3 Finite fields and integer arithmetic

In order to perform explicit computations with elliptic curves over finite fields, we first need to understand how to compute in finite fields. In many of the applications we will consider, the finite fields involved will be quite large, so it is important to understand the computational complexity of finite field operations. This is a huge topic, one to which an entire course could be devoted, but we will spend just two lectures on this topic, with the goal of understanding the most commonly used algorithms and analyzing their asymptotic complexity. This will force us to omit many details.

Our first step is to fix an explicit representation of finite field elements. This might seem like a technical detail, but it is actually quite crucial; questions of computational complexity are meaningless otherwise.

Example 3.1. By Theorem 3.11 below, the multiplicative group of a finite field $\mathbb{F}_q$ is cyclic. One way to represent the nonzero elements of a finite field is as explicit powers of a fixed generator, in which case it is enough to know the exponent, an integer in $[0, q-1]$. With this representation multiplication and division are easy, solving the discrete logarithm problem is trivial, but addition is costly (not even polynomial time). We will instead choose a representation that makes addition (and subtraction) very easy, multiplication slightly harder but still easy, division slightly harder than multiplication but still easy (all these operations take quasi-linear time). But solving the discrete logarithm problem will be hard (no polynomial-time algorithm is known).

For they sake of brevity, we will focus primarily on finite fields of large characteristic, and prime fields in particular, although the algorithms we describe will work in any finite field of odd characteristic (and most will also work in characteristic 2). Fields of characteristic 2 are quite important in many applications (coding theory in particular), and there are many specialized algorithms that are optimized for such fields, but we will not address them here.[1]

3.1 Finite fields

We begin with a quick review of some basic facts about finite fields, all of which are straightforward but necessary for us to establish a choice of representation; we will also need them when we discuss algorithms for factoring polynomials over finite fields in the next lecture. Those already familiar with this material should feel free to skim this section.

Definition 3.2. For each prime p we define $\mathbb{F}_p$ to be the quotient ring $\mathbb{Z}/p\mathbb{Z}$.

Theorem 3.3. *The ring $\mathbb{F}_p$ is a field, and every field of characteristic p contains a canonical subfield isomorphic to $\mathbb{F}_p$. In particular, all fields of cardinality p are isomorphic.*

Proof. To show that the ring $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ is a field we just need to show that every nonzero element is invertible. If $[a] := a + p\mathbb{Z}$ is a nontrivial coset in $\mathbb{Z}/p\mathbb{Z}$ then a and p are coprime and $(a, p) = (1)$ is the unit ideal. Therefore $ua + vp = 1$ for some $u, v \in \mathbb{Z}$ with $ua \equiv 1 \bmod p$, so $[u][a] = [1]$ in $\mathbb{Z}/p\mathbb{Z}$ and $[a]$ is invertible. To justify the second claim, note that in any field of characteristic p the subring generated by 1 is isomorphic to $\mathbb{Z}/p\mathbb{Z} = \mathbb{F}_p$, and this subring is clearly unique (any other must also contain 1), hence canonical. □

[1]With the recent breakthrough [1] computing discrete logarithms in finite fields of small characteristic in quasi-polynomial time, there is much less enthusiasm for using such fields in cryptographic applications.

The most common way to represent $\mathbb{F}_p$ for computational purposes is to pick a set of unique coset representatives for $\mathbb{Z}/p\mathbb{Z}$, such as the integers in the interval $[0, p-1]$.

Definition 3.4. For each prime power $q = p^n$ we define $\mathbb{F}_q = \mathbb{F}_{p^n}$ to be the field extension of $\mathbb{F}_p$ generated by adjoining all roots of $x^q - x$ to $\mathbb{F}_p$ (the splitting field of $x^q - x$ over $\mathbb{F}_p$).

Theorem 3.5. *Let $q = p^n$ be a prime power. The field $\mathbb{F}_q$ has cardinality q and every field of cardinality q is (non-canonically) isomorphic to $\mathbb{F}_q$.*

Proof. The map $x \mapsto x^q = x^{p^n}$ is an automorphism σ_q of $\mathbb{F}_q$, since in characteristic p we have

$$(a+b)^{p^n} = a^{p^n} + b^{p^n} \qquad \text{and} \qquad (ab)^{p^n} = a^{p^n} b^{p^n},$$

where the first identity follows from the binomial theorem: $\binom{p^n}{r} \equiv 0 \bmod p$ for $0 < r < p^n$. Let $k := \mathbb{F}_q^{\sigma_q}$ be the subfield of $\mathbb{F}_q$ fixed by σ_q. We have $\mathbb{F}_p \subseteq k$, since

$$(1 + \cdots + 1)^q = 1^q + \cdots + 1^q = 1 + \cdots + 1,$$

and it follows that $\mathbb{F}_q = \mathbb{F}_q^{\sigma_q}$, since σ_q fixes $\mathbb{F}_p$ and every root of $x^q - x$. The polynomial $x^q - x$ has no roots in common with its derivative $(x^q - x)' = qx^{q-1} - 1 = -1$, so it has q distinct roots, which are precisely the elements of $\mathbb{F}_q$ (they lie in $\mathbb{F}_q$ be definition, and every element of $\mathbb{F}_q = \mathbb{F}_q^{\sigma_q}$ is fixed by σ_q and therefore a root of $x^q - x$).

Now let k be a field of cardinality $q = p^n$. Then k must have characteristic p, since the set $\{1, 1+1, \ldots\}$ is a subgroup of the additive group of k, so the characteristic divides $\#k = p^n$, and in a finite ring with no zero divisors the characteristic must be prime. By Theorem 3.3, the field k contains $\mathbb{F}_p$. The order of each $\alpha \in k^\times$ divides $\#k^\times = q - 1$; thus $\alpha^{q-1} = 1$ for all $\alpha \in k^\times$, so every $\alpha \in k$, including $\alpha = 0$, is a root of $x^q - x$. It follows that k is isomorphic to a subgroup of $\mathbb{F}_q$, and $\#k = \#\mathbb{F}_q$, so $k \simeq \mathbb{F}_q$ (this isomorphism is not canonical because when q is not prime there are many ways to embed k in $\mathbb{F}_q$). □

Remark 3.6. Now that we know all finite fields of cardinality q are isomorphic, we will feel free to refer to any and all of them as *the* finite field $\mathbb{F}_q$, with the understanding that there are many ways to represent $\mathbb{F}_q$ and we will need to choose one of them.

