TD 11 : Universal Hashing, Rabin-Karp algorithm

1 ε-Universal hashing

Define a family $\mathscr{H}$ of hash functions from a finite set U to a finite set B to be *ε-universal* if for all pairs of distinct elements k and ℓ in U,

$$\mathbb{P}(h(k) = h(\ell)) \leq \varepsilon,$$

where the probability is over the choice of the hash function h drawn uniformly at random from the family $\mathscr{H}$.

Problem 1.1. Show that an ε-universal family of hash functions must have

$$\varepsilon \geq \frac{1}{|B|} - \frac{1}{|U|}.$$

Problem 1.2. Let U be the set of n-tuples of values drawn from $\mathbb{Z}_p$, and let $B = \mathbb{Z}_p$, where p is prime. Define the hash function $h_b : U \to B$ for $b \in \mathbb{Z}_p$ on an input n-tuple $\langle a_0, a_1, \ldots, a_{n-1}\rangle$ from U as

$$h_b(\langle a_0, a_1, \ldots, a_{n-1}\rangle) = \left(\sum_{j=0}^{n-1} a_j b^j\right) \mod p,$$

and let $\mathscr{H} = \{h_b : b \in \mathbb{Z}_p\}$. Argue that $\mathscr{H}$ is $((n-1)/p)$-universal.

2 k-Universal hashing

Let $\mathscr{H}$ be a class of hash functions in which each hash function $h \in \mathscr{H}$ maps the universe U of keys to $\{0, 1, \ldots, m-1\}$. We say that $\mathscr{H}$ is *k-universal* if, for every fixed sequence of k distinct keys $\langle x^{(1)}, \ldots, x^{(k)}\rangle$ and for any h chosen at random from $\mathscr{H}$, the sequence $\langle h(x^{(1)}), \ldots, h(x^{(k)})\rangle$ is equally likely to be any of the m^k sequences of length k with elements drawn from $\{0, 1, \ldots, m-1\}$.

Problem 2.1. Show that if the family $\mathscr{H}$ of hash functions is 2-universal, then it is universal.

Problem 2.2. Suppose that the universe U is the set of n-tuples of values drawn from $\mathbb{Z}_p = \{0, 1, \ldots, p-1\}$, where p is prime. Consider an element $x = \langle x_0, x_1, \ldots, x_{n-1}\rangle \in U$. For any n-tuple $a = \langle a_0, a_1, \ldots, a_{n-1}\rangle \in U$, define the hash function h_a by

$$h_a(x) = \left(\sum_{j=0}^{n-1} a_j x_j\right) \mod p.$$

Let $\mathscr{H} = \{h_a : a \in \mathbb{Z}_p\}$. Show that $\mathscr{H}$ is universal, but not 2-universal.

Problem 2.3. Suppose that we modify $\mathscr{H}$ slightly from the previous question : for any $a \in U$ and for any $b \in \mathbb{Z}_p$, define

$$h'_{ab}(x) = \left(\sum_{j=0}^{n-1} a_j x_j + b\right) \mod p$$

and $\mathscr{H}' = \{h'_{ab} : a, b \in \mathbb{Z}_p\}$. Argue that $\mathscr{H}'$ is 2-universal.

Problem 2.4. Suppose that Alice and Bob secretly agree on a hash function h from a 2-universal family $\mathscr{H}$ of hash functions. Each $h \in \mathscr{H}$ maps from a universe of keys U to $\mathbb{Z}_p$, where p is prime. Later, Alice sends a message m to Bob over the Internet, where $m \in U$. She authenticates this message to Bob by also sending an authentication tag $t = h(m)$, and Bob checks that the pair (m, t) he receives indeed satisfies $t = h(m)$. Suppose that an adversary intercepts (m, t) en route and tries to fool Bob by replacing the pair (m, t) with a different pair (m', t'). Argue that the probability that the adversary succeeds in fooling Bob into accepting (m', t') is at most $1/p$, no matter how much computing power the adversary has, and even if the adversary knows the family $\mathscr{H}$ of hash functions used.

3 Rabin-Karp string matching algorithm

Let $T[1 \dots n]$ be a text and $P[1 \dots m]$ be a pattern over an alphabet $\Sigma = \{0, 1, \dots, |\Sigma| - 1\}$. We treat each character as an integer in $\{0, 1, \dots, |\Sigma| - 1\}$ and interpret each m-length substring as a number in base $|\Sigma|$. For a string $S[1..m]$, define its *fingerprint* (or *hash value*) as

$$h(S[1..\ell]) = \sum_{j=1}^{\ell} S[j] \cdot |\Sigma|^{\ell - j} \pmod{p} \tag{1}$$

where p is a prime number chosen uniformly at random from all primes up to M (the value of M will be chosen later).

Problem 3.1. Define $t_i := h(T[i \dots i + m - 1])$. Derive a formula to compute t_{i+1} from t_i in $O(1)$ time.

Problem 3.2. Design a randomized algorithm using fingerprints that finds all occurrences of P in T as efficiently as possible.

Problem 3.3. For a random prime p chosen uniformly from all primes up to M, show that the probability of a false positive at any single position i where $T[i \dots i + m - 1] \neq P[1 \dots m]$ is at most

$$\frac{m \log_2 |\Sigma|}{\pi(M)},$$

where $\pi(M)$ is the number of primes up to M.

Problem 3.4. Show that for large enough M, the expected running time of the algorithm is $O(n + m)$. *Hint :* