THE GAUSSIAN INTEGERS I: THE FUNDAMENTAL THEOREM

PETE L. CLARK

1. Introducing the Gaussian Integers

Earlier we justified our proof of the fundamental theorem of arithmetic by arguing that the result was not obvious, since its analogue in the ring $\mathbb{E}$ of even integers did not hold. Here we will explore a far more exciting explanation: essentially the same proof can be used to show that certain other "integer-like rings" have the unique factorization property.

The unique factorization property can be formulated in any integral domain: that is, one has the notion of a **unique factorization domain** (or UFD). Rather than begin with this general (purely algebraic) notion here, we choose to keep a more properly number-theoretic focus and start with a particular ring which will be very important for us in the rest of the course.

Namely, we consider the ring $\mathbb{Z}[i]$ of Gaussian integers. Here $i = \sqrt{-1}$, and the easiest way to view $\mathbb{Z}[i]$ is as the subring of complex numbers $\{a+bi \mid a, b \in \mathbb{Z}\}$, i.e., those complex numbers whose x and y coordinates are integers. In other words, the Gaussian integers are just the lattice points in the plane, added and multiplied as complex numbers.

Let us check that this *is* a subring of $\mathbb{C}$. It contains $0 = 0 + 0i$ and $1 = 1 + 0i$. Moreover it is certainly closed under addition: $(a+bi)+(c+di) = (a+c)+(b+d)i$, and if all of a, b, c, d are integers, so are $a + c$ and $b + d$. Finally we have to check closure under multiplication:

$$(a + bi)(c + di) = ac + bci + adi + bdi^2 = (ac - bd) + (ad + bc)i,$$

where in the second equality we have used, of course, that $i^2 = -1$.

Remark: That the integer lattice is closed under addition – i.e., that it forms a subgroup of the additive group of the complex numbers – seems geometrically obvious, since any element can be written as the sum of a horizontal element $a \cdot 1$ and a vertical element $b \cdot i$. That it is closed under multiplication also has a geometric meaning: we can write complex numbers in polar form as $re^{i\theta}$, where $r = \sqrt{a^2 + b^2}$ and $\theta = \arctan(b/a)$, and then

$$r_1 e^{i\theta_1} \cdot r_2 e^{i\theta_2} = r_1 r_2 e^{i(\theta_1+\theta_2)}.$$

In other words, to multiply two nonzero complex numbers, we multiply their lengths and add their angles. In particular, multiplication by i is a 90 degree counterclockwise rotation. We can see that this symmetry preserves the standard integer lattice and also that this is a very special property of the lattice: given any two linearly independent vectors $v_1,\ v_2$ in the plane we can consider the lattice that they generate: $\{av_1 + bv_2 \mid a, b \in \mathbb{Z}\}$ (a picture would be quite helpful here!). Most of these lattices will not have this extra rotational symmetry, or indeed any other symmetry other than reflection through the origin.

Note that $\mathbb{Z}[i]$, being a subring of a field, is an integral domain.

Exercise: Let D be an integer which is not a square.
Define the quotient ring $R_D = \mathbb{Z}[X]/(X^2 - D)$. Show that elements of Q_D can be written in the form $a + bX$, where $a, b \in \mathbb{Z}$ and $X^2 = D$. For $D = -1$ we get the Gaussian integers.

What would it even mean for $\mathbb{Z}[i]$ to have unique factorization? We would like to have a notion of "primes" in this ring. The most naive definition, of a nonzero element which is only divisible by 1 and itself, doesn't quite work, since every $a+bi$ is divisible by $\pm i$:

$$a + bi = i(b - ai) = -i(-b + ai).$$

Of course any integer is also divisible by -1, but this is more easily ignored in $\mathbb{Z}$ just by restricting to positive integers. One might think to define "positive" Gaussian integers as those $a+bi$ with $a, b > 0$ – i.e., those in the first quadrant of $\mathbb{C}$ – but this notion of positivity isn't stable under multiplication: $(1 + i)^2 = 2i,\ (i + 1)^4 = -4$ and so on. In fact there is no ordering on $\mathbb{Z}[i]$ compatible with its ring structure in a sense which we make precise in the exercises.

