ESO207A: Data Structures and Algorithms	
Notes:	*Hashing and Basic Number Theory/Algorithms*

This is an extremely superficial introduction to the deep and wonderful area of algebraic algorithms, and computational number theory, with applications to hashing and cryptography.

1 Groups

The abstract algebra structure *group* plays a very special role in understanding numbers.

Definition 1. *A group G is a nonempty set of elements on which a binary product operation denoted by $\cdot$ is defined. (That is, the product operation takes any two elements $a, b \in G$ and returns an element $a \cdot b \in G$). G is a group if the following properties are satisfied.*

1. *G is closed under $\cdot$, that is, for any $a, b \in G$, $a \cdot b \in G$.*
2. *Product operation is associative, that is, for any $a, b, c \in G$, $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.*
3. *Existence of identity element: there exists an element $e \in G$ such that for all $a \in G$, $a \cdot e = e \cdot a = a$.*
4. Every *element has an inverse: For every $a \in G$, there exists an element denoted $a^{-1} \in G$ such that $a \cdot a^{-1} = a^{-1} \cdot a = e$.*

Groups are fundamental in algebra. For our simple purposes, we will be looking at the group $\{1, 2, \ldots, n-1\}$ with the operation multiplication modulo n, for n prime. This is called the *multiplicative group* modulo n. For example, let $n = 5$. The group is then the set $\{1, 2, 3, 4\}$ together with the operation $a \cdot b = ab \mod n$. If $a, b \neq 0$, then the product is non-zero and taking mod n places it in the set $\{1, \ldots, n-1\}$. Note that in this group $2^{-1} = 3$, since, $2 \cdot 3 = 1 \mod 5$ (and so $3^{-1} = 2$) and also $4^{-1} = 4$). Of course, $1^{-1} = 1$ since it is the identity element. We will later prove that multiplicative groups modulo prime are actually groups.

A group is called abelian or commutative if the product operation is commutative. For our simple applications, the product operation will be commutative, for example, $a \cdot b \mod n$, so we will not worry too much about it. But there is a rich theory for non-abelian groups.

The notation $o(G)$, called the order of G, denotes the number of elements in G. For example, for our multiplicative group modulo n, the order is $n - 1$.

The following lemma states some fundamental properties of groups that we will take for granted in the future. The proofs are given for your reference.

Lemma 1. *If G is a group, then the following hold.*

1. *The identity element of G is unique.*
2. *Every element $a \in G$ has a unique inverse.*

Proof. 1. Suppose there are two items e and f such that for all $a \in G$, $a \cdot e = e \cdot a = a$ and $a \cdot f = f \cdot a = a$. Then,

$$e = e \cdot f = f \ .$$

2. Suppose there are two elements $b, c \in G$ such that $a \cdot b = b \cdot a = e$ and $a \cdot c = c \cdot a = e$. Then, multiplying both sides of $a \cdot b = e$ by c we have

$$c \cdot (a \cdot b) = c \cdot e$$

But the *LHS* equals $(c \cdot a) \cdot b = e \cdot b = b$ and the *RHS* equals c. Thus, $b = c$, or that a^{-1} is unique.

$\square$

A consequence is the nice "cancellation property" that groups imply. Given $a \cdot x = a \cdot y$, we can "cancel" a (i.e., multiply by a^{-1} both sides) to conclude $x = y$ in G. Similarly, for $x \cdot a = y \cdot a$, we conclude that $x = y$ in G.

1.1 Subgroups

A nonempty subset H of a group G is said to form a *subgroup* of G, if under the product operation of G, H itself forms a group. Another equivalent way of saying this is the following. A non-empty subset H of a group G is a subgroup of G if (i) H is closed under $\cdot$ operation, and (ii) every element $a \in H$ has an inverse $a^{-1} \in H$ (existence of inverse).

Here is a simple example. Consider the multiplicative group modulo 5, namely $G = \{1, 2, 3, 4\}$ with multiplication modulo 5 as the product operation. For any element $a \in G$, consider the sequence, $1, a, a^2, \ldots,$. For example, the powers of 2 sequence is $1, 2, 4, 3$, after which the sequence repeats ($2^2 = 4, 2^3 = 3 \mod 5$ and $2^4 = 1 \mod 5$).

In any finite group G, the sequence $1, a, a^2, a^3, \ldots$ must cycle. Consider the earliest index k such that $a^k = a^j$ for some earlier index $j < k$. Then, by the cancellation property, $a^{k-j} = e$. Thus, the sequence is $1, a, a^2, \ldots, a^{k-j-1}$, after which the sequence will repeat itself. Here $k - j$ is called the order of a, denoted $o(a)$, namely, the smallest power such that when a is raised to this power, we get the identity element.