Theorem 3.7. *The finite field $\mathbb{F}_{p^m}$ is a subfield of $\mathbb{F}_{p^n}$ if and only if m divides n.*

Proof. If $\mathbb{F}_{p^m} \subseteq \mathbb{F}_{p^n}$ then $\mathbb{F}_{p^n}$ is an $\mathbb{F}_{p^m}$-vector space of (integral) dimension n/m, so $m|n$. If $m|n$ then $p^n - p^m = (p^m - 1)(p^{n-m} + p^{n-2m} + \cdots + p^{2m} + p^m)$ is divisible by $p^m - 1$ and

$$x^{p^n} - x = (x^{p^m} - x)(1 + x^{p^m - 1} + x^{2(p^m-1)} + \cdots + x^{p^n - p^m})$$

is divisible by $x^{p^m} - x$. Thus every root of $x^{p^m} - x$ is also a root of $x^{p^n} - x$, so $\mathbb{F}_{p^m} \subseteq \mathbb{F}_{p^n}$. □

Theorem 3.8. *For any irreducible $f \in \mathbb{F}_p[x]$ of degree $n > 0$ we have $\mathbb{F}_p[x]/(f) \simeq \mathbb{F}_{p^n}$.*

Proof. The ring $k := \mathbb{F}_p[x]/(f)$ is an $\mathbb{F}_p$-vector space with basis $1, \ldots, x^{n-1}$ and therefore has dimension n and cardinality p^n. The ring $\mathbb{F}_p[x]$ is a principal ideal domain and f is irreducible and not a unit, so (f) is a maximal ideal and $\mathbb{F}_p[x]/(f)$ is a field with p^n elements, hence isomorphic to $\mathbb{F}_{p^n}$ by Theorem 3.5. □

Theorem 3.8 allows us to explicitly represent $\mathbb{F}_{p^n}$ as $\mathbb{F}_p[x]/(f)$ using any irreducible polynomial $f \in \mathbb{F}_p[x]$ of degree n, and it does not matter which f we pick; by Theorem 3.5 we always get the same field (up to isomorphism). We also note the following corollary.

Corollary 3.9. *Every irreducible polynomial $f \in \mathbb{F}_p[x]$ of degree n splits completely in $\mathbb{F}_{p^n}$.*

Proof. We have $\mathbb{F}_p[x]/(f) \simeq \mathbb{F}_{p^n}$, so every root of f must be a root of $x^{p^n} - x$, hence an element of $\mathbb{F}_{p^n}$. □

Remark 3.10. This corollary implies that $x^{p^n} - x$ is the product over the divisors d of n of all monic irreducible polynomials of degree d in $\mathbb{F}_p[x]$. This can be used to derive explicit formulas for the number of irreducible polynomials of degree d in $\mathbb{F}_p[x]$ using Möbius inversion. It also implies that, even though we defined $\mathbb{F}_{p^n}$ as the splitting field of $x^{p^n} - x$, it is also the splitting field of every irreducible polynomial of degree n.

Theorem 3.11. *Every finite subgroup of the multiplicative group of a field is cyclic.*

Proof. Let k be a field, let G be a subgroup of $k^\times$ of order n, and let m be the exponent of G (the least common multiple of the orders of its elements), which necessarily divides n. Every element of G is a root of $x^m - 1$, which has at most m roots, so $m = n$. Every finite abelian group contains an element of order equal to its exponent, so G contains an element of order $m = n = \#G$ and is therefore cyclic. □

Corollary 3.12. *The multiplicative group of a finite field is cyclic.*

If α is a generator for the multiplicative group $\mathbb{F}_q^\times$, then it generates $\mathbb{F}_q$ as an extension of $\mathbb{F}_p$, that is, $\mathbb{F}_q = \mathbb{F}_p(\alpha)$, and we have $\mathbb{F}_q \simeq \mathbb{F}_p[x]/(f)$, where $f \in \mathbb{F}_p[x]$ is the minimal polynomial of α, but the converse need not hold. This motivates the following definition.

Definition 3.13. A monic irreducible polynomial $f \in \mathbb{F}_p[x]$ whose roots generate the multiplicative group of the finite field $\mathbb{F}_p[x]/(f)$ is called a *primitive polynomial*.

Theorem 3.14. *For every prime p and positive integer n there exist primitive polynomials of degree n in $\mathbb{F}_p[x]$. Indeed, the number of such polynomials is $\phi(p^n - 1)/n$.*

Here $\phi(m)$ is the Euler function that counts the generators of a cyclic group of order m, equivalently, the number of integers in $[1, m-1]$ that are relatively prime to m.

Proof. Let α be a generator for $\mathbb{F}_{p^n}^\times$ with minimal polynomial $f_\alpha \in \mathbb{F}_p[x]$; then f_α is primitive. There are $\phi(p^n - 1)$ possible choices for α. Conversely, if $f \in \mathbb{F}_p[x]$ is a primitive polynomial of degree n then each of its n roots is a generator for $\mathbb{F}_q^\times$. We thus have a surjective n-to-1 map $\alpha \to f_\alpha$ from the set of generators of $\mathbb{F}_{p^n}^\times$ to the set of primitive polynomials over $\mathbb{F}_p$ of degree n; the theorem follows. □

The preceding theorem implies that there are plenty of irreducible (and even primitive) polynomials $f \in \mathbb{F}_p[x]$ that we can use to represent $\mathbb{F}_q = \mathbb{F}_p[x]/(f)$ when q is not prime. The choice of the polynomial f has some impact on the cost of reducing polynomials in $\mathbb{F}_p[x]$ modulo f; ideally we would like f to have as few nonzero coefficients as possible. We can choose f to be a binomial only when its degree divides $p - 1$, but we can usually (although not always) choose f to be a trinomial; see [8]. Finite fields in cryptographic standards are often specified using an $f \in \mathbb{F}_p[x]$ that makes reduction modulo f particularly efficient.

