth recording.

Proposition 4.34. *Let* $A \in M_{m\times n}(R)$*. Then*

$$\mathcal{I}_{k+1}(A) \subseteq \mathcal{I}_k(A)$$

for $1 \le k \le \min\{m, n\} - 1$*.*

Problem 50. Prove this. *Hint:* Take any $(k+1)\times(k+1)$ sub-determinant of A and expand it along its first column to express it as a linear combination of $k \times k$ sub-determinants of A. But if every $(k+1)\times(k+1)$ sub-determinant is a linear combination of $k\times k$ sub-determinants of A we have $\mathcal{I}_{k+1}(A) \subseteq \mathcal{I}_k(A)$. □

5. The Smith normal from.

5.1. **Row and column operations and elementary matrices in $M_{n\times n}(R)$.** Let R be a commutative ring and $A \in M_{m\times n}(R)$. Then we wish to simplify A by doing elementary row and column operations.

A ***type I elementary matrix*** is a square matrix of the form

$$E := \begin{bmatrix} 1 & & & \cdots & & & 0\\ & \ddots & & & & & \\ & & 1 & & & & \\ \vdots & & & u & & & \vdots\\ & & & & 1 & & \\ & & & & & \ddots & \\ 0 & & & \cdots & & & 1\end{bmatrix}$$

Where u is a unit in the (i,i) position.

Then is easy to check that the inverse of E is also a type I elementary matrix:

$$E^{-1} := \begin{bmatrix} 1 & & & \cdots & & & 0 \\ & \ddots & & & & & \\ & & 1 & & & & \\ \vdots & & & u^{-1} & & & \vdots \\ & & & & 1 & & \\ & & & & & \ddots & \\ 0 & & & \cdots & & & 1 \end{bmatrix}$$

Where u^{-1} exists as u is a unit.

We record for future use the effect of multiplying on the left or right by a type I elementary matrix.

Proposition 5.1. *Let $E \in M_{n\times n}(R)$ be an elementary matrix of type I as above. Then the inverse of E is also an elementary matrix of type I. If $A \in M_{n\times p}(R)$ and $B \in M_{m\times n}$ then EA is A with the i-th row multiplied by u and BE is BE with the i column multiplied by u.* $\square$

To be more explicit about what multiplication by E does if

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & & \vdots \\ a_{i1} & \cdots & a_{ip} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{np} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} b_{11} & \cdots & b_{1i} & \cdots & b_{1n} \\ \vdots & & \vdots & & \\ b_{m1} & \cdots & b_{mi} & \cdots & b_{mn} \end{bmatrix}$$

then

$$EA = \begin{bmatrix} a_{11} & \cdots & a_{1p} \\ \vdots & & \vdots \\ ua_{i1} & \cdots & ua_{ip} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{np} \end{bmatrix} \quad \text{and} \quad BE = \begin{bmatrix} b_{11} & \cdots & ub_{1i} & \cdots & b_{1n} \\ \vdots & & \vdots & & \\ b_{m1} & \cdots & ub_{mi} & \cdots & b_{mn} \end{bmatrix}.$$

Also if we take $u = 1$ in the definition of an elementary matrix of type I we see that the identity matrix I_n is an elementary matrix of type I.

An ***elementary row operation of type I*** on the matrix A is multiplying one of the rows of A by a unit of R. Likewise an ***elementary column operation of type I*** on the matrix A is multiplying one of the columns by a unit. Note that doing an elementary row or column operation of type I on A is the same as multiplying A by an elementary matrix of type I.

An ***elementary matrix of type II*** is just the identity matrix with two of its rows interchanged. Let $1 \leq i < j \leq n$ and E be the identity matrix with its i-th and j-th rows interchanged. Then

$$E = \begin{array}{cl} \begin{array}{cccccc} & & i\text{-th col.} & & j\text{-th col.} & \end{array} & \\ \begin{bmatrix} 1 & & & & & \\ & \ddots & & & & \\ & & 0 & & 1 & \\ & & & \ddots & & \\ & & 1 & & 0 & \\ & & & & & \ddots \\ & & & & & & 1 \end{bmatrix} & \begin{array}{l} \\ \\ i\text{-th row} \\ \\ j\text{-th row} \\ \\ \\ \end{array} \end{array}$$

Note that E is can also be obtained from interchanging the i-th and j-columns of I_n, so we could also have defined a type II elementary matrix to be the identity matrix with two of its columns interchanged. When $n = 2$ we have

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}^2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_2.$$

This calculation generalizes easily and we see for any elementary matrix of type II that $E^2 = I_n$. Thus E is invertible with $E^{-1} = E$. We summarize the basic properties of type II elementary matrices.

Proposition 5.2. *Let $E \in M_{n\times n}(R)$ be an elementary matrix of type II. Then the inverse of E is its own inverse. If $A \in M_{n\times p}(R)$ and $B \in M_{m\times n}$ then EA is A with the i-th and j-th rows interchanged and BE is B with the i-th and j-th columns interchanged.* □

An ***elementary row operation of type II*** on the matrix A interchanging is interchanging two of the rows of A. Likewise an ***elementary column operation of type II*** on the matrix A is interchanging two of the columns of A. Thus doing an elementary row or column operation of type II on A is the same as multiplying A by an elementary matrix of type II. Note that interchanging the i-th and j-th rows of a matrix twice leaves the matrix unchanged. This is another way of seeing that for an elementary matrix of type II that $E^2 = I$.

An ***elementary matrix of type III*** differs from the identity matrix by having one one off diagonal entry nonzero. If the off diagonal element is r appearing at the $i\,j$ place then E is

$$E = \begin{array}{c} \begin{matrix} & & & j\text{-th} & \\ & & & \text{col.} & \end{matrix} \\ \begin{bmatrix} 1 & & & & \\ & \ddots & & r & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 1 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ i\text{-th row.} \\ \\ \\ \\ \end{array}$$

(This is the form of E when $i < j$. If $j < i$ then then r is below the diagonal.)

If $A \in M_{n\times p}(R)$ then A has n rows. Let $A^1, \dots, A^n$ be the rows of A so that

$$A = \begin{bmatrix} A^1 \\ A^2 \\ \vdots \\ A^n \end{bmatrix}.$$

If E is the $n \times n$ elementary matrix of type III that has r in the $i\,j$ place (with $i \neq j$) then multiplying A on the left by E adds r times the j-th row of A to the i-th row and leaves the other rows unchanged. That is

$$EA = \begin{bmatrix} A^1 \\ \vdots \\ A^i + rA^j \\ \vdots \\ A^j \\ \vdots \\ A^n \end{bmatrix}$$

For example when $n = 4$, $i = 3$ and $j = 1$ this is

$$EA = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ r & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} A^1 \\ A^2 \\ A^3 \\ A^4 \end{bmatrix} = \begin{bmatrix} A^1 \\ A^2 \\ A^3 + rA^1 \\ A^4 \end{bmatrix}$$

If $B \in M_{m\times n}(R)$ then B has n columns, say $B = [B_1, B_2, \dots, B_n]$. Then multiplication of B on the right by E adds r times the i-th column of B to the j-th column and leaves the other columns unchanged. That is

$$\begin{aligned} BE &= [B_1, \dots, B_i, \dots, B_j, \dots, B_n]E \\ &= [B_1, \dots, B_i, \dots, B_j + rB_i, \dots, B_n] \end{aligned}$$

Again looking at the case of $n = 4$, $i = 3$ and $j = 1$ this is

$$\begin{aligned} BE &= [B_1, B_2, B_3, B_4] \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ r & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ &= [B_1 + rB_3, B_2, B_3, B_4]. \end{aligned}$$

As to the inverse of this 4×4 example just change the r to a $-r$:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ r & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -r & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

In general if E is the elementary matrix of type III with r in the $i\,j$-th place (with $i \neq j$) then the inverse, E^{-1}, of E is the elementary matrix of type III with $-r$ in the $i\,j$ place. This can also be seen as follows. Multiplication of A on the left by E adds r times the j-th row of A to the i-th row and leaves the other rows unchanged. If A' is the resulting matrix, then subtracting r times the j-th row of A' to the i-th row of A' is A (as the j-th row of A' is A_j and the i-th row of A' is $A_i + rA_j$).

An ***elementary row operation of type III*** on the matrix A interchanging is adding a scalar multiple of one row to another. Likewise an ***elementary column operation of type III*** on the matrix A is adding a scalar multiple of one column to another column. So doing an elementary row or column operation of type III on A is the same as multiplying A by an elementary matrix of type III.