The set $\{1, a, a^2, \ldots, a^{o(a)-1}\}$ is a group and is called the group generated by a. The order of this group is the order of a. In the previous example, the order of the group generated by 2 was 4 (namely, the sequence $1, 2, 4, 3$). The group generated by 4 is $\{1, 4\}$ and the group generated by 3 is $\{1, 3, 4, 2\}$, namely, the whole group.

Definition 2. *Let G be a group and H be a subgroup of G. For $a, b \in G$, we say that a is congruent to b modulo H, denoted $a \equiv_H b$, if $ab^{-1} \in H$.*

The crucial insight is that $\equiv_H$ is an equivalence relation. This is easy to see. (1) Reflexivity: $aa^{-1} = e \in H$, since, H is a subgroup. (2) Symmetry: if $ab^{-1} \in H$, then, $(ab^{-1})^{-1} = ba^{-1} \in H$, or that $b \equiv_H a$. (3) Transitivity: Suppose $ab^{-1} \in H$ and $bc^{-1} \in H$, then, the product $ab^{-1} \cdot bc^{-1} = ac^{-1} \in H$, since, H is closed under product, and so $a \equiv_H c$.

By property of equivalence relation, the set G is partitioned into equivalence classes. Let $[a]_H$ denote the equivalence class to which a belongs, that is,

$$[a]_H = \{b \in G \mid a \equiv_H b\}$$

We will now give an alternative and very interesting definition of the equivalence classes above $[a]_H$.

Definition 3. *If H is a subgroup of G and $a \in H$, then the set $Ha = \{ha \mid h \in H\}$ is called a right coset of H.*

The key property is that $Ha = [a]_H$, that is, the right coset of H generated by a is exactly the equivalence class of a. Let us prove this now.

Lemma 2. *Show that $Ha = [a]_H = \{b \in G \mid a \equiv_H b\}$.*

Proof. $b \in Ha$ iff $ba^{-1} \in H$ iff $b \equiv_H a$. □

Since the right cosets of H are the equivalence classes, *any two right cosets are either identical or do not share any element.*

We will now show a 1-1 correspondence between any two right cosets. For simplicity consider the right cosets Ha and $He = H$. Consider the mapping $h \to ha$. This maps H to the right coset Ha. Clearly, this is onto. It is 1-1 since $h_1 a = h_2 a$ implies, by cancellation of a, that $h_1 = h_2$. Hence, there is a set isomorphism between H and Ha. Hence, all right cosets Ha are set isomorphic to each other, that they can be placed in a 1-1 correspondence with each other.

In particular, for finite groups, it means that all right cosets have the same size, and they partition the elements of G. Let $\imath(H, G)$ denote the number of right cosets of H in G. Each element of G belongs to one and only one right coset. So counting the elements of G, we have the fundamental equation that

$$o(G) = o(H) \times \ \imath(H, G)$$

This gives us the wonderful Lagrange's theorem.

Theorem 3. *Let G be a finite group and H be a subgroup of G. Then, $o(H)$ is a divisor of $o(G)$.*

1.2 An application to number theory

Let us apply the little bit of what we have learnt so far to the multiplicative group of numbers modulo p, where, p is a prime number. [1] Let G be the group $\{1, 2, \ldots, p-1\}$ with the $\cdot$ operation being multiplication $\mod n$. Let $a \in G$ and consider the group generated by a, namely, $H = \{1, a, a^2, \ldots, a^{o(a)-1}\}$. By Lagrange's theorem, we have, $o(a) = o(H)$ divides $o(G) = p - 1$. So $p - 1 = ko(a)$, or that,

$$a^{p-1} = a^{o(a)k} = (a^{o(a)})^k = 1^k = 1$$

This is Fermat's little theorem.

Theorem 4. *For any prime p, and $1 \leq a \leq p - 1$, $a^p = 1 \mod p$.*

[1] We haven't yet proved that it is a group, but will do so shortly after introducing gcd.

2 Very! Elementary Number Theory

Numbers in this section will mean integers, positive or zero or negative, that is, elements of $\mathbb{Z}$.

Division operation. Given any two numbers, a and b, one can always divide a by b to obtain a unique quotient and remainder as follows:

$$a = qb + r, \quad \text{where, } 0 \le r < |b| \ .$$

2.1 GCD and its computation

The *gcd* of two numbers a and b is defined as the largest common divisor of a and b. We will denote it by $\gcd(a, b)$.

Given two numbers a, b let $I(a, b)$ denote the set of all integer linear combinations of a and b, that is, $I(a, b) = \{ax + by \mid x, y \in \mathbb{Z}\}$. The set $I(a, b)$ is called the *ideal generated by a and b*, although we will not use much of its properties. Observe that (i) the ideal is closed under addition, that is, the sum of two elements $ax + by$ and $ax' + by'$ is again of the same type and hence a member of $I(a, b)$. (ii) the ideal absorbs multiplication by any integer, that is, given any member $ax + by \in I$ and any integer z, then, $z(ax + by)$ is a member of the ideal.