For mathematical purposes it is more useful to fix a universal choice of primitive polynomials once and for all; this simplifies the task of migrating data from one computer algebra system to another, as well as the restoration of archived data. A simple way to do this is to take the lexicographically minimal primitive polynomial $f_{p,n} \in \mathbb{F}_p[x]$ of each degree n, where we represent each $f_{p,n}(x) = \sum a_i x^i$ as a sequence of integers $(a_0, \ldots, a_{n-1}, 1)$ with $0 \le a_i < p$.

There are two downsides to this simple-minded approach. First (and most significantly), we would like to be able to easily embed $\mathbb{F}_p^m$ in $\mathbb{F}_p^n$ when $m|n$, which means that if α is a root of $f_{p,n}(x)$ then we would really like $\alpha^{n/m}$ to be a root of $f_{p,m}(x)$, including when $m = 1$. Secondly (and less significantly), we would like the root r of $f_{p,1} = x - r$ to be the least primitive root modulo p, which will not be the case if we use the lexicographic ordering defined above, but will be the case if we tweak our sign convention and take $(a_0, \ldots, a_{n-1}, 1)$ to represent the polynomial $x^n - a_{n-1}x^{n-1} + \cdots + (-1)^n a_0$. This leads to the following recursive definition due to Richard Parker (named in honor of John Conway).

Definition 3.15. Order polynomials $f(x) = x^n - a_{n-1}x^{n-1} + \cdots + (-1)^n a_0 \in (\mathbb{Z}/p\mathbb{Z})[x]$ with $0 \le a_i < p$ according to the lexicographic order on integer sequences $(a_0, \ldots, a_{n-1}, 1)$. For each prime p and $n > 0$ the *Conway polynomial* $f_{p,n}(x)$ is defined by:

- For $n = 1$, let $f_{p,1}(x) := x - r$, where r is the least positive integer generating $(\mathbb{Z}/p\mathbb{Z})^\times$;
- For $n > 1$, let $f_{p,n}(x)$ be the least primitive polynomial of degree n such that for $0 < m < n$ and every root α of $f_{p,m}(x)$ we have $f_{p,n}(\alpha^{n/m}) = 0$.

That $f_{p,n}(x)$ exists is a straight-forward proof by induction that we leave as an exercise.

Conway polynomials are now used by most computer algebra systems, including GAP, Magma, Macaulay2, and SageMath. One downside to their recursive definition is that it is quite time consuming to compute any particular Conway polynomial on demand; instead, each of these computer algebra systems includes a list of precomputed Conway polynomials. The key point is that, even in a post-apocalyptic scenario where all these tables are lost, they can all be readily reconstructed from the succinct definition above.

Having fixed a representation for $\mathbb{F}_q$, every finite field operation can ultimately be reduced to integer arithmetic: elements of $\mathbb{F}_p$ are represented as integers in $[0, p-1]$, and elements of $\mathbb{F}_q = \mathbb{F}_p[x]/(f)$ are represented as polynomials of degree less than $\deg f$ whose coefficients are integers in $[0, p-1]$. We will see exactly how to efficiently reduce arithmetic in $\mathbb{F}_q$ to integer arithmetic in the next lecture. In the rest of this lecture we consider the complexity of integer arithmetic.

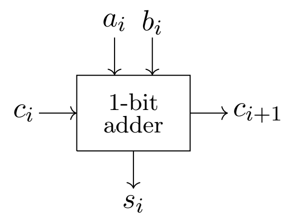
3.2 Integer addition

Every nonnegative integer a has a unique *binary representation* $a = \sum_{i=0}^{n-1} a_i 2^i$ with $a_i \in \{0, 1\}$ and $a_{n-1} \ne 0$. The binary digits a_i are called *bits*, and we say that a is an *n-bit integer*; we can represent negative integers by including an additional sign bit.

To add two integers in their binary representations we apply the "schoolbook" method, adding bits and carrying as needed. For example, we can compute 43+37=80 in binary as

$$\begin{array}{r} {\scriptstyle 101111} \\ 101011 \\ +100101 \\ \hline 1010000 \end{array}$$

The carry bits are shown in red. To see how this might implemented in a computer, consider a 1-bit adder that takes two bits a_i and b_i to be added, along with a carry bit c_i.



$$c_{i+1} = (a_i \wedge b_i) \vee (c_i \wedge a_i) \vee (c_i \wedge b_i)$$

$$s_i = a_i \otimes b_i \otimes c_i$$

The symbols $\wedge$, $\vee$, and $\otimes$ denote the boolean functions AND, OR, and XOR (exclusive-or) respectively, which we may regard as primitive components of a boolean circuit. By chaining $n+1$ of these 1-bit adders together, we can add two n-bit numbers using $7n+7 = O(n)$ boolean operations on individual bits.

Remark 3.16. Chaining adders is known as *ripple* addition and is no longer commonly used, since it forces a sequential computation. In practice more sophisticated methods such as *carry-lookahead* are used to facilitate parallelism. This allows most modern microprocessors to add two 64 (or even 128) bit integers in a single clock cycle, and with the SIMD (Single Instruction Multiple Data) instruction sets available on newer AMD and Intel processors, one may be able to perform four (or even eight) 64 bit additions in a single clock cycle.