Problem 51. Show the following:

(1) An elementary matrix of type I is the result of doing an elementary row operation of type I on the identity matrix I_n.
(2) An elementary matrix of type II is the result of doing an elementary row operation of type II on the identity matrix I_n.
(3) An elementary matrix of type III is the result of doing an elementary row operation of type III on the identity matrix I_n. □

Definition 5.3. An ***elementary matrix*** is a matrix that is an elementary matrix of type I, II, or III. □

Definition 5.4. An ***elementary row operation*** on a matrix is either an elementary row operation of type I, II, or III. An An ***elementary column operation*** on a matrix is either an elementary column operation of type I, II, or III. □

5.2. **Equivalent matrices in $M_{m\times n}(R)$.** We now wish to see how much we can simply matrices by doing row and column operations.

Definition 5.5. Let $A, B \in M_{m\times n}(R)$. Then

(1) A and B are ***row-equivalent*** iff B can be obtained from A by a finite number of elementary row operations.
(2) A and B are ***column-equivalent*** iff B can be obtained from A by a finite number of elementary column operations.
(3) A and B are ***equivalent*** iff B can be obtained from A by a finite number of of both row and column operations. We will use the notation $A \cong B$ to indicate that A and B are equivalent. □

Our discussion of the relationship between elementary row and column operations and multiplication by elementary matrices makes the following clear.

Proposition 5.6. *Let $A, B \in M_{m\times n}(R)$.*

(1) *A and B are row equivalent if and only if there is a finite sequence $P_1, P_2, \ldots, P_k$ elementary matrices of size $m \times m$ so that $B = P_kP_{k-1}\cdots P_1A$.*
(2) *A and B are row equivalent if and only if there is a finite sequence $Q_1, Q_2, \ldots, Q_k$ elementary matrices of size $n \times n$ so that $B = AQ_1Q_2\cdots Q_k$.*
(3) *A and B are equivalent if and only if there is a finite sequence $P_1, P_2, \ldots, P_k$ elementary matrices of size $m \times m$ and a finite sequence $Q_1, Q_2, \ldots, Q_l$ elementary matrices of size $n \times n$ so that $B = P_kP_{k-1}\cdots P_1AQ_1Q_2\cdots Q_l$.* □

Proposition 5.7. *All three of the relations of row-equivalence, column-equivalence, and equivalence are equivalence relations.*

Proof. We prove this for the case of equivalence, the other two cases being similar and a bit easier. We use the version of equivalence in terms of multiplication by elementary matrices given in Proposition 5.6. As I_m and I_n are elementary matrices and $A = I_mAI_n$ we have that $A \cong A$. Thus $\cong$ is reflective. If $A \cong B$ then there are elementary matrices $P_1, \ldots, P_k$ and $Q_1, \ldots, Q_l$ of the appropriate size so that $B = P_kP_{k-1}\cdots P_1AQ_1Q_2\cdots Q_l$. But we can solve for A and get $A = P_1^{-1}P_2^{-1}\cdots P_k^{-1}BQ_l^{-1}\cdots Q_2^{-1}Q_1^{-1}$. As the inverse of an elementary matrix is also an elementary matrix, this implies $B \cong A$. Therefore

$\cong$ is symmetric. Finally if $A \cong B$ and $B \cong C$ then there are elementary matrices $P_1, \ldots, P_k$, $P'_1, \ldots, P'_{k'}$, $Q_1, \ldots, Q_l$, and $Q'_1, \ldots, Q'_{l'}$ so that $B = P_k P_{k-1} \cdots P_1 A Q_1 Q_2 \cdots Q_l$ and $C = P'_{k'} \cdots P'_1 B Q'_1 \cdots Q'_{l'}$. Therefore

$$C = P'_{k'} \cdots P'_1 P_k P_{k-1} \cdots P_1 A Q_1 Q_2 \cdots Q_l Q'_1 \cdots Q'_{l'}$$

which shows that $A \cong C$. This shows that $\cong$ is transitive and completes the proof. $\square$

5.3. **Existence of the Smith normal form.** Our goal is to simplify matrices $A \in M_{m\times n}(R)$ as much as possible by use of elementary row and columns. For general rings this is a hard problem, but in the case that R is a Euclidean domain (which for us means the integers, $\mathbf{Z}$, or the polynomials, $\mathbf{F}[x]$, over a field $\mathbf{F}$) this has a complete solution: Every matrix $A \in M_{m\times n}(R)$ is equivalent to a diagonal matrix. Moreover by requiring that the diagonal elements satisfy some extra conditions on the diagonal elements this diagonal form is unique. This will allow us to understand when two matrices over a field are similar as $A, B \in M_{n\times n}(\mathbf{F})$ are similar if and only if they have the matrices $xI_n - A$ and $xI_n - B$ are equivalence in $M_{n\times n}(\mathbf{F}[x])$ (cf. Theorem 6.1).

Before stating the basic result we recall that if R is a commutative, and $a, b \in R$ then we write $a \mid b$ to mean that "a divides b" (cf. 2.6).

Theorem 5.8 (Existence of Smith normal form)**.** *Let R be an Euclidean domain. Then every $A \in M_{m\times n}(R)$ is equivalent to diagonal matrix of the form*

$$S = \begin{bmatrix} f_1 & & & & & \\ & f_2 & & & & \\ & & \ddots & & & \\ & & & f_r & & \\ & & & & 0 & \\ & & & & & \ddots \end{bmatrix} \qquad \textit{This is } m\times n \textit{ and all off diagonal elements are } 0.$$

where $f_1 \mid f_2 \mid \cdots \mid f_{r-1} \mid f_r$.

Definition 5.9. The matrix S is called a ***Smith normal form*** of A. (Uniqueness will be discussed in Theorem 5.12.) $\square$

Remark 5.10. This result also holds when R is a principle ideal domain. However it is not true for general rings. It is straightforward to show if R is a ring where each $A \in M_{m\times n}(R)$ has a Smith normal form, then every finitely generated ideal of R (that is an ideal of the form $\langle a_1, a_2, \dots, a_n\rangle$ for some $a_1, a_2, \dots, a_n \in R$) is principle.[6] This is a strong condition on a ring and fails even for such concrete and interesting examples as $\mathbf{Z}[x]$ and $\mathbf{F}[x, y]$. A ring R where every matrix A over R has a Smith normal form is called a ***elementary divisor ring***. A detailed study of such rings was made by Irving Kaplansky in [1]. $\square$

Proof. We use induction on $m + n$. The base case is $m + n = 2$ in which case the matrix A is 1×1 and there is nothing to prove. So let $A \in M_{m\times n}(R)$ and assume that the result is true for all matrices in any $M_{m'\times n'}(R)$ where $m' + n' < m + n$. If $A = 0$ then A is already in the required form, so assume that $A \neq 0$. Let $\delta\colon R \to \{0, 1, 2, \dots\}$ be as in the definition of Euclidean domain and let $\mathcal{A}$ be the set of all entries of elements of matrices equivalent to A, and let $f_1 \in \mathcal{A}$ be a nonzero element of $\mathcal{A}$ that minimizes δ. That is $\delta(f_1) \le \delta(a)$ for all $0 \neq a \in \mathcal{A}$. (Recall that $\delta(0)$ is undefined, so we leave it out of the competition for minimizer.) Let B be a matrix equivalent to A that has f_1 as an element. If f_1 is in the i, j-th place of B, then we can can interchange the first and i-th row of B and then the first and j-th column of B and assume that f_1 is in the $1, 1$ place of B. (Interchanging rows and columns are elementary row and column operations and so the resulting matrix is still equivalent to A.) So B is of the form

$$B = \begin{bmatrix} f_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ b_{21} & b_{22} & b_{23} & \cdots & b_{2n} \\ b_{31} & b_{32} & b_{33} & \cdots & b_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{m1} & b_{m2} & b_{m3} & \cdots & b_{mn} \end{bmatrix}$$

We can use the division algorithm in R to find a quotient and remainder when the elements $b_{21}, b_{31}, \dots, b_{m1}$ of the first column are divided by f_1. That is there are $q_2, \dots, q_m, r_2, \dots, r_m \in R$ so that $b_{i1} = q_i f_1 + r_i$ where either $r_i = 0$ or $\delta(r_i) < \delta(f_1)$. Then $r_i = b_{i1} - q_j f_1$. Now doing the $m - 1$ row operations of taking $-q_i$ times the first row of A and adding to the i-th row we get that B (and thus also A) is equivalent

$$\begin{bmatrix} f_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ b_{21} - q_2 f_1 & * & * & \cdots & * \\ b_{31} - q_3 f_1 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ b_{m1} - q_m f_1 & * & * & \cdots & * \end{bmatrix} = \begin{bmatrix} f_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ r_2 & * & * & \cdots & * \\ r_3 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ r_m & * & * & \cdots & * \end{bmatrix}$$

