

4. (25%) This problem is an example of *universal hashing*, a strategy for picking hash functions for a hash table randomly so that no input always exhibits bad hashing behavior.

Let  $p$  be a prime number. I want to hash pairs of numbers  $(x, y)$ , where  $x$  and  $y$  are always between 0 and  $p - 1$  inclusive. I decide to use a chained hash table with hash function

$$h_{a,b}(x, y) = (ax + by) \bmod p$$

where  $a$  and  $b$  also lie between 0 and  $p - 1$ .

- (a) Suppose that my  $a$  and  $b$  are fixed, and that you've discovered what they are (perhaps by hacking into my computer). Describe how to generate  $p$  distinct input pairs  $(x_i, y_i)$  for which  $h_{a,b}(x_i, y_i)$  yields the same value. That is, all the inputs  $(x_i, y_i)$  will hash to the same slot of my table.

(**Hint:** for any  $c$ ,  $1 \leq c < p$ , and every  $i$ ,  $0 \leq i < p$ , there exists exactly one  $j$ ,  $0 \leq j < p$  for which  $cj \equiv i \pmod{p}$ .)

*Proof.* First, note that for any given  $p, z \in \{0, 1, 2, \dots, p - 1\} = \mathbb{Z}_p$ , we have a unique choice of  $q \in \mathbb{Z}_p$  such that  $p + q = z$ . Further, we know that for any given  $p$ , the  $q$  described above always exists. That is, to any number in  $\mathbb{Z}_p$ , we can always add another, unique number to construct any other number in  $\mathbb{Z}_p$ . Therefore, because we want to construct a series of pairs of values  $(x_1, y_1), (x_2, y_2) \dots (x_p, y_p)$  such that they all hash to the same value, we must first choose some value in  $\{0, 1, 2, \dots, p - 1\} = \mathbb{Z}_p$  that we want every value to hash to. Given that sum value, say  $z$ , we can allow  $x_i$  to vary over each value in  $\mathbb{Z}_p$ , in total,  $p$  values. Given that we know both  $a, b$ , we know the value for each  $ax_i$ , so we wish to construct  $by_i$  such that  $by_i = z - ax_i \bmod p$ . Given that this  $by_i$  is unique, and  $b$  is already fixed, we can simply choose the unique  $y_i$  as described in the hint such that  $by_i \bmod p = (z - ax_i) \bmod p$ . Therefore, we have now constructed  $p$  different pairs  $(x_i, y_i)$  such that  $(ax_i + by_i) \bmod p = z$ ,  $i = 1, 2, \dots, p$ .  $\square$

- (b) To defend against your malicious hackery, I have decided not to fix  $a$  and  $b$  once and for all, but rather to choose them randomly every time I instantiate my hash table class. Each value will be chosen uniformly at random (with replacement) from the range  $0 \dots p - 1$ .

Fix two non-identical inputs  $(x, y)$  and  $(x', y')$ . (They may have the same  $x$  or  $y$  values, but not both.) For how many distinct pairs  $(a, b)$  will these two inputs hash to the same slot?

(Use the hint from part (a)).

*Proof.* Given any  $(x, y)$  and  $(x', y')$  we wish to find the number of pairs  $(a, b)$  such that:

$$\begin{aligned} (ax + by) \bmod p &= (ax' + by') \bmod p \text{ and} \\ (ax + by) \bmod p - (ax' + by') \bmod p &= 0 \end{aligned}$$

given that  $a, b, x, y, x', y' \in \mathbb{Z}_p = \{0, 1, 2, \dots, p - 1\}$ . This is the same as the following expression:

$$(ax + by) - (ax' + by') = mp$$

where  $m \in \mathbb{Z}$  because  $mp$  is congruent to  $0 \bmod p \forall m \in \mathbb{Z}$ . Therefore, we can say:

$$\begin{aligned} a(x - x') + b(y - y') &= mp \\ &= 0 \bmod p \end{aligned}$$

Based on the initial assumptions, we are left with 2 cases:

- i.  $x = x'$  or  $y = y'$ .

Without loss of generality, assume that  $x = x'$  and  $y \neq y'$ . In this case, from the previous expression, we have:

$$a(0) + b(y - y') = 0 \pmod{p}$$

Because  $a(0) = 0$ , we must have that  $b(y - y') = 0$ .  $y - y' \neq 0$  is fixed, therefore, by the hint in part **(a)**, we must choose the unique  $b$  such that  $b(y - y') = 0$ . Therefore, there can only be a single choice for  $b$ , but  $p$  choices for  $a$ , because  $a$  is trivial. Therefore, there are trivially  $p$  choices for  $(a, b)$ , as  $a$  varies over all of  $\mathbb{Z}_p$  and  $b$  is fixed.

- ii.  $x \neq x'$  and  $y \neq y'$ .

Let  $x - x' = q$  and  $y - y' = r$ . In this case, we have:

$$\begin{aligned} a(x - x') + b(y - y') &= \\ a(q) + b(r) &= 0 \pmod{p} \end{aligned}$$

By the hint in **(a)**, we know that because  $q, r \neq 0$  are fixed, the equation  $aq = z \pmod{p}$  has exactly one solution for every element  $z \in \mathbb{Z}_p$ . Similarly, for every  $z \in \mathbb{Z}_p$ , there exists exactly one  $-z \in \mathbb{Z}_p$  such that  $-z + z = 0$ . Therefore, as  $a$  is allowed to vary, over the entire set,  $b$  must be chosen uniquely such that  $aq + br = 0 \pmod{p}$  holds. Because there are  $p$  possible choices for  $a$  (i.e. every element in  $\mathbb{Z}_p$ ), there are  $p$  unique combinations  $(a, b)$  such that  $(ax + by) \pmod{p} = (ax' + by') \pmod{p}$ .

□

- (c) If I choose each of  $a$  and  $b$  uniformly at random from  $0 \dots p-1$ , what is the probability that  $(x, y)$  and  $(x', y')$  will hash to the same value?

We have a  $\frac{1}{p}$  chance that  $(x, y)$  and  $(x', y')$  will hash to the same value.

*Proof.* Notice that for  $\{0, 1, 2, \dots, p-1\}$  there are  $p^2$  ordered pairs  $(a, b)$  that can be formed. From **(b)**, we know that there are  $p$  unique pairs  $(a, b)$  such that  $(ax + by) \pmod{p} = (ax' + by') \pmod{p}$ . Therefore, the probability that we choose  $a, b$  randomly from  $\{0, 1, 2, \dots, p-1\}$  such that  $(x, y)$  and  $(x', y')$  hash to the same values is given by:

$$P[(ax + by) \pmod{p} = (ax' + by') \pmod{p}] = \frac{p}{p^2} = \frac{1}{p}.$$

□

- (d) Given an arbitrary set of  $n$  distinct inputs  $(x_i, y_i)$ , what is the expected number (over my random choices of  $a$  and  $b$ ) of pairs  $i, j$ ,  $i < j$ , for which  $h_{a,b}(x_i, y_i) = h_{a,b}(x_j, y_j)$ ?

**(Hint: use linearity of expectation!)**