We could instead represent the same integer a as a sequence of words rather than bits. For example, write $a = \sum_{i=0}^{k-1} a_i 2^{64i}$, where $k = \left\lceil \dfrac{n}{64} \right\rceil$. We may then add two integers using a sequence of $O(k)$, equivalently, $O(n)$, operations on 64-bit words. Each word operation is ultimately implemented as a boolean circuit that involves operations on individual bits, but since the word-size is fixed, the number of bit operations required to implement any particular word operation is a constant. So the number of bit operations is again $O(n)$, and if we ignore constant factors it does not matter whether we count bit or word operations.

Subtraction is analogous to addition (now we need to borrow rather than carry), and has the same complexity, so we will not distinguish these operations when analyzing the complexity of algorithms. With addition and subtraction of integers, we have everything we need to perform addition and subtraction in a finite field. To add two elements of $\mathbb{F}_p \simeq \mathbb{Z}/p\mathbb{Z}$ that are uniquely represented as integers in the interval $[0, p-1]$ we simply add the integers and check whether the result is greater than or equal to p; if so we subtract p to obtain a value in $[0, p-1]$. Similarly, after subtracting two integers we add p if the result is negative. The total work involved is still $O(n)$ bit operations, where $n = \lg p$ is the number of bits needed to represent a finite field element.

To add or subtract two elements of $\mathbb{F}_q \simeq (\mathbb{Z}/p\mathbb{Z})[x]/(f)$ we simply add or subtract the corresponding coefficients of the polynomials, for a total cost of $O(d \lg p)$ bit operations, where $d = \deg f$, which is again $O(n)$ bit operations, if we put $n = \lg q = d \lg p$.

Theorem 3.17. *The time to add or subtract two elements of $\mathbb{F}_q$ in our standard representation is $O(n)$, where $n = \lg q$ is the size of a finite field element.*

Remark 3.18. We will discuss the problem of reducing an integer modulo a prime p using fast Euclidean division in the next lecture. But this operation is not needed to reduce the sum or difference of two integers in $[0, p-1]$ to a representative in $[0, p-1]$; it is faster (both in theory and practice) to simply subtract or add p as required (at most once).

3.3 A quick refresher on asymptotic notation

Let f and g be two real-valued functions whose domains include the positive integers. The *big-O* notation "$f(n) = O(g(n))$" is shorthand for the statement:

There exist constants c and N such that for all $n \geq N$ we have $|f(n)| \leq c|g(n)|$.

This is equivalent to

$$\limsup_{n\to\infty} \frac{|f(n)|}{|g(n)|} < \infty.$$

Warning 3.19. "$f(n) = O(g(n))$" is an abuse of notation; in words we would say $f(n)$ is $O(g(n))$, where the word "is" does not imply equality (e.g., "Aristotle is a man"), and it is generally better to write this way. Symbolically, it would make more sense to write $f(n) \in O(g(n))$, regarding $O(g(n))$ as a set of functions. Some do, but the notation $f(n) = O(g(n))$ is far more common and we will occasionally use it in this course, with one caveat: we will never write a big-O expression to the left of the equal sign. It may be true that $f(n) = O(n \log n)$ implies $f(n) = O(n^2)$, but we avoid writing $O(n \log n) = O(n^2)$ because $O(n^2) \neq O(n \log n)$.

We also have *big-*Ω notation "$f(n) = \Omega(g(n))$", which means $g(n) = O(f(n))$,[2] as well as *little-*o notation "$f(n) = o(g(n))$," which is shorthand for

$$\lim_{n\to\infty} \frac{|f(n)|}{|g(n)|} = 0.$$

An alternative notation that is sometimes used is $f \ll g$, but depending on the author this may mean $f(n) = o(g(n))$ or $f(n) = O(g(n))$ (computer scientists tend to mean the former, while number theorists usually mean the latter, so we will avoid this notation). There is also a little-omega notation, but the symbol ω already has so many uses in number theory that we will not burden it further (we can always use little-o notation instead). The notation $f(n) = \Theta(g(n))$ means that $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$ both hold.

It is easy to see that the complexity of integer addition is $\Theta(n)$, since we have shown it is $O(n)$ and it is clearly $\Omega(n)$ because it takes this long to output n bits (in a Turing machine model one can show that for most inputs the machine will have to write to $\Omega(n)$ cells on the Turing tape, no matter what algorithm it uses).

Warning 3.20. Don't confuse a big-O statement with a big-Θ statement; the former implies only an upper bound. If Alice has an algorithm that is $O(2^n)$ this does not mean that Alice's algorithm requires exponential time, and it does not mean that Bob's $O(n^2)$ algorithm is better; Alice's algorithm could be $O(n)$ for all we know. But if Alice's algorithm is $\Omega(2^n)$ then we would definitely prefer to use Bob's algorithm for all sufficiently large n.

Big-O notation can also be used for multi-variable functions: "$f(m,n) = O(g(m,n))$" is shorthand for the statement:

There exist constants c and N such that for all $m, n \geq N$ we have $|f(m,n)| \leq c|g(m,n)|$.

This statement is weaker than it appears. For example, it says nothing about the relationship between $f(m,n)$ and $g(m,n)$ if we fix one of the variables. However, in virtually all of the examples we will see it will actually be true that if we regard $f(m,n) = f_m(n)$ and $g(m,n) = g_m(n)$ as functions of n with a fixed parameter m, we have $f_m(n) = O(g_m(n))$, and similarly, $f_n(m) = O(g_n(m))$. In this situation one says that $f(m,n) = O(g(m,n))$ *uniformly* in m and n.

[2] The Ω-notation originally defined by Hardy and Littlewood had a slightly weaker definition, but modern usage generally follows our convention, which is due to Knuth.