[6] Let A be the row vector $A = [a_1, a_2, \dots, a_n]$. If A has a Smith normal form, there are invertible $P \in M_{1\times 1}(R)$ and $Q \in M_{n\times n}(R)$ such that $PAQ = [f_1, 0, \dots, 0]$. Using Corollary 1.13 it is not hard to check $\langle a_1, \dots, a_n\rangle = \langle f_1\rangle$.

where $*$ is used to represent unspecified elements of R. As this matrix is equivalent to A and by the way that f_1 was chosen we must have $r_2 = r_3 = \cdots r_m = 0$ (as otherwise $\delta(r_j) < \delta(f_1)$ and f_1 was choosen so that $\delta(f_1) \le \delta(b)$ for any nonzero element of a matrix equivalent to A). Thus our matrix is of the form

$$\begin{bmatrix} f_1 & b_{12} & b_{13} & \cdots & b_{1n} \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix}$$

We now clear out the first row in the same manner. There are p_j and s_j so that $b_{1j} = p_j f_1 + s_j$ and either $s_j = 0$ or $\delta(s_j) < \delta(f_1)$. Then by doing the $n-1$ column operations of taking $-p_j$ times the first column and adding to the j-th column we can farther reduce our matrix to

$$\begin{bmatrix} f_1 & a_{12} - p_2 f_1 & a_{13} - p_3 f_1 & \cdots & a_{1n} - p_n f_1 \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix} = \begin{bmatrix} f_1 & s_2 & s_3 & \cdots & s_n \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix}$$

Exactly as above this the minimality of $\delta(f_1)$ over all elements in matrices equivalent to A implies that $s_j = 0$ for $j = 2, \ldots, n$. So we now have that A is equivalent to the matrix

$$C = \begin{bmatrix} f_1 & 0 & 0 & \cdots & 0 \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix} = \begin{bmatrix} f_1 & 0 & 0 & \cdots & 0 \\ 0 & c_{22} & c_{23} & \cdots & c_{2n} \\ 0 & c_{32} & c_{33} & \cdots & c_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & c_{m2} & c_{m3} & \cdots & c_{mn} \end{bmatrix}$$

If either $m = 1$ or $n = 1$ then C is of one of the two forms

$$[f_1, 0, 0, \ldots, 0], \quad \text{or} \quad \begin{bmatrix} f_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

and we are done.

So assume that $m, n \ge 2$. We claim that every element in this matrix is divisable by f_1. To see this consider any element c_{ij} in the i-th row (where $i, j \geq 2$). Then we can add the i-th row to the first row to get the matrix:

$$\begin{bmatrix} f_1 & c_{i1} & c_{i2} & \cdots & c_{in} \\ 0 & c_{22} & c_{23} & \cdots & c_{2n} \\ 0 & c_{32} & c_{33} & \cdots & c_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & c_{m2} & c_{m3} & \cdots & c_{mn} \end{bmatrix}$$

which is equivalent to A. We use the same trick as above. There are $t_j, \rho_j \in R$ for $2 \leq j \leq n$ so that $c_{ij} = t_j f_1 + \rho_j$ with $\rho_j = 0$ or $\delta(\rho_j) < \delta(f_1)$. Then add $-t_j$ times the first column of to the j-th column to get

$$\begin{bmatrix} f_1 & a_{i2} - t_2 f_1 & a_{i3} - t_3 f_1 & \cdots & a_{in} - t_n f_1 \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix} = \begin{bmatrix} f_1 & \rho_2 & \rho_3 & \cdots & \rho_n \\ 0 & * & * & \cdots & * \\ 0 & * & * & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & * & * & \cdots & * \end{bmatrix}.$$

As this matrix is equivalent to A again the minimality of $\delta(f_1)$ implies that $\delta(\rho_j) = 0$ for $j = 2, \ldots, n$. Therefore $c_{ij} = t_j f_1$ which implies that c_{ij} is divisible by f_1.

As each element of C is divisible by f_1 we can write $c_{ij} = f_1 c'_{ij}$. Factor the f_1 out of the elements of C implies that we can write C in block form as

$$C = \begin{bmatrix} f_1 & 0 \\ 0 & f_1 C' \end{bmatrix} \tag{5.1}$$

where C' is $(m-1) \times (n-1)$.

At last we get to use the induction hypothesis. As $(m-1)+(n-1) < m+n$ the matrix C' is equivalent to a matrix of the form

$$\begin{bmatrix} f'_2 & & & & & \\ & f'_2 & & & & \\ & & \ddots & & & \\ & & & f'_r & & \\ & & & & 0 & \\ & & & & & \ddots \end{bmatrix}$$

where $f'_2, f'_3, \ldots, f'_r$ satisfy $f'_2 \mid f'_3 \mid \cdots \mid f'_r$. (We start at f'_2 rather than f'_1 to make later notation easier.) This means there is a $(m-1)\times(m-1)$ matrix P and an $(n-1) \times (n-1)$ matrices Q so that each of P and Q

are products of elementary matrices and so that

$$PC'Q = \begin{bmatrix} f_2' & & & & & \\ & f_3' & & & & \\ & & \ddots & & & \\ & & & f_r' & & \\ & & & & 0 & \\ & & & & & \ddots \end{bmatrix}$$

This in turn implies

$$Pf_1C'Q = f_1PC'Q = f_1 \begin{bmatrix} f_2' & & & & & \\ & f_3' & & & & \\ & & \ddots & & & \\ & & & f_r' & & \\ & & & & 0 & \\ & & & & & \ddots \end{bmatrix}$$
$$= \begin{bmatrix} f_1f_2' & & & & & \\ & f_1f_3' & & & & \\ & & \ddots & & & \\ & & & f_1f_r' & & \\ & & & & 0 & \\ & & & & & \ddots \end{bmatrix}$$

The block matrices

$$\begin{bmatrix} 1 & 0 \\ 0 & P \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & 0 \\ 0 & Q \end{bmatrix}$$

are of size $m\times m$ and $n\times n$ respectively and are products of elementary matrices. Using our calculation of $Pf_1C'Q$ in equation (5.1) gives

$$\begin{aligned}\begin{bmatrix}1 & 0\\ 0 & P\end{bmatrix} C \begin{bmatrix}1 & 0\\ 0 & Q\end{bmatrix} &= \begin{bmatrix}1 & 0\\ 0 & P\end{bmatrix}\begin{bmatrix}f_1 & 0\\ 0 & f_1C'\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & Q\end{bmatrix}\\ &= \begin{bmatrix}f_1 & 0\\ 0 & Pf_1C'Q\end{bmatrix}\\ &= \begin{bmatrix}f_1 &&&&&\\ & f_1f_2' &&&&\\ && \ddots &&&\\ &&& f_1f_r' &&\\ &&&& 0 &\\ &&&&& \ddots\end{bmatrix}\\ &= \begin{bmatrix}f_1 &&&&&\\ & f_2 &&&&\\ && \ddots &&&\\ &&& f_r &&\\ &&&& 0 &\\ &&&&& \ddots\end{bmatrix}\end{aligned}$$

where $f_2 = f_1f_2', f_3 = f_1f_3', \dots, f_r = f_1f_r'$. As this matrix is equivalent to A to finish the proof it it enough to show that $f_1 \mid f_2 \mid f_3 \mid \cdots \mid f_r$. As $f_2 = f_1f_2'$ it is clear that $f_1 \mid f_2$. If $2 \le j \le r-1$ then we have that $f_j' \mid f_{j+1}'$ so by definition there is a $c_j \in R$ so that $f_{j+1}' = c_jf_j'$. Multiply by f_1 and use $f_j = f_1f_j'$ and $f_{j+1} = f_1f_{j+1}'$ to get $f_{j+1} = f_1f_{j+1}' = f_1c_jf_j' = c_jf_j$. This implies that $f_j \mid f_{j+1}$ and we are done. □

5.3.1. *An application of the existence of the Smith normal form: invertible matrices are products of elementary matrices.* Theorem 5.8 lets us give a very nice characterization of invertible matrices.