It turns out to be most fruitful just to accept this: to regard the four integers $\pm a + bi,\ \pm i(a + bi)$ as being in some sense equivalent, just as we regard $\pm n$ as being in some sense equivalent in $\mathbb{Z}$. But what sense is this?

Well, for given nonzero n, $-n$ is the unique element such that $n|(-n)$ and $-n|n$. Equivalently, $n/(-n) = -1$ is a *unit* in $\mathbb{Z}$ – its inverse is also in $\mathbb{Z}$, and ± 1 are the only units in $\mathbb{Z}$. We call two elements x and y of an integral domain R **associates** if $x = uy$ for some unit $u \in R^\times$.

Exercise: a) Being associates is an equivalence relation.
b) x and y are associates iff they generate the same principal ideal: $(x) = (y)$.

So what is happening is that $\mathbb{Z}[i]$ has at least four units: $\pm 1, \pm i$. In fact these are the only units. To see this, observe that the inverse of $a + bi$ in $\mathbb{C}$ is

$$(a + bi)^{-1}\frac{a - bi}{a^2 + b^2} = \frac{a}{a^2 + b^2} + \frac{b}{a^2 + b^2}i.$$

So if $(a + bi)^{-1}$ is in $\mathbb{Z}[i]$ then $\frac{a}{a^2+b^2}$ and $\frac{b}{a^2+b^2}$ are both integers. But, since $|a^2 + b^2| \geq \max\{|a|, |b|\}$ with equality iff $ab = 0$, one easily sees that this can only happen for $(a, b) = (\pm 1, 0)$ or $(0, \pm 1)$. So we have found all the units in $\mathbb{Z}[i]$.

Now we can make the key definition: a nonzero, nonunit element x of $\mathbb{Z}[i]$ is *irreducible* if whenever it factors as $x = yz$ with $y,\ z \in \mathbb{Z}[i]$, one of y and z is a unit (and in fact only one, since otherwise x would be a unit).

Two key questions to ask are:

Q1: Does every (nonzero nonunit) $x \in \mathbb{Z}[i]$ admit a factorization into irreducibles?

Q2: If so, is the irreducible factorization essentially unique, in the sense that if

$$x = p_1 \cdots p_r = q_1 \cdots q_s$$

are two irreducible factorizations, then there exists a one-to-one correspondence $\sigma : \{1, \ldots r\} \to \{1, \ldots, s\}$ such that for all $1 \leq i \leq r$, p_i and $\sigma(p_i)$ are associates?

The first question ought not to be so hard: intuitively, in a nontrivial factorization $x = yz$, the two factors get "smaller." Indeed, the length of x is precisely the length of y times the length of z, and the nonzero nonunits are precisely the elements $a + bi$ of $\mathbb{Z}[i]$ whose length $\sqrt{a^2 + b^2}$ is greater than 1. Thus, if $x = yz$ is a nontrivial factorization, then indeed y and z have smaller lengths than x.

This does not exactly tell us that the factorization process must terminate, because the length is a real number, and there is no smallest real number, nor a smallest real number which is greater than 1. But the length $\sqrt{a^2 + b^2}$ is not just any old real number: it has quite a special form. After ruminating a bit on this, inspiration will strike sooner or later: for arithmetic purposes, it is better to consider the length squared, say

$$N(a + bi) = a^2 + b^2,$$

which we call the **norm** of $a+bi$. We can immediately hail this as a wonderful thing: for instance, writing $\overline{z} = a - bi$, we have $z\overline{z} = N(z)$, which explains for instance our formula for the inverse: $z^{-1} = \frac{\overline{z}}{N(z)}$. Moreover, since lengths are multiplicative, so are lengths squared: for $x, y \in \mathbb{C}$

$$N(xy) = N(x)N(y).$$

Moreover, $N(\mathbb{Z}[i]) \subset \mathbb{N}$, and for $z \in \mathbb{Z}[i]$,

$$N(z) = 0 \iff z = 0,$$

$$N(z) = 1 \iff z \in \mathbb{Z}[i]^\times = \{\pm 1,\ \pm i\}.$$

Therefore we have the following wonderfully convenient criterion for testing the nontriviality of a factorization: $z = xy$ is a nontrivial factorization in $\mathbb{Z}[i]$ iff $N(z) = N(x)N(y)$ is a nontrivial factorization in $\mathbb{Z}$ iff $N(x), N(y) > 1$. This gives the "clear progress" we want to have to be sure that the factorization process will eventually terminate: at each stage the norms get smaller, and the smallest possible value is $N(x) = 2$.