Let d be the smallest positive element of $I(a, b)$. The fundamental observation is that $d = \gcd(a, b)$ and that $I(a, b) = I(d)$. (Here, $I(d)$ is the set of all integer multiples of d, that is, $I(d) = \{xd \mid x \in \mathbb{Z}\}$.

Lemma 5. *$d = \gcd(a, b)$ and $I(d) = I(a, b)$.*

Proof. Since d is an element of $I(a, b)$, and $I(a, b)$ is closed under integer multiplications, $I(d) \subset I(a, b)$.

Divide a by d. Then, $a = qd + r$, where, $0 \le r < d$. If $r \ne 0$, then, $r = a - qd \in I(a, b)$ and this contradicts the definition of d. So, $r = 0$ or that d divides a. Analogously, d divides b. So (i) d is a common divisor of a and b, and (ii) d divides every element of $I(a, b)$. By (ii) $I(d) \supset I(a, b)$ and therefore $I(d) = I(a, b)$.

Since d is of the form $ax + by$ for some $x, y \in \mathbb{Z}$, if d' is a common divisor of a and b, then, d' divides $ax + by = d$. Hence, every common divisor of a and b divides d.

Hence, $d = \gcd(a, b)$.

□

Suppose $a \ge b$. Then, $I(a, b) = I(a - b, b)$, since, $a = a - b + b$, each member $xa + yb \in I(a, b)$ can be equivalently expressed as $x'(a - b) + y'b \in I(a - b, b)$ for appropriate x', y' and vice-versa. Hence, $\gcd(a, b) = \gcd(a - b, b)$. Like wise, for $a \ge 2b$, $\gcd(a, b) = \gcd(a - b, b) = \gcd(a - 2b, b)$. So in general, we get $\gcd(a, b) = \gcd(a \mod b, b)$. This is Euclid's famous algorithm.

GCD(a, b) // assumes $a \ge b \ge 0$
1. **if** $b == 0$ **return** a
2. **else return** GCD$(b, a \mod b)$

So what is the complexity of the above operation. Though we have not discussed integer operations, $a \mod b$ can be computed in quadratic time, that is, quadratic in the sum of the bits needed to express a and b.

To see how many times the recursion is called, note the following. If $b \leq a/2$, then $a \mod b < b \leq a/2$. If $b > a/2$, then, $a \mod b = a - b < a/2$. That is, in each case, the larger argument a becomes at most $a/2$ in one recursive step. Analogously, the second argument in a call b becomes the first argument after the first recursive call, and becomes at most $b/2$ in the second recursive call. Hence after two recursive calls, the sizes of a and b, each reduce by at least a factor of 2. Hence after $2(\lceil \log_2 b \rceil + 1)$ steps or sooner, one of the arguments becomes 0 and the recursion bottoms out.

The complexity of Euclid's gcd algorithm is therefore $O((\log a + \log b)^3)$, that is, cubic.

Let $d = \gcd(a, b)$. Then, as argued in Lemma 5, $d = ax + by$ for some integers x, y. It would be really convenient for many applications to generalize the gcd algorithm to not only compute $\gcd(a, b)$ but also produce these multiplicators x and y. Let $a' = a \mod b = a - (a/b)b$, where, a/b is the integer quotient when a is divided by b. Let $d = \gcd(b, a') = bx' + a'y'$. Then,

$$d = bx' + (a - (a/b)b)y' = ay' + b(x' - (a/b)y')$$

This gives the following generalized gcd algorithm below.

GEN_GCD(a, b) // returns the triple $(d = \gcd(a, b), x, y)$ such that $d = ax + by$
if $b == 0$ **return** $(a, 1, 0)$
else
    $(d, x', y') =$ GEN_GCD$(b, a \mod b)$
    return $(d, y', (x' - (a/b)y'))$

Finally, we note that if any number d is a common divisor of a and b can be expressed as $d = ax + by$, then, d is $\gcd(a, b)$.

2.2 Modular Division

Consider the set $G = \{1, 2, \ldots, p-1\}$ equipped with the product operation multiplication $\mod p$. Let p be prime. We have earlier claimed that this is a group. Of course, it is closed under multiplication and 1 is the identity element. So it remains to show that every $a \in G$ has an inverse. Note that p being prime, it has no non-trivial common factors with $1 \leq a \leq p - 1$. In other words, $\gcd(a, p) = 1$, for each $a \in G$. Hence, there exists integer multipliers x and y such that $ax + py = 1$, or that $ax = 1 - py$. Taking $\mod p$ of both sides, we have,

$$ax = 1 \mod p$$

The number x is the inverse a^{-1} of $a \in G$. This can be found efficiently by the generalized gcd algorithm. We have thus shown that the multiplicative group G is indeed a group when p is prime.

Consider $\gcd(4, 6) = 2$. Then, the equation $4x = 1 \mod 6$ has no solution, since, if it did, then, for some $x, y \in \mathbb{Z}$, we would have, $4x = 1 + 6y$ or, that $4x - 6y = 1$. But this would mean that the gcd of $4, 6$ is 1, which is not true. In general, $ax = 1 \mod n$ has a solution iff $\gcd(a, n) = 1$. Since, this is true for each $a \in \{1, \ldots, p - 1\}$, the multiplicative group modulo prime is indeed a group.

The general multiplicative group modulo n, for any number n is defined as follows. Its elements are

$$G = \{a \mid 1 \leq a \leq n-1 \text{ and } \gcd(a,n) = 1\}$$

That is, G consists of all the non-zero numbers less than n that are relatively prime to n. Thus, $G_6 = \{1, 5\}$ with multiplication $\mod 6$.

3 Universal Hashing

Suppose we want to hash a collection of m IP addresses. IP addresses are 32 bit numbers (logical address of a computer) written in dot notation, split up 8 bits at a time, for example 23.145.17.202. Each 8-bit segment can take $2^8 = 256$ values. The universe of IP addresses is therefore 2^{32} big. At any time, we may want to hash say all the IP addreses within a lab , which may not be very large, say at most 255.

Let us denote an IP address as a tuple (x_1, x_2, x_3, x_4) corresponding to each of the 8-bit segments– $0 \leq x_i \leq 255$, for $i = 1, 2, 3, 4$. Choose any function h that maps an IP address to a table of size $m = 255$, that is, for $0 \leq x_i \leq 255$, for $1 \leq i \leq 4$, $h(x_1, x_2, x_3, x_4) \in \{0, 1, \ldots, m-1\}$.

The problem with a deterministic function is that since the hash function maps 2^{32} IP addresses to $m = 255$ buckets, there will be a particular bucket to which at least $2^{32}/255$ IP addresses get mapped.

To do anything better, we need randomization. We instead, define a class $\mathcal{H}$ of hash functions, instead of just one. We then select a hash function at random from this class and use it. Here is an example. Instead of $m = 255$, let $m = 257$, a prime number. Now choose 4 numbers a_1, a_2, a_3, a_4 independently and randomly from the set $G = \{0, 1, 2, \ldots, 256\}$. That is $a_i \in G$ randomly, for each $i = 1, 2, \ldots, 4$. Let $a = (a_1, a_2, a_3, a_4)$. Now consider the hash function defined as

$$h_a(x_1, x_2, x_3, x_4) = a_1x_1 + a_2x_2 + \ldots + a_4x_4 \mod m = \sum_{i=1}^{4} a_ix_i \mod m \ .$$

What would we consider a good hash function? Suppose $x = (x_1, \ldots, x_4)$ and $y = (y_1, \ldots, y_4)$ are distinct IP addresses. Ideally, if h is a "random" hash function, then, the probability that x and y would map to the same bucket under h, that is, $\Pr(h(x) = h(y))$ should be $\frac{1}{m}$, since m is the number of buckets available.

If the coefficients a are picked at random, then, h_a is likely to a good hash function in the above sense. Let us state and prove the lemma.

Lemma 6. *Let $x = (x_1, \ldots, x_4)$ and $y = (y_1, \ldots, y_4)$ be different IP addresses. If the coefficients $a = (a_1, a_2, a_3, a_4)$ are chosen uniformly at random from $\{0, 1, \ldots, m-1\}$. Then,*

$$Pr(h_a(x_1, \ldots x_4) = h_a(y_1, \ldots, y_4)) = \frac{1}{m} \ .$$

Proof. Since, $x \neq y$, at least one of the quadruples must be different, say $x_4 \neq y_4$. Then, the equation $h_a(x) = h_a(y)$ can be written as

$$\sum_{i=1}^{3} a_i(x_i - y_i) = a_4(x_4 - y_4) \mod m$$

How many solutions $a = (a_1, a_2, a_3, a_4)$ satisfy this equation. Suppose we choose a_1, a_2, a_3 arbitrarily, that is in m^3 ways. Since, $x_4 - y_4 \neq 0$, there is a (unique) inverse $(x_4 - y_4)^{-1}$ in the multiplicative group $\mod m$. Therefore, a_4 is unique. The probability that a_4 assumes this unique value is $1/m$, which is what the lemma claims. □

The hash family $\mathcal{H} = \{h_a : a = (a_1, a_2, a_3, a_4), a_i \in \{0, 1 \ldots, m-1\}\}$ has the following property. Given any two distinct IP addresses x and y, there are exactly $|\mathcal{H}|/m$ hash functions that map x and y to the same bucket.

Such hash function families are called *universal*.