Theorem 5.11. *Let $A \in M_{n\times n}(R)$ be a square matrix over an Euclidean domain. Then A is invertible if and only if it is a product of elementary matrices.*

Proof. One direction is clear: Elementary matrices are invertible, so a product of elementary matrices is invertible.

Now assume that A in invertible. Then by Theorem 5.8 A is equivalent to a diagonal matrix

$$D = \operatorname{diag}(f_1, f_2, \dots, f_r, 0, \dots, 0).$$

There is here are matrices P and Q, each a product of elementary matrices, so that

$$A = PDQ.$$

As A, P and Q are invertible their determinants are units (Theorem 4.21) and therefore from $\det(A) = \det(P)\det(D)\det(Q)$ it follows that $\det(D) = \det(A)\det(P)^{-1}\det(Q)^{-1}$ is a unit. But the determinant of a diagonal matrix is the product of its diagonal elements. Thus in the definition of D if $r < n$ there will be a zero on the diagonal and so $\det(D) = 0$, which is not a unit. Thus $r = n$ and so $\det(D) = f_1 f_2 \cdots f_n$. But then $f_1(f_2 \cdots f_n \det(D)^{-1}) = 1$ so that f_1 is a unit with inverse $f_1^{-1} = (f_2 \cdots f_n \det(D)^{-1})$. Likewise each f_k is a unit with inverse $f_k^{-1} = \det(D)^{-1} \prod_{j \neq k} f_j$. But then letting E_k be the diagonal matrix

$$E_k = \operatorname{diag}(1, 1, \ldots, f_k, \ldots, 1)$$

(all ones on the diagonal except at the k-th place where f_k appears) we have that E_k is a an elementary matrix and that D factors as

$$D = E_1 E_2 \cdots E_n.$$

Thus D is a product of elementary matrices. But then $A = PDQ$ is a product of elementary matrices. This completes the proof. $\square$

5.4. **Uniqueness of the Smith normal form.** Recall, Theorem 2.16, that in a Euclidean domain R that any finite set of elements $\{a_1, a_2, \ldots, a_\ell\}$ has a greatest common divisor and that the greatest common divisor of $\{a_1, a_2, \ldots, a_\ell\}$ is the generator of the ideal $\langle a_1, a_2, \ldots, a_\ell \rangle$ (which is a principal ideal by Theorem 2.7). Recall, Definition 4.30, for $A \in M_{m\times n}$ that $\mathcal{I}_k(A)$ is the ideal of R generated by all $k \times k$ sub-determinants of A. Therefore

$$\text{gcd of } k \times k \text{ sub-determinats of } A = \text{genartor of } \mathcal{I}_K(A). \tag{5.2}$$

Theorem 5.12 (Uniqueness of Smith Normal Form)**.** *Let R be a Euclidean domain and let $A \in M_{m\times n}(R)$ and let*

$$S = \begin{bmatrix} f_1 & & & & & \\ & f_2 & & & & \\ & & \ddots & & & \\ & & & f_r & & \\ & & & & 0 & \\ & & & & & \ddots \end{bmatrix}$$

be a Smith normal form of A. Then
(5.3)

$$\textit{gcd of } k\times k \textit{ sub-determinats of } A = \begin{cases} f_1 f_2 \dots f_k, & 1 \le k \le r; \\ 0, & r < k \le k \le \min\{m,n\}. \end{cases}.$$

Therefore the elements $f_1, f_2, \dots, f_r$ are unique up to multiplication by units.

Proof. As S is a Smith Normal form of A there are invertible matrices P and Q such that $PAQ = S$. By Theorem 4.33

$$\mathcal{I}_k(A) = \mathcal{I}_k(PAQ) = \mathcal{I}_k(S).$$

But a direct calculation (left as an exercise) shows

$$\mathcal{I}_k(S) = \begin{cases} \langle f_1 f_2 \dots f_k \rangle, & 1 \le k \le r; \\ \langle 0 \rangle, & r < k \le \min\{m,n\}. \end{cases}.$$

This, along with $\mathcal{I}_k(A) = \mathcal{I}_k(S)$, implies (5.3).

If

$$S' = \begin{bmatrix} f_1' & & & & & \\ & f_2' & & & & \\ & & \ddots & & & \\ & & & f_r' & & \\ & & & & 0 & \\ & & & & & \ddots \end{bmatrix}$$

is another Smith normal form of A then we have

$$\mathcal{I}_k(S') = \mathcal{I}_k(A) = \mathcal{I}_K(S)$$

and therefore, as greatest common divisors are unique up to multiplication by units, there are units $u_1, u_1, \dots, u_r$ of R such that

$$\begin{aligned} f_1' = u_1 f_1, f_1' f_2' = u_2 f_1 f_2, f_1' f_2' f_2' \\ = u_2 f_1 f_2 f_3, , \dots, f_1' f_2', \dots, f_k' = u_k f_a f_2, \dots, f_k. \end{aligned}$$

This implies $f_1' = u_1 f_1$ and

$$f_j = u_{j-1}^{-1} u_j f_j \quad \text{for} \quad 2 \le j \le k.$$

which show $f_1, \dots, f_r$ are unique up to multiplication by units. $\square$

6. Similarity of matrices and linear operators over a field.

6.1. Similarity over R is and equivalence over $R[x]$.

Theorem 6.1. *Let R be a commutative ring and $A, B \in M_{n\times n}(R)$. Then there is an invertible $S \in M_{n\times n}(R)$ so that $B = SAS^{-1}$ if and only if there are invertible $P, Q \in M_{n\times n}(R[x])$ so that $P(xI_n - A) = (xI_n - B)Q$.*

Proof. One direction is easy. If $B = SAS^{-1}$ then $SA = BS$. But then $S(xI_n - A) = xS - SA = xS - BS = (xI_n - B)S$. So letting $P = Q = S$ we have that P and Q are invertible elements of $M_{n\times n}(R[x])$ and $P(xI_n - A) = (xI_n - B)Q$.

Conversely assume that $P, Q \in M_{n\times n}(R[x])$ are invertible and $P(xI_n - A) = (xI_n - B)Q$. Write

$$P = x^m P_m + x^{m-1}P_{m-1} + \cdots + xP_1 + P_0$$

and

$$Q = x^k Q_k + x^{k-1}Q_{k-1} + \cdots + xQ_1 + Q_0$$

where $P_m \neq 0 \neq Q_k$. Then the highest power of x that occurs in $P(xI_n - A)$ is $m+1$ and the highest power of x that occurs in $(xI_n - B)Q$ is $k+1$. As these must be equal we have $k = m$.

The next part of the argument looks very much like the proof of the Cayley-Hamilton Theorem. Writing out both $P(xI_n - A)$ and $(xI_n - B)Q$ in terms of powers of x we find

$$\begin{aligned} P(xI_n - A) &= (x^m P_m + x^{m-1}P_{m-1} + \cdots + xP_1 + P_0)(xI_n - A) \\ &= x^{m+1}P_m + x^m(P_{m-1} - P_mA) + x^{m-1}(P_{m-2} - P_{m-1}A) \\ &\qquad + \cdots + x^2(P_1 - P_2A) + x(P_0 - P_1A) - P_0A \end{aligned}$$

and

$$\begin{aligned} (xI_n - B)Q &= (xI_n - B)(x^m Q_m + x^{m-1}Q_{m-1} + \cdots + xQ_1 + Q_0) \\ &= x^{m+1}Q_m + x^m(Q_{m-1} - BQ_m) + x^{m-1}(Q_{m-2} - BQ_{m-1}) \\ &\qquad + \cdots + x^2(Q_1 - BQ_2) + x(Q_0 - BQ_1) - BQ_0. \end{aligned}$$

Comparing the coefficients of powers of x gives

$$\begin{aligned} P_m &= Q_m \\ P_{m-1} - P_m A &= Q_{m-1} - BQ_m \\ P_{m-2} - P_{m-1}A &= Q_{m-2} - BQ_{m-1} \\ \vdots \quad &= \quad \vdots \\ P_1 - P_2 A &= Q_1 - BQ_2 \\ P_0 - P_1 A &= Q_0 - BQ_1 \\ P_0 A &= BQ_0 \end{aligned}$$