We can now concentrate on the more subtle and interesting question of the uniqueness of the factorization into irreducible elements, up to order and associates. The idea is to try to imitate the proof in $\mathbb{Z}$. For that proof, the first step was to show that every ideal of $\mathbb{Z}$ was principal, and that was probably the easiest step: we just took (except for the zero ideal!) the element of least positive absolute value. At this point the rest of the proof must look more worrisome, so let us first nail down that this is absolutely not the case.

2. Unique Factorization in a PID

First make the following definition: a **principal ideal domain** (PID) is – as it should be – a commutative ring without zero divisors in which every ideal is principal, i.e., of the form $(x) = \{rx \mid r \in R\}$ for some element of R.[1] In fact we have essentially already proved the following result:

Theorem 1. *Let R be a PID, and let x be a nonzero nonunit of R. Then any two irreducible factorizations*

$$x = p_1 \cdots p_r = q_1 \cdots q_s$$

of x are equivalent in the above sense: there is a bijective correspondence from the p_i's to the q_j's in which corresponding elements are associates.

Proof: Well, we've seen it all before. The first step is to realize that the ideal $I_{a,b}$ generated by two elements of R, being principal, must be generated by a gcd of a and b: i.e., $I_{a,b} = (d)$ is a common divisor of a and b with the property that $e|a \& e|b \implies e|d$. (Note that there is no ambiguity in the sense of "greatest" in this level of generality: we don't necessarily have an additive ordering $\leq$ on R, and indeed we do not on the Gaussian integers.) Just to recall the proof: say $d = xa + yb$. Then since a and b are in the ideal, $a = Ad$ and $b = Bd$, say, so $d|a$ and $d|b$. And if $e|a$ and $e|b$, then $e|xa + yb$, so $e|d$: done!

Now we prove Euclid's Lemma in R. It says that if p is an irreducible element of R and $p|xy$, then $p|x$ or $p|y$. Well, suppose that p does not divide x. Then the gcd d of p and x is a proper divisor of p, i.e., we have that the ideal (d) strictly contains (p), but a little reflection reveals that this means that d is the unit ideal $R = (1)$: if its generator were not a unit, that would lead to a nontrivial factorization of p. So there exist r, s in R such that $rp + sx = 1$, so

$$rpy + sxy = y$$

so that p divides $rpy + sxy$, so $p|y$: same proof!

Finally, the deduction of the uniqueness of the factorization is again exactly the same, with the small proviso that since we are only concluding that the primes p_i and q_j are associates, when we cancel p_i and q_j we have to correct by a unit.

Note that we have not quite shown that every PID is a Unique Factorization Domain (UFD), since if R is an arbitrary PID we have not shown that every nonzero nonunit x has at least one factorization into irreducibles (note however that we did show this for $\mathbb{Z}[i]$ in the last section, so we are worrying only about "the general case" now). This is true, and not too hard to prove; however, is a little bit abstract, so is relegated to Algebra Handout 3.

3. The Gaussian Integers Form a PID

Let us now establish the key fact: namely that every ideal in $\mathbb{Z}[i]$ is principal. As above, since we have already explained that the properties of the norm map N imply that every Gaussian integer can be factored into irreducibles, this will imply that irreducibles are *prime* and that every nonzero nonunit Gaussian integer can

[1] See algebra handout A2 for more information on principal rings and principal ideal domains.

be factored uniquely into irreducibles.

Recall that in the proof of the fundamental theorem in $\mathbb{Z}$ we observed that every ideal of $\mathbb{Z}$ is generated by an element which was "smallest" in a reasonable sense. In $\mathbb{Z}$, this sense was just smallest in absolute value. So it is very natural to try to show that every ideal in $\mathbb{Z}[i]$ is generated by any of its elements of minimal norm (notice that if $N(z) = N(w)$, then z and w are associates). For this we establish the following important result:

Theorem 2. *(Division Theorem in $\mathbb{Z}[i]$) Let $\alpha, \beta \in \mathbb{Z}[i]$, with $\beta \neq 0$. Then there exist $\gamma,\ \delta \in \mathbb{Z}[i]$ such that $\alpha = \gamma\beta + \delta$ with $N(\delta) < N(\beta)$.*

Proof: Recall that N is defined and multiplicative – $N(zw) = N(z)N(w)$ – on all complex numbers: in particular, we get a map $N : \mathbb{Q}[i] \rightarrow \mathbb{Q}^{\geq 0}$, so the conclusion can be usefully rephrased as: find Gaussian integers γ and δ such that

$$\frac{\alpha}{\beta} - \gamma = \frac{\delta}{\beta}$$

with $N(\frac{\delta}{\beta}) < 1$. This makes it clear that what we are trying to do is approximate an element of $\mathbb{Q}[i]$ – an $a + bi$ with $a, b \in \mathbb{Q}$ – by a Gaussian integer, in a sufficiently efficient way. Well, there's pretty much only one thing to do: given any rational number $x \in \mathbb{Q}$, we can find an integer X such that $|x - X| \leq \frac{1}{2}$ and this is the best we can do in general (in particular, if X is equal to $\frac{1}{2}$ plus an integer, then there are two nearest integers and they have distance $\frac{1}{2}$; otherwise there is a unique nearest integer and the inequality can be made strict but not otherwise improved). So for $\alpha/\beta = a + bi$, choose nearest integers A to a and B to b and put $\gamma = A + Bi$. Then

$$N(\frac{\alpha}{\beta} - \gamma) = N((a - A) + (b - B)i) = (a - A)^2 + (b - b)^2 \leq (\frac{1}{2})^2 + (\frac{1}{2})^2 = \frac{1}{2} < 1.$$

This proves the theorem.

Remark: In some sense it was pure luck that the norm of the difference came out less than 1: it is not the sort of "estimate" that can be improved with further work: either it comes out or it doesn't. The significance of this will become clear later.

Finally, we prove that $\mathbb{Z}[i]$ is a PID using the argument alluded to above: let I be a nonzero ideal of $\mathbb{Z}[i]$. Then the set $\{N(z) \mid z \in I \setminus 0\}$ is a nonempty set of $\mathbb{Z}^{+}$ so has a least element n; let z_0 be one (of the four) elements of I with $N(z_0) = n$. We claim that $I = (z_0)$. Since $z_0 \in I$ and I is an ideal, we have $(z_0) \subset I$. Conversely, let $z \in I$, and apply the division theorem with $\alpha = z$, $\beta = z_0$:

$$z = \gamma z_0 + \delta,$$

where $N(\delta) < N(z_0) = n$, so δ is an element of I of norm smaller than n, meaning $\delta = 0$ and $z \in (z_0)$.

4. A failure of the fundamental theorem

Consider now the subring

$$\mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} \mid a, b \in \mathbb{Z}\}$$

of the complex numbers. It is again an integral domain, and again the restriction of $|z|^2$ gives a pleasant looking norm function: we have

$$N(a+b\sqrt{-5}) = (a+b\sqrt{-5})(a-b\sqrt{-5}) = a^2+5b^2.$$

In particular the units of this ring are now the solutions to $a^2+5b^2 = 1$, i.e., $a = \pm 1, b = 0$, so just ± 1.

However, consider:

$$6 = 2 \cdot 3 = N(1+\sqrt{-5}) = (1+\sqrt{-5})(1-\sqrt{-5}).$$

I claim that $2, 3$ and $1 \pm \sqrt{-5}$ are all irreducible elements. Indeed since $N(2) = 4$, a nontrivial factorization of 2 would involve elements of norm 2, which would involve solutions to $a^2+5b^2 = 2$, of which there are none. Similarly, since $N(3) = 9$, a nontrivial factorization of 3 would involve elements of norm 3, hence solutions to $a^2+5b^2 = 3$: still none. Finally, a nontrivial factorization of $1 \pm \sqrt{-5}$ would involve elements either of norm 2 or norm 3! Finally, since the associates of an element $z \in \mathbb{Z}[\sqrt{-5}]$ are just $\pm z$, no two of 2, 3, $1+\sqrt{-5}$ and $1-\sqrt{-5}$ are associates, so we have indeed found two essentially different factorizations of 6 in $\mathbb{Z}[\sqrt{-5}]$.