So far we have spoken only of *time complexity*, but *space complexity* plays a crucial role in many algorithms that we will see in later lectures. Space complexity measures the amount of memory an algorithm requires; this can never be greater than its time complexity (it takes time to use space), but it may be smaller. When we speak of “the complexity" of an algorithm, we should really consider both time and space. An upper bound on the time complexity is also an upper bound on the space complexity but it is often possible (and desirable) to obtain a better bound for the space complexity.

For more information on asymptotic notation and algorithmic complexity, see [2].

Warning 3.21. In this class, unless explicitly stated otherwise, our asymptotic bounds always count bit operations (as opposed to finite field operations, or integer operations). When comparing complexity bounds found in the literature, one must be sure to understand exactly what is being counted. For example, a complexity bound that counts operations in finite fields may need to be converted to a bit complexity to get an accurate comparison, and this conversion is going to depend on exactly which finite field operations are being used and how the finite fields are represented. A lack of care in this regard has led to more than one erroneous claim in the literature.

3.4 Integer multiplication

We now consider the problem of integer multiplication. Unlike addition, this is an open problem (no one knows how fast we can multiply) and remains an active area of research. Because we do not know the complexity of integer multiplication, it is common to use the notation $\mathsf{M}(n)$ to denote the time to multiply two n-bit integers; this allows one to state bounds for algorithms that depend on the complexity of integer multiplication in a way that does not depend on whatever the current state of the art is.

3.4.1 Schoolbook method

Let us compute $37 \times 43 = 1591$ with the “schoolbook” method, using a binary representation.

$$\begin{array}{r} 101011 \\ \times\ 100101 \\ \hline 101011 \\ 101011 \\ +101011 \\ \hline 11000110111 \end{array}$$

Multiplying individual bits is easy (just use an AND gate), but we need to do n^2 bit multiplications, followed by n additions of n-bit numbers (suitably shifted). The complexity of this algorithm is thus $\Theta(n^2)$. This gives us the upper bound $\mathsf{M}(n) = O(n^2)$. The only lower bound known is the trivial one, $\mathsf{M}(n) = \Omega(n)$, so one might hope to do better than $O(n^2)$, and indeed we can.

3.4.2 Karatsuba's algorithm

Before presenting Karatsuba's algorithm, it is worth making a few remarks regarding its origin. In the first half of the twentieth century it was widely believed that $\mathsf{M}(n) = \Omega(n^2)$; indeed, no less a mathematician than Kolmogorov formally stated this conjecture in a 1956 meeting of the Moscow Mathematical Society [7, §5]. This conjecture was one of the topics at a 1960 seminar led by Kolmogorov, with Karatsuba in attendance. Within the first week of the seminar, Karatsuba was able to disprove the conjecture. Looking back on the event, Karatsuba writes [7, §6]

> *After the next seminar I told Kolmogorov about the new algorithm and about the disproof of the n^2 conjecture. Kolmogorov was very agitated because this contradicted his very plausible conjecture. At the next meeting of the seminar, Kolmogorov himself told the participants about my method and at this point the seminar was terminated.*

Karatsuba's algorithm is based on a divide-and-conquer approach. Rather than representing n-bit integers using n digits in base 2, we may instead write them in base $2^{n/2}$ and may compute their product as follows

$$\begin{aligned} a &= a_0 + 2^{n/2}a_1, \\ b &= b_0 + 2^{n/2}b_1, \\ ab &= a_0b_0 + 2^{n/2}(a_1b_0 + b_1a_0) + 2^n a_1b_1, \end{aligned}$$

As written, this reduces an n-bit multiplication to four multiplications of $(n/2)$-bit integers and three additions of $O(n)$-bit integers (multiplying an intermediate result by a power of 2 can be achieved by simply writing the binary output "further to the left" and is effectively free). However, as observed by Karatsuba one can use the identity

$$a_0b_1 + b_0a_1 = (a_0 + a_1)(b_0 + b_1) - a_0b_0 - a_1b_1$$

to compute $a_0b_1 + b_0a_1$ using just one multiplication in addition to computing the products a_0b_0 and a_1b_1. By reusing the common subexpressions a_0b_0 and a_1b_1, we can compute ab using three multiplications and six additions (we count subtractions as additions). We can use the same idea to recursively compute the three products a_0b_0, a_1b_1, and $(a_0+a_1)(b_0+b_1)$; this recursive approach yields Karatsuba's algorithm.

If we let $T(n)$ denote the running time of this algorithm, we have

$$\begin{aligned} T(n) &= 3T(n/2) + O(n) \\ &= O(n^{\lg 3}) \end{aligned}$$

It follows that $\mathsf{M}(n) = O(n^{\lg 3})$, where $\lg 3 \coloneqq \log_2 3 \approx 1.59$.[3]

3.4.3 The Fast Fourier Transform (FFT)

The fast Fourier transform is widely regarded as one of the top ten algorithms of the twentieth century [3, 5], and has applications throughout applied mathematics. Here we focus on the discrete Fourier transform (DFT), and its application to multiplying integers and polynomials, following the presentation in [9, §8]. It is actually more natural to address the problem of polynomial multiplication first.

Let R be a commutative ring containing a primitive nth root of unity ω, by which we mean that $\omega^n = 1$ and $\omega^i - \omega^j$ is not a zero divisor for $0 \le i < j < n$ (when R is a field this coincides with the usual definition). We shall identify the set of polynomials in $R[x]$

[3] In general we shall use $\lg n$ to denote $\log_2 n$.

of degree less than n with the set of all n-tuples with entries in R. Thus we represent the polynomial $f(x) = \sum_{i=0}^{n-1} f_i x^i$ by its coefficient vector $(f_0, \ldots, f_{n-1}) \in R^n$ and may speak of the polynomial $f \in R[x]$ and the vector $f \in R^n$ interchangeably.

The discrete Fourier transform $\mathrm{DFT}_\omega : R^n \to R^n$ is the R-linear map

$$(f_0, \ldots, f_{n-1}) \xrightarrow{\mathrm{DFT}_\omega} (f(\omega^0), \ldots, f(\omega^{n-1})).$$

You should think of this map as a conversion between two types of polynomial representations: we take a polynomial of degree less than n represented by n coefficients (its *coefficient-representation* and convert it to a representation that gives its values at n known points (its *point-representation*).

One can use Lagrange interpolation to recover the coefficient representation from the point representation, but our decision to use values $\omega^0, \ldots, \omega^{n-1}$ that are nth roots of unity allows us to do this more efficiently. If we define the Vandermonde matrix

$$V_\omega := \begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{n-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2n-2} \\ 1 & \omega^3 & \omega^6 & \cdots & \omega^{3n-3} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & \omega^{n-1} & \omega^{2n-2} & \cdots & \omega^{(n-1)^2} \end{pmatrix},$$

then $\mathrm{DFT}_\omega(f) = V_\omega f^{\mathrm{tr}}$. Our assumption that none of the differences $\omega^i - \omega^j$ is a zero divisor in R ensures that the matrix V_ω is invertible, and its inverse is simply $\frac{1}{n} V_{\omega^{-1}}$. It follows that

$$\mathrm{DFT}_\omega^{-1} = \frac{1}{n} \mathrm{DFT}_{\omega^{-1}}.$$

Thus if we have an algorithm to compute DFT_ω we can use it to compute DFT_ω^{-1}: just replace ω by $\omega^{-1} = \omega^{n-1}$ and multiply the result by $\frac{1}{n}$.

We now define the *cyclic convolution* $f * g$ of two polynomials $f, g \in R^n$:

$$f * g = fg \bmod (x^n - 1).$$

Reducing the product on the right modulo $x^n - 1$ ensures that $f * g$ is a polynomial of degree less than n, thus we may regard the cyclic convolution as a map $R^n \times R^n \to R^n$. If $h = f * g$, then $h_i = \sum f_j g_k$, where the sum is over $j + k \equiv i \bmod n$. If f and g both have degree less than $n/2$, then $f * g = fg$; thus the cyclic convolution of f and g can be used to compute their product, provided that we make n big enough.

We also define the *pointwise product* $f \cdot g$ of two vectors in $f, g \in R^n$:

$$f \cdot g = (f_0 g_0,\ f_1 g_1,\ \ldots,\ f_{n-1} g_{n-1}).$$

We have now defined three operations on vectors in R^n: the binary operations of convolution and point-wise product, and the unary operation DFT_ω. The following theorem relates these three operations and is the key to the fast Fourier transform.

Theorem 3.22. $\mathrm{DFT}_\omega(f * g) = \mathrm{DFT}_\omega(f) \cdot \mathrm{DFT}_\omega(g)$.

Proof. Since $f * g = fg \bmod (x^n - 1)$, we have

$$f * g = fg + q \cdot (x^n - 1)$$

for some polynomial $q \in R[x]$. For every integer i from 0 to $n-1$ we then have

$$\begin{aligned}(f * g)(\omega^i) &= f(\omega^i)g(\omega^i) + q(\omega^i)(\omega^{in} - 1)\\ &= f(\omega^i)g(\omega^i),\end{aligned}$$

where we have used $(\omega^{in} - 1) = 0$, since ω is an nth root of unity. □

The theorem implies that if f and g are polynomials of degree less then $n/2$ then

$$fg = f * g = \mathrm{DFT}_\omega^{-1}(\mathrm{DFT}_\omega(f) \cdot \mathrm{DFT}_\omega(g)). \tag{1}$$

This identify allows us to multiply polynomials using the discrete Fourier transform. In order to put this into practice, we need an efficient way to compute DFT_ω. This is achieved by the following recursive algorithm.

Algorithm: Fast Fourier Transform (FFT)
Input: A positive integer $n = 2^k$, a vector $f \in R^n$, and the vector $(\omega^0, \ldots, \omega^{n-1}) \in R^n$.
Output: $\mathrm{DFT}_\omega(f) \in R^n$.

1. If $n = 1$ then return (f_0) and terminate.
2. Write the polynomial $f(x)$ in the form $f(x) = g(x) + x^{\frac{n}{2}}h(x)$, where $g, h \in R^{\frac{n}{2}}$.
3. Compute the vectors $r = g + h$ and $s = (g - h) \cdot (\omega^0, \ldots, \omega^{\frac{n}{2}-1})$ in $R^{\frac{n}{2}}$.
4. Recursively compute $\mathrm{DFT}_{\omega^2}(r)$ and $\mathrm{DFT}_{\omega^2}(s)$ using $(\omega^0, \omega^2, \ldots, \omega^{n-2})$.
5. Return the vector $(r(\omega^0), s(\omega^0), r(\omega^2), s(\omega^2), \ldots, r(\omega^{n-2}), s(\omega^{n-2}))$

Let $T(n)$ be the number of operations in R used by the FFT algorithm. Then

$$\begin{aligned}T(n) &= 2T(n/2) + O(n)\\ &= O(n \log n).\end{aligned}$$

This shows that the FFT is fast (justifying its name); let us now prove that it is correct.

Theorem 3.23. *The FFT algorithm outputs* $\mathrm{DFT}_\omega(f)$.

Proof. We must verify that the kth entry of the output vector is $f(\omega^k)$, for $0 \le k < n$. For even $k = 2i$ we have:

$$\begin{aligned}f(\omega^{2i}) &= g(\omega^{2i}) + (\omega^{2i})^{n/2}h(\omega^{2i})\\ &= g(\omega^{2i}) + h(\omega^{2i})\\ &= r(\omega^{2i}).\end{aligned}$$

For odd $k = 2i + 1$ we have:

$$\begin{aligned}
f(\omega^{2i+1}) &= \sum_{0\le j<n/2} f_j\omega^{(2i+1)j} + \sum_{0\le j<n/2} f_{n/2+j}\omega^{(2i+1)(n/2+j)}\\
&= \sum_{0\le j<n/2} g_j\omega^{2ij}\omega^j + \sum_{0\le j<n/2} h_j\omega^{2ij}\omega^{in}\omega^{n/2}\omega^j\\
&= \sum_{0\le j<n/2} (g_j - h_j)\omega^j\omega^{2ij}\\
&= \sum_{0\le j<n/2} s_j\omega^{2ij}\\
&= s(\omega^{2i}),
\end{aligned}$$

where we have used the fact that $\omega^{n/2} = -1$. □

Corollary 3.24. *Let R be a commutative ring containing a primitive nth root of unity, with $n = 2^k$, and assume $2 \in R^\times$. We can multiply two polynomials in $R[x]$ of degree less than $n/2$ using $O(n \log n)$ operations in R.*

Proof. From (1) we have

$$fg = \mathrm{DFT}_\omega^{-1}(\mathrm{DFT}_\omega(f)\cdot\mathrm{DFT}_\omega(g)) = \frac{1}{n}\,\mathrm{DFT}_{\omega^{-1}}(\mathrm{DFT}_\omega(f)\cdot\mathrm{DFT}_\omega(g))$$

and we note that $n = 2^k \in R^\times$ is invertible. We can compute $\omega^0, \ldots, \omega^{n-1}$ using $O(n)$ multiplications in R (this also gives us $(\omega^{-1})^0, \ldots, (\omega^{-1})^{n-1}$). Computing DFT_ω and $\mathrm{DFT}_{\omega^{-1}}$ via the FFT algorithm uses $O(n \log n)$ operations in R, computing the pointwise product of $\mathrm{DFT}_\omega(f)$ and $\mathrm{DFT}_\omega(g)$ uses $O(n)$ operations in R, and computing $1/n$ and multiplying a polynomial of degree less than n by this scalar uses $O(n)$ operations in R. □

What about rings that do not contain an nth root of unity? By extending R to a new ring $R' := R[\omega]/(\omega^n - 1)$ we can obtain a formal nth root of unity ω, and one can then generalize Corollary 3.24 to multiply polynomials in any ring R in which 2 is invertible using $O(n \log n \log\log n)$ operations in R; see [9, §8.3] for details.

The need for 2 to be invertible can be overcome by considering a 3-adic version of the FFT algorithm that works in rings R in which 3 is invertible. For rings in which neither 2 nor 3 is invertible we instead compute $2^k fg$ and $3^m fg$ (just leave out the multiplication by $1/n$ at the end). Once we know both $2^k fg$ and $3^m fg$ we can recover the coefficients of fg by using the Euclidean algorithm to compute $u, v \in \mathbb{Z}$ such that $u2^k + v3^m = 1$ and applying $u2^k fg + v3^m fg = fg$.

3.5 Integer multiplication

To any positive integer $a = \sum_{i=0}^{n-1} a_i 2^i$ we may associate the polynomial $f_a(x) = \sum_{i=0}^{n} a_i x^i \in \mathbb{Z}[x]$, with $a_i \in \{0, 1\}$, so that $a = f_a(2)$. We can then multiply positive integers a and b via

$$ab = f_{ab}(2) = (f_a f_b)(2).$$

Note that the polynomials $f_a(x)f_b(x)$ and $f_{ab}(x)$ may differ (the former may have coefficients greater than 1), but they take the same value at $x = 2$; in practice one typically uses base 2^{64} rather than base 2 (the a_i and b_i are then integers in $[0, 2^{64} - 1]$).

Applying the generalization of Corollary 3.24 discussed above to the ring $\mathbb{Z}$, Schönhage and Strassen [11] obtain an algorithm to multiply two n-bit integers in time $O(n \log n \log\log n)$, which gives us a new upper bound

$$\mathsf{M}(n) = O(n \log n \log\log n).$$

Remark 3.25. In 2007 Fürer [4] showed that this bound can been improved to

$$\mathsf{M}(n) = O\left(n \log n \, 2^{O(\log^* n)}\right)$$

where $\log^* n$ denotes the iterated logarithm, which counts how many times the log function must be applied to n before the result is less than or equal to 1. In 2016 Harvey, van der Hoeven and Lercer [10] proved the sharper bound

$$\mathsf{M}(n) = O\left(n \log n \, 8^{\log^* n}\right),$$

and in 2018 Harvey and van der Hoeven [6] further improved this to

$$\mathsf{M}(n) = O\left(n \log n \, 4^{\log^* n}\right).$$

But these improvements, and even the original Schönhage–Strassen algorithm, are primarily of theoretical interest: in practice one uses the "three primes" algorithm sketched below, which for integers with $n \le 2^{62}$ bits has a "practical complexity" of $O(n \log n)$ (this statement is mathematically meaningless but gives a rough indication of how the complexity scales as n varies in this bounded range).

The details of the Schoönhage and Strassen algorithm and its subsequent improvements are rather involved. There is a simpler approach used in practice to multiply integers less than $2^{2^{62}}$; this includes integers that would require 500 petabytes (500,000 terabytes) to write down and is more than enough for any practical application that is likely to arise in the near future. Let us briefly outline this approach.

Write the positive integers $a, b < 2^{2^{62}}$ that we wish to multiply in base 2^{64} as $a = \sum a_i 2^{64i}$ and $b = \sum b_i 2^{64i}$, with $0 \le a_i, b_i < 2^{64}$, and define the polynomials $f_a = \sum a_i x^i \in \mathbb{Z}[x]$ and $f_b = \sum b_i x^i \in \mathbb{Z}[x]$ as above. Our goal is to compute $f_{ab}(2^{64}) = (f_a f_b)(2^{64})$, and we note that the polynomial $f_a f_b \in \mathbb{Z}[x]$ has less than $2^{62}/64 = 2^{56}$ coefficients, each of which is bounded by $2^{56}2^{64}2^{64} < 2^{184}$.

Rather than working over a single ring R we will use three finite fields $\mathbb{F}_p$ of odd characteristic, where p is one of the primes

$$p_1 := 71 \cdot 2^{57} + 1, \qquad p_2 := 75 \cdot 2^{57} + 1, \qquad p_3 := 95 \cdot 2^{57} + 1.$$

Note that if p is any of the primes p_1, p_2, p_3, then $\mathbb{F}_p^\times$ is a cyclic group whose order $p-1$ is divisible by 2^{57}, which implies that $\mathbb{F}_p$ contains a primitive 2^{57}th root of unity ω; indeed, for $p = p_1, p_2, p_3$ we can use $\omega = \omega_1, \omega_2, \omega_3$, respectively, where $\omega_1 = 287, \omega_2 = 149, \omega_3 = 55$.

We can thus use the FFT Algorithm above with $R = \mathbb{F}_p$ to compute $f_a f_b \bmod p$ for each of the primes $p \in \{p_1, p_2, p_3\}$. This gives us the values of the coefficients of $f_a f_b \in \mathbb{Z}[x]$ modulo three primes whose product $p_1 p_2 p_3 > 2^{189}$ is more than large enough to uniquely the coefficients via the Chinese Remainder Theorem (CRT); the time to recover the integer coefficients of $f_a f_b$ from their values modulo p_1, p_2, p_3 is negligible compared to the time to apply the FFT algorithm over these three fields. If a and b are significantly smaller, say $a, b \le 2^{2^{44}}$, a "one prime" approach suffices.

3.6 Kronecker substitution

We now note an important converse to the idea of using polynomial multiplication to multiply integers: we can use integer multiplication to multiply polynomials. This is quite useful in practice, as it allows us take advantage of very fast implementations of FFT–based integer multiplication that are now widely available. If f is a polynomial in $\mathbb{F}_p[x]$, we can lift f to $\hat{f} \in \mathbb{Z}[x]$ by representing its coefficients as integers in $[0, p-1]$. If we then consider the integer $\hat{f}(2^m)$, where $m = \lceil 2\lg p + \lg(\deg f + 1)\rceil$, the coefficients of $\hat{f}$ will appear in the binary representation of $\hat{f}(2^m)$ separated by blocks of $m - \lceil \lg p \rceil$ zeros. If g is a polynomial of similar degree, we can easily recover the coefficients of $\hat{h} = \hat{f}\hat{g} \in \mathbb{Z}[x]$ in the integer product $N = \hat{f}(2^m)\hat{g}(2^m)$; we then reduce the coefficients of $\hat{h}$ modulo p to get $h = fg$. The key is to make m large enough so that the kth block of m binary digits in N contains the binary representation of the kth coefficient of $\hat{h}$.

This technique is known as *Kronecker substitution*, and it allows us to multiply two polynomials of degree d in $\mathbb{F}_p[x]$ in time $O(\mathsf{M}(d(n + \log d))$, where $n = \log p$. Typically we have $\log d = O(n)$, in which case this simplifies to $O(\mathsf{M}(dn))$ In particular, we can multiply elements of $\mathbb{F}_q \simeq \mathbb{F}_p[x]/(f)$ in time $O(\mathsf{M}(n))$, where $n = \log q$, provided that either $\log \deg f = O(n)$ or $\log p = O(1)$, which are the two most typical cases, corresponding to large characteristic and small characteristic fields, respectively.

Remark 3.26. When $\log d = O(n)$, if we make the standard assumption that $\mathsf{M}(n)$ grows super-linearly then using Kronecker substitution is strictly faster (by more than any constant factor) than a layered approach that uses the FFT to multiply polynomials and then recursively uses the FFT for the coefficient multiplications; this is because $\mathsf{M}(dn) = o(\mathsf{M}(d)\,\mathsf{M}(n))$.

3.7 Complexity of integer arithmetic

To sum up, we have the following complexity bounds for arithmetic on n-bit integers:

addition/subtraction	$O(n)$
multiplication (schoolbook)	$O(n^2)$
multiplication (Karatsuba)	$O(n^{\lg 3})$
multiplication (FFT)	$O(n \log n \log\log n)$

References

[1] Razvan Barbulescu, Pierrick Gaudry, Antoine Joux, Emmanuel Thomé, *A heuristic quasi-polynomial algorithm for discrete logarithm in finite fields of small characteristic*, in *Advances in Cryptology — EUROCRYPT 2014*, LNCS **8441** (2014), 1–16.

[2] Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein, *Introduction to algorithms*, third edition, MIT Press, 2009.

[3] Jack Dongarra, Francis Sullivan, *Top ten algorithms of the century*, Computing in Science and Engineering **2** (2000), 22–23.

[4] Martin Fürer, *Faster integer multiplication*, Proceedings of the thirty-ninth annual ACM Symposium on the Theory of Computing (STOC), 2007.

[5] Dan Givoli, *The top 10 computational methods of the 20th century*, IACM Expressions **11** (2001), 5–9.

[6] David Harvey and Joris van der Hoeven, *Faster integer multiplication using short lattice vectors*, Thirteenth Algorithmic Number Theory Symposium (ANTS XIII), Open Book Series **2** (2018), 293–310.

[7] A. A. Karatsuba, *The complexity of computations*, Proceedings of the Steklov Institute of Mathematics **211** (1995), 169–193 (English translation of Russian article).

[8] Jaochim von zur Gathen, *Irreducible trinomials over finite fields*, Mathematics of Computation **72** (2003), 1787–2000.

[9] Joachim von zur Gathen and Jürgen Gerhard, *Modern computer algebra*, third edition, Cambridge University Press, 2013.

[10] David Harvey, Joris van der Hoeven, and Grégoire Lecerf, *Even faster integer multiplication*, J. Complexity **36** (2016), 1–30.

[11] Arnold Schönhage and Volker Strassen, *Schnelle Multiplikation großer Zahlen*, Computing, **7** (1971), 281–292.