Multiply the first of these on the right by A^{m+1}, the second by A^m, the third by A^{m-1} etc. to get

$$\begin{aligned} P_m A^{m+1} &= Q_m A^{m+1} \\ P_{m-1}A^m - P_m A^{m+1} &= Q_{m-1}A^m - BQ_m A^m \\ P_{m-2}A^{m-1} - P_{m-1}A^m &= Q_{m-2}A^{m-1} - BQ_{m-1}A^{m-1} \\ \vdots \quad &= \quad \vdots \\ P_1 A^2 - P_2 A^3 &= Q_1 A^2 - BQ_2 A^2 \\ P_0 A - P_1 A^2 &= Q_0 A - BQ_1 A \\ P_0 A &= BQ_0 \end{aligned}$$

Adding these equations we see that the terms on the left each term and its negative occurs exactly once to the sum will be zero. Grouping the terms on the right of the sum that contain a B together:

$$\begin{aligned} 0 &= (Q_m A^{m+1} + Q_{m-1}A^m + \cdots + Q_1 A^2 + Q_0 A) \\ &\qquad - B(Q_m A^m + Q_{m-1}A^{m-1} + \cdots + Q_2 A^2 + Q_1 A + Q_0) \\ &= (Q_m A^m + Q_{m-1}A^{m-1} + \cdots + Q_1 A^2 + Q_0 A + P_0 A)A \\ &\qquad - B(Q_m A^m + Q_{m-1}A^{m-1} + \cdots + Q_2 A^2 + Q_1 A + Q_0) \\ &= SA - BS \end{aligned}$$

where

$$S = Q_m A^m + Q_{m-1}A^{m-1} + \cdots + Q_2 A^2 + Q_1 A + Q_0.$$

Thus for this S

$$SA = BS.$$

We now show S is invertible. First, using that $SA = BS$, we find $SA^2 = BSA = B^2 S$, and that generally $SA^k = B^k S$. Let $G = Q^{-1} \in$

$M_{n\times n}(R[x])$ be the inverse of Q. Write

$$G = x^l G_l + x^{l-1}G_{l-1} + \cdots + xG_1 + G_0.$$

Then in the product $GQ = I_n$ the coefficient of x^p is $\sum_{i+j=p} G_i Q_j$ and therefore $GQ = I_n$ implies

$$\sum_{i+j=p} G_i Q_j = \delta_{0p} I_n = \begin{cases} I_n, & p = 0; \\ 0, & p \neq 0. \end{cases}$$

Let

$$T = G_l B^l + G_l B^{l-1} + \cdots + G_1 B + G_0.$$

Then (using at the third step that $B^k S = A^k S$)

$$\begin{aligned}
TS &= (G_l B^l + G_l B^{l-1} + \cdots + G_1 B + G_0)S \\
&= G_l B^l S + G_l B^{l-1} S + \cdots + G_1 BS + G_0 S \\
&= G_l S A^l + G_l S A^{l-1} + \cdots + G_1 SA + G_0 S \\
&= \sum_{k=0}^{m} G_l Q_k A^{l+k} + \sum_{k=0}^{m} G_{l-1} Q_k A^{l-1+k} \\
&\qquad\qquad + \cdots + \sum_{k=0}^{m} G_1 A^{1+k} + \sum_{k=0}^{m} G_0 A^k \\
&= \sum_{p=0}^{m+l} \left(\sum_{i+j=p} G_i Q_j \right) A^p \\
&= \sum_{p=0}^{m+l} \delta_{0p} I_n A^p \\
&= A^0 = I_n.
\end{aligned}$$

Therefore $TS = I_n$. By Theorem 4.23 this implies that $ST = I_n$ and so S is invertible with inverse T. To finish the proof we note that $SA = BS$ now implies $B = SAS^{-1}$. □

7. Modules over Euclidean Domains.

7.1. **Modules over a commutative ring.** While the main theorems here only are true for modules over Euclidean domains, or more generally principle ideal domains, the definitions and basic properties work over more general rings.

Recall the definition of a vector space, V, over a field $\mathbf{F}$: V is an Abelian group with addition $+$ and there is a function $(r, v) \mapsto rv$

(called scalar mutilation) such that for all $r, s \in \mathbf{F}$ and all $v, w \in V$

$$\begin{aligned} 1v &= v, \\ r(sv) &= (rs)v, \\ (r+s)v &= rv + sv, \\ r(v+w) &= rv + rw. \end{aligned}$$

It is not unnatural to replace the scalers $\mathbf{F}$ by a ring R to get a "vector space over the ring R".

Definition 7.1. Let R be a commutative[7] ring. Then V is a ***module over*** $\mathbf{R}$ iff V is V is an Abelian group with addition $+$ and there is a function $(r, v) \mapsto rv$, called ***scalar multiplication***, such that

$$\begin{aligned} 1v &= v, \\ r(sv) &= (rs)v, \\ (r+s)v &= rv + sv, \\ r(v+w) &= rv + rw. \end{aligned}$$

for all $r, s \in R$ and $v, w \in V$. $\square$

We will often just say V is a R***-module*** rather than V is a module over R. Note that when R is a field, then this just becomes the definition of a vector space over R.

As with most new definitions some elementary (and boring) facts need to be verified.

Proposition 7.2. *Let V be an R module. Then for all $r \in R$ and $v \in V$*

$$\begin{aligned} 0v &= 0 \\ r0 &= 0 \\ (-r)v &= -(rv) = r(-v). \end{aligned}$$

Problem 52. Interpret and prove these. By interpret we mean to resolve the ambiguities of notation. For example in $0v = 0$ the first 0 is the zero in the ring R and the second 0 is the zero in V. In $r0 = 0$ both 0's are the zero element in V. In $(-r)v = -(rv) = r(-v)$, $-r$ is the additive inverse of r in R, $-(rv)$ is the additive inverse of (rv) in V and $-v$ is additive inverse of v in V. $\square$

A couple of examples are particularly important to us.

[7]This definition still makes sense, and is still interesting, when R is a noncommunative. But we will restrict ourselves to the commutative case.

Proposition 7.3. *Let V be an Abelian group with group operation $+$. Then V is a $\mathbf{Z}$ module by the rule*

$$(n, v) \mapsto nv$$

with the usual convention that

$$nv = \begin{cases} \underbrace{v + v + \cdots + v}_{n \text{ terms}}, & n > 0; \\ 0, & n = 0; \\ \underbrace{(-v) + (-v) + \cdots + (-v)}_{-n \text{ terms}}, & n < 0. \end{cases}$$

Problem 53. If this is new to you, verify the module axioms. □

The following definitions are the same as in a vector space. The problems following the examples show that in modules corresponding properties are very different.

Definition 7.4. If V is an R module, then $v_1, \ldots, v_n \in V$ is an ***independent*** set iff $r_1, \ldots, r_n \in R$ and $r_1 v_1 + \cdots + r_n v_n = 0$ implies $r_1 = r_2 = \cdots = r_n = 0$. □

Problem 54. If V is a finite Abelian group considered as a $\mathbf{Z}$ module, then V contains no non-empty independent subsets. *Hint:* If n is the order of V, then $nv = 0$ for all $v \in V$. □

Definition 7.5. If V is an R module, then $v_1, \ldots, v_n \in V$ ***span*** V iff each $v \in V$ is a linear combination of $v_1, \ldots, v_n$. That is if $v \in V$ there are $r_1, \ldots, r_n \in R$ such that $v = r_1 v_1 + \cdots + r_n v_n$. □

Definition 7.6. A spanning subset $\{v_1, \ldots, v_n\}$ of an R-module V is ***minimal*** iff the result of removing any element of $\{v_1, \ldots, v_n\}$ no longer spans V.

Recall that in a finite dimensional vector space, V, a minimal spanning set is independent and in fact a basis of V. Thus any two minimal spanning subsets have the same number of elements. This does not generalize to modules:

Problem 55. Show that in the $\mathbf{Z}$ module $\mathbf{Z}_6$ (the cyclic group of order 6) that $\{v_1\} = \{1\}$ and $\{w_1, w_2\} = \{2, 3\}$ are both minimal spanning subsets of V. (More generally if $n_1, \ldots, n_k \in \mathbf{Z}$ are pairwise relatively prime and $n = n_1 n_2 \cdots n_k$, then $\{v_1\} = \{1\}$ and $\{w_1, \ldots, w_k\} = \{n_1, \ldots, n_k\}$ are both minimal spanning subsets of the $\mathbf{Z}$ module $V = \mathbf{Z}_n$.) □

The next few definitions and results follow a pattern that should be familiar by now.

Proposition 7.7. *If V is an R-module and W is a submodule of V, then let $v + W = \{v + w : w \in W\}$ be the* ***coset of*** *v. Let V/W : $0\{v+W : v \in V\}$. Then V/W with the natural operations $((v_1+W)+(v_1+W) = (v+1+v_2)+W$ and $r(v+W) = rv+W)$ makes V/W is an R-module.*

Problem 56. Make this precise and prove it. ☐

Definition 7.8. Let V and W be R modules. Then a $\varphi\colon V \to W$ is a ***homomorphism*** iff $\varphi(v_1 + v_2) = \varphi(v_1) + \varphi(v_2)$ and $\varphi(rv) = r\varphi(v)$ for all $r \in R$ and $v, v_1, v_2 \in V$. The ***kernel*** of φ is $\ker(\varphi) := \{v \in V : \varphi(v) = 0\}$ and the ***image*** of φ is $\mathrm{Image}(\varphi) = \varphi[V] := \{\varphi(v) : v \in V\}$. If φ is bijective, then it is an bi isomorphism. ☐

Proposition 7.9 (The first homomorphism theorem for modules)**.** *Let $\varphi\colon V \to W$ be a homomorphism of R-modules. Then $\ker(\varphi)$ is a submodule of V and $\varphi[V]$ is a submodule of W and $\varphi[V]$ is isomorphic to the quotient module $V/\ker(\varphi)$. Thus if φ is surjective, W is isomorphic to to $V/\ker(\varphi)$.*

Problem 57. Prove this. ☐

If $\mathbf{F}$ is a field, then the most elementary, and maybe also the most natural, vector spaces over $\mathbf{F}$ are the spaces $\mathbf{F}^n$ of length n column vectors over $\mathbf{F}$. In the case of a general commutative ring this is still one of the easiest examples.

Problem 58. Show that R^n with the natural operations, that is for

$$v = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}, \qquad w = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix}$$

and $r \in R$ the operations of sum and scalar multiplication are given by

$$v + w = \begin{bmatrix} v_1 + w_1 \\ v_1 + w_2 \\ \vdots \\ v_n + w_n \end{bmatrix} \qquad \text{and} \qquad rv = \begin{bmatrix} rv_1 \\ rv_2 \\ \vdots \\ rv_n \end{bmatrix},$$

then R^n is an R-module. *Hint:* If this is clear to you skip it. ☐

To give a first difference between modules over a field $\mathbf{F}$ and a ring R, let us consider $R = R^1$ as a R-module in the sense of the last problem.

When $R = \mathbf{F}$ is a field, R, has no submodules over than R and $\{0\}$. This is no longer true over general rings. In fact it is only true when R is a field. This follows from the next proposition and that a ring with no ideals other than $\{0\}$ and itself is a field.

Proposition 7.10. *Let R be a ring. Then, viewing R as a module over itself, a nonempty subset $I \subseteq R$ is a submodule of R if and only if I is an ideal of R.*

Problem 59. Prove this. *Hint:* This is just a definition chase. $\square$

Over a field, $\mathbf{F}$, every finitely generated (that is finite dimensional) vector space is isomorphic to one of the space $\mathbf{F}^n$. Again things are different over general rings. To look at the simplest case, consider a module V with is generated by a single non-zero element $v_0 \in V$. That is every element of V is of the form rv_0 with $r \in R$. In the case $R = \mathbf{Z}$, so that an R module is just an Abelian group, then a R-module is cyclic if and only if it is generated by a single element. The following generalizes this to general rings.

Definition 7.11. A R-module, V, is ***cyclic*** iff there is a $v_0 \in V$ so that $V = \langle v_0 \rangle$ where $\langle v_0 \rangle := \{rv_0 : r \in R\}$ is the submodule of V generated by v_0. Any v_1 with $\langle v_1 \rangle = V$ is a ***generator*** of V. $\square$

Proposition 7.12. *Let V be a cyclic R-module generated by v_0. Then*

$$I := \{r \in R : rv_0 = 0\}$$

is an ideal in R (thus also a submodule of R) and the R-modules V and R/I are isomorphic.

Problem 60. Prove this. *Hint:* Show $\varphi \colon R \to V$ given by $\varphi(r) = rv_0$ is a surjective module homomorphism and use the first isomorphism for modules. $\square$

7.2. **Noetherian Rings, Definition and Basic Results.** While the main structure theorems 5.8 and 5.4 for matrices over a ring, R, only hold over principal ideal domains, many of the preliminary results hold over a more general class of rings.

Definition 7.13. The commutative ring R is ***Noetherian*** iff each ideal of R is finitely generated. Explicitly, if I is an ideal of R, then there is a finite number set $\{a_1, \ldots, a_n\} \subseteq I$ with $I = \langle a_1, \ldots, a_n \rangle$. $\square$

Thus principal ideal domains are the special case of Noetherian ring where each ideal is generated by just one element. Most examples of rings you are likely to encounter in beginning algebra classes are Noetherian. Here is an example showing non-Neotherian rings exist.

Example 7.14. Let $R = \mathbf{Q}[x_1, x_2, \ldots]$ be the ring of polynomials in the countable set of variables $x_1, x_2, \ldots$. Then the ideal $I = \langle x_1, x_2, \ldots \rangle$ is not finitely generated. □

However starting with a Noetherian ring and doing many of the standard constrictions of ring theory produces more Noetherian rings.

Proposition 7.15. *If R is a Noetherian ring and I an ideal of R, then the quotient ring R/I is also Noetherian.*

Problem 61. Prove this. *Hint:* To Simplify notation let $\overline{R} := R/J$. Let $\pi\colon R \to \overline{R}$ be the natural projection, that is $\pi(r) = r + I$. Let $\overline{J}$ be an ideal in $\overline{R}$. Then $J := \pi^{-1}[\overline{J}]$ is an ideal in R and therefore $J = \langle a_1, \ldots, a_n \rangle$ for some $a_1, \ldots, a_n \in J$. Show $\overline{J} = \langle \pi(a_1), \ldots, \pi(a_n) \rangle$. □

Definition 7.16. A commutative ring R satisfies the ***ascending chain condition*** (which will be referred as the ACC) on ideals iff any ascending chain of ideals $I_1 \subseteq I_2 \subset I_3 \subseteq \cdots$ eventually ***stabilizes***. That is there is an n so that $I_n = I_{n+1} = I_{n+2} = \cdots$.

The following is due to Emmy Noether and is the reason this class of rings is called "Noetherian".

Theorem 7.17. *A commutative ring is Noetherian if and only if it satisfies the ACC on ideals.*

Problem 62. Prove this along the following lines.

(1) First assume that R is Noetherian and let $I_1 \subseteq I_2 \subseteq I_3 \subseteq \cdots$ be an ascending sequence of ideals in R. Show the union $I := \bigcup_{k=1}^{\infty} I_k$ is an ideal. As R is Noetherian, $I = \langle a_1, \ldots, a_m \rangle$ for some elements $a_1, \ldots, a_m \in I$. Show there is an n that contains all the elements $a_1, \ldots, a_m$ and that the chain stabilizes at this n (that is $I_n = I_{n+1} = I_{n+2} = \cdots$).
(2) Conversely assume R satisfies the ACC on ideals, and let I be an ideal of R. It is required to show I is finitely generated. If $I = \langle 0 \rangle$, then I is finitely generated. So assume that $I \neq \langle 0 \rangle$. Then choose a sequence of elements as follows. Let $a_1 \in I$ with $a_1 \neq 0$. If $a_1, \ldots, a_k$ have been chosen, then either $\langle a_1, \ldots, a_k \rangle = I$, in which case I is finitely generated and we are done, or if $\langle a_1, \ldots, a_k \rangle \neq I$ then choose $a_{k+1} \in I$ with $a_{k+1} \notin I$. Show for some n the equality $\langle a_1, \ldots, a_n \rangle = I$ must hold, or otherwise the set of ideals $I_n := \langle a_1, \ldots, a_n \rangle$ would be an ascending sequence of ideals that does not stabilize, contradicting the ACC on ideals. □

The following is a foundational result. Note that it has content even when the ring R is a principal ideal domain, in fact even in the case where $R = \mathbf{Z}$, the integers.

Theorem 7.18 (Hilbert Basis Theorem)**.** *If R is a Noetherian ring, so is the polynomial ring $R[x]$.*

Problem 63. Prove this. *Hint:* Let I be an ideal in $R[x]$. If $0 \neq f(x) \in R[x]$ is

$$f(x) = a_k x^k + a_{k-1}x^{k-1} + \cdots + a_1 x + a_0$$

with $a_k \neq 0$, then a_k is the ***lead coefficient*** of $f(x)$ and k is the ***degree*** of $f(x)$. Set

$$I_n := \{0\} \cup \{a \in R : a \text{ is the lead coefficient of an } f(x) \in I \text{ of degree } n.\}$$

(1) Show $a \in I_n$ if and only if there exists a $p(x) \in R[x]$ with $\deg p(x) < n$ (or $p(x) = 0$) such that
$$f(x) = ax^n + p(x) \in I.$$
(2) Use the last fact to show I_n is an ideal of R.
(3) Show that $I_n \subseteq I_{n+1}$. (If $a \in I_n$ there there is a $p(x)$ as in (1) with $ax^n + p(x) \in I$. But I is an ideal so $xf(x) = ax^{n+1} + xp(x) \in I$.)
(4) Use that R is Noetherian to show that there is a N such that $I_N = I_{N+1} = I_{N+2} = \cdots$.
(5) For each n with $0 \leq n \leq N$, use that R is Noetherian to find $a_{n,1}, \dots, a_{n,k_n}$ such that $I_n = \langle a_{n,1}, \dots, a_{n,k_n} \rangle$. For each $a_{n,j}$ there is a $p_{n,j}$ with $p_{n,j}(x) = 0$ or $\deg p_{n,j}(x) < n$ such that
$$f_{n,j}(x) := a_{n,j}x^n + p_{n,j}(x) \in I.$$
(6) Let $f(x) \in I$ have $\deg f(x) = n$ where $0 \leq n \leq N$ and let a be the leading coefficient of $f(x)$. Then $a \in I_n = \langle a_{n,1}, \dots, a_{n,k_n} \rangle$. Thus there are $r_1, \dots, r_{k_n} \in R$ with $a = r_1 a_{n,1} + \cdots + r_{k_n} a_{n,k_n}$. With $f_{n,j}$ as in (5) show
$$f(x) - \sum_{j=1}^{k_n} r_j f_{n,j}(x) \in I$$
and
$$f(x) - \sum_{j=1}^{k_n} r_j f_{n,j}(x) = 0 \quad \text{or} \quad \deg\left(f(x) - \sum_{j=1}^{k_n} r_j f_{n,j}(x) \right) < n.$$

(7) Let

$$\mathcal{B} = \bigcup_{n=0}^{N} \{f_{n,1}(x), \ldots, f_{n,k_n}(x)\}$$

Let $f(x) \in I$ with $\deg f(x) \le N$. Show that there are elements $g_1(x), \ldots, g_m(x) \in \mathcal{B}$ and elements of $r_1, \ldots, r_m \in R$ such that

$$f(x) - \sum_{j=1}^{m} r_j g_j(x) = 0 \quad \text{or} \quad \deg\left(f(x) - \sum_{j=1}^{m} r_j g_j(x)\right) < \deg f(x).$$

(Do not make this hard; it is just a restatement of (6).)

(8) If $n > N$, then $I_n = I_N$. Show that it $\deg f(x) > N$ then there are $g_1(x), \ldots, g_m(x) \in \mathcal{B}$ and $r_1, \ldots, r_m \in R$ with

$$f(x) - \sum_{j=1}^{m} r_j x^{n-N} g_j(x) = 0$$
$$\text{or} \quad \deg\left(f(x) - \sum_{j=1}^{m} r_j x^{n-N} g_j(x)\right) < \deg f(x).$$

(Again this is not hard, the elements $g_1(x), \ldots, g_m(x)$ can be chosen to be a subset of $\{f_{N,1}(x), \ldots, f_{N,k_N}(x)\}$.)

(9) Complete the proof by using induction on $\deg f(x)$, along with (7) and (8), to show if $f(x) \in I$ then $f(x)$ is a linear combination with coefficients in $R[x]$ of elements of $\mathcal{B}$ and thus I is generated by the finite set $\mathcal{B}$. □

While the following is an easy corollary of the last theorem and induction, it is important enough to be recorded as a theorem.

Theorem 7.19. *If R is a field, or more generally a Noetherian ring, then the polynomial ring $R[x_1, \ldots, x_n]$ over R in the n variables $x_1, \ldots, x_n$ is also Noetherian.*

Problem 64. Prove this. *Hint:* Use induction on n. If $R_n := R[x_1, \ldots, x_n]$ is Noetherian, then $R[x_1, \ldots, x_n, x_{n+1}] = R_n[x_{n+1}]$ is a polynomial ring over a Noetherian ring. □

A natural question is if we know something about the number of generators in ideals in a ring R, then can we say something about the number of generators in $R[x]$? For example if R is a principal ideal domain, where every ideal is generated by just one element, then is there an upper bound on the number of generators for ideas in $R[x]$? The next problem shows that the answer is no.

Problem 65. Let $\mathbf{F}$ be a field and $R = \mathbf{F}[x]$. Then R is a Euclidean domain and so R is a principal ideal domain. Then $R[y] = \mathbf{F}[x, y]$. For each $n = 1, 2, \ldots$ let $I_n := \langle x^n, x^{n-1}y, x^{n-2}y^2, \ldots, xy^{n-1}, y^n\rangle$ be the ideal generated by the monomials that are homogenous of degree n. Show that any set of generators of I_n has at least $(n+1)$ elements. *Hint:* I don't know any very intuitive proof of this, but here is an outline of one way to think about it. Both I_n and I_{n+1} are vector spaces over $\mathbf{F}$ with I_{n+1} being a subspace of I_n. Show that the quotient vector space I_n/I_{n+1} has as a basis

$$x^n + I_{n+1},\ x^{n-1}y + I_{n+1},\ \cdots, xy^{n-1} + I_{n+1},\ y^n + I_{n+1}$$

and thus the dimension of over $\mathbf{F}$ of I_n/I_{n+1} is $\dim_{\mathbf{F}}(I_n/I_{n+1}) = n+1$. If $I_n = \langle f_1, \ldots, f_m\rangle$, then show

$$f_1 + I_{n+1},\ f_2 + I_{n+1}, \ldots, f_m + I_{n+1}$$

spans I_n/I_{n+1} and thus $m \geq \dim_{\mathbf{F}}(I_n/I_{n+1}) = n+1$. $\square$

7.3. **Submodules of R^n with R Noetherian.** For a commutative ring R let R^n be the set of column vectors of length n with entires from R. Then, as in Problem 58, R^n is a module over R in a natural way.

Proposition 7.20. *If R is Noetherian, then any submodule, W, of R^n is finitely generated over R. (That is there are $w_1, \ldots, w_N \in W$ such that every element of W is of the form $r_1w_1 + \cdots + r_Nw_N$ with $r_k \in R$.)*

Problem 66. Prove this. *Hint:* Use induction on n. The base case of $n = 1$ is just the definition of R being Noetherian. Assume true for $n-1$, and let W be a submodule of R^n. Let $V \subseteq \mathbf{R}^{n-1}$ be the set of $v = (r_1, \ldots, r_{n-1})$ so that there exists $r_n \in R$ with $(r_1, \ldots, r_{n-1}, r_n) \in W$. Show this is a submodule of R^{n-1}. By the induction hypothesis there are $v_1, \ldots, v_K \in V$ that span V over R. For each v_j there is an $a_j \in R$ with $(v_j, a_j) \in W$. Let $I \subseteq R$ be the set of r so that $(0, \ldots, 0, r) \in W$. Show I an ideal in R. As R is Noetherian $I = \langle x_1, \ldots, x_L\rangle$ for some $x_1, \ldots, x_L \in R$. Show

$$\{(v_j, a_j) : 1 \leq j \leq K\} \cup \{(0, x_i) : 1 \leq i \leq L\}$$

generates W. $\square$

Definition 7.21. Let R be a commutative ring and V a module over R. Then V is ***Noetherian over*** R iff every submodule of V is finitely generated. (That is if W is a submodule of V then there are $x_1, \ldots, x_n \in W$ so that every element of W is of the form $r_1x_1 + \cdots + r_nx_n$ with $r_1, \ldots, r_n \in R$.)

Theorem 7.22. *Every finitely generated module, V, over a Noetherian ring R is a Noetherian module.*

Problem 67. Prove this. *Hint:* Let W be a submodule of V. As V is finitely generated, there are $v_1, \ldots, v_n \in V$ such that every element of V is a linear combination of $v_1, \ldots, v_n$ with coefficients from R. Thus the module homorphism $\varphi\colon R^n \to V$ given by $\varphi(r_1, \ldots, r_n) = r_1v_1 + \cdots + r_nv_n$ is surjective. Then $\varphi^{-1}[W]$ is submodule of R^n and by Proposition 7.20 $\varphi^{-1}[W]$ is finitely generated, say with generators $x_1, \ldots, x_N$. Show $\varphi(x_1), \ldots, \varphi(x_N)$ generate W. □

7.4. **Presentation Matrices of Finitely Generated Modules over Noetherian Rings.** Let R be a Noetherian ring and V a finitely generated module over R. Let $v_1, \ldots, v_m$ be a set of generators for V. Let R^m be the set of column vectors over R of length m. Then there is a module homomorphism $\varphi\colon R^m \to V$ given by

$$\varphi \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{bmatrix} = r_1v_1 + \cdots + r_mv_m.$$

$\ker(\varphi)$ is a submodule or $\mathbf{R}^m$ and therefore by Proposition 7.20 there is a finite set $a_1, \ldots, a_n \in \ker(\varphi)$ that spans $\ker(\varphi)$ over R. The matrix A with columns $a_1, \ldots, a_n$, that is,

$$A = \big[a_1, a_2, \ldots, a_n\big]$$

is a ***presentation matrix*** for the R-module V. From the first homomorphism theorem for modules.

$$V = \operatorname{Image}(\varphi) = R^n/\ker(\varphi) = R^n/\langle a_1, \ldots, a_n\rangle$$

where $\langle a_1, \ldots, a_N\rangle$ is the submodule of V generated by $a_1, \ldots, a_N$ (which is just the set of linear combinations of $a_1, \ldots, a_N$ with coefficients from R.) The matrix $A \in M_{m\times n}(R)$ can be viewed as a module homomorphism $A\colon R^n \to R^m$ by matrix multiplication. Then

$$\langle a_1, \ldots, a_n\rangle = A[R^n] = \{Av : v \in \mathbf{R}^n\}.$$

Problem 68. Prove the last equality. □

To summarize this discussion:

Proposition 7.23. *If R is a Noetherian ring, then every finitely generated R-module, V, is isomorphic to a quotient module*

$$R^m/A[R^n]$$

for some positive integers m and n and some matrix $A \in M_{m\times n}(R)$. □

Problem 69. Here we give some examples of what modules, V, defined by particularly simple, basically diagonal, presentation matrices, A, look like. Prove the following

(1) If $A = \begin{bmatrix} a \end{bmatrix}$ is 1×1, then V is isomorphic to $R/\langle a \rangle$.

(2) More generally if A is a square diagonal matrix

$$A = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \end{bmatrix}$$

then V is isomorphic to

$$R/\langle a_1 \rangle \oplus R/\langle a_2 \rangle \oplus \cdots \oplus R/\langle a_n \rangle.$$

(3) If A is

$$A = \begin{bmatrix} a_1 & & & & 0 & \cdots & 0 \\ & a_2 & & & 0 & \cdots & 0 \\ & & \ddots & & 0 & \cdots & 0 \\ & & & a_n & 0 & \cdots & 0 \end{bmatrix}$$

Some columns of consisting of zeros added to last example.

then V is still isomorphic to

$$R/\langle a_1 \rangle \oplus R/\langle a_2 \rangle \oplus \cdots \oplus R/\langle a_n \rangle.$$

(4) But if the presentation matrix of part (2) has rows of zeros added:

$$A = \begin{bmatrix} a_1 & & & \\ & a_2 & & \\ & & \ddots & \\ & & & a_n \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & 0 \end{bmatrix}$$

$n+m$ rows, the last n of which are all zeros.

then

$$V = R/\langle a_1 \rangle \oplus R/\langle a_2 \rangle \oplus \cdots \oplus R/\langle a_n \rangle \oplus R^m$$

□

Proposition 7.23 is a nice result in that it shows finite generated modules over a Noetherian ring, R, have very concrete realizations, as quotient modules of some R^m by a finitely generated submodule. But it does little to settle when two such modules are isomorphic as just about nothing in the construction is unique. For example if $R = \mathbf{Z}$, the integers, then the following presentation matrices all determine the cyclic group of order 6.

$$\begin{bmatrix}6\end{bmatrix}, \quad \begin{bmatrix}2 & 0\\ 0 & 3\end{bmatrix}, \quad \begin{bmatrix}2 & 0 & 2\\ 0 & 3 & 6\end{bmatrix}, \quad \begin{bmatrix}6 & -6 & 18\\ 2 & -3 & 9\end{bmatrix}, \quad \begin{bmatrix}6 & -6 & 18\\ 2 & -3 & 9\\ 0 & 0 & 1\end{bmatrix}.$$

So it is interesting to give conditions which imply two presentation matrices define isomorphic R-modules. First a couple of almost trivial results.

Proposition 7.24. *If A is a $m \times n$ presentation matrix of a module over the ring R, then the $(m+1)\times(n+1)$ matrix*

$$B = \begin{bmatrix} A & 0\\ 0 & 1\end{bmatrix} \qquad \begin{array}{l}\textit{B is A bordered to the right and}\\ \textit{below by 0's and with a 1 in the}\\ (m+1)\times(n+1) \textit{ position.}\end{array}$$

is a presentation matrix of a module isomorphic to the module with presentation matrix A. (More briefly A and B are presentation matrices for isomorphic R-module.)

Problem 70. Prove this. □

Proposition 7.25. *If A is a $m \times n$ presentation matrix of a module over the ring R, and B is the $m \times (n+1)$ matrix*

$$B = \begin{bmatrix} A & 0\end{bmatrix} \qquad \textit{B is A with a column of 0's added.}$$

Then A and B define isomorphic R-module.

Problem 71. Prove this. □

The next result is less trivial.

Theorem 7.26. *Let R be a commutative ring, and let $A \in M_{m\times n}(R)$, $P \in M_{m\times m}(R)$, and $Q \in M_{n\times n}(R)$. Assume that P and Q have inverses in $M_{m\times m}(R)$ and $M_{n\times n}(R)$ respectively. Then A and*

$$B = PAQ$$

are presentation matrices for isomorphic R-modules.

Problem 72. Prove this. *Hint:* It is required to show $R^m/A[R^n]$ and $R^m/B[R^n]$ are isomorphic. Use that Q is invertible to show $Q[R^n] = R^n$ and therefore $B[R^n] = PAQ[R^n] = PA[R^n]$. So we only need show $R^m/A[R^n]$ and $R^m/PA[R^n]$ are isomorphic. Define a map $\varphi\colon R^m/A[R^n] \to R^m/PA[R^n]$ by $\varphi(v + A[R^n]) = Pv + PA[R^n]$. Show this is well defined and a module homomorphism. Show $\psi\colon R^m/PA[R^n] \to R^n/A[R^n]$ given by $\psi(v + PA[R^n]) = P^{-1}v + A[R^n]$ is an inverse to φ. □

Problem 73. As a, hopefully enlightening, variant on a part of the proof of the last theorem, let V be a module over the commutative ring R and let W be a submodule of V. If $\alpha\colon V \to V$ is a R-module automorphism of V, then $\alpha[W] := \{\alpha(w) : w \in W\}$ is a submodule of V and the quotient modules V/W and $V/\alpha[W]$ are isomorphic. □

In the case that R is a Euclidean domain, then by Theorem 5.8, p. 74 for any matrix $A \in M_{m\times n}(R)$ there are invertible $P \in M_{m\times m}(R)$ and $Q \in M_{n\times n}(R)$ such that PAQ is in Smith normal form, that is

$$PMQ = \begin{bmatrix} f_1 & & & & & \\ & f_2 & & & & \\ & & \ddots & & & \\ & & & f_r & & \\ & & & & 0 & \\ & & & & & \ddots \end{bmatrix} \qquad \text{This is } m \times n \text{ and } f_1 \mid f_2 \mid \cdots \mid f_{r-1} \mid f_r. \tag{7.1}$$

Proposition 7.27. *A module with a presentation matrix as in* (7.1) *is isomorphic to*

$$R/\langle f_1\rangle \oplus R/\langle f_2\rangle \oplus \cdots \oplus R/\langle f_r\rangle \oplus R^{n-r}.$$

Problem 74. Prove this. *Hint:* Theorem 7.26 and Problem 69. □

References

1. Irving Kaplansky, *Elementary divisors and modules*, Trans. Amer. Math. Soc. **66** (1949), 464–491. MR 0031470 (11,155b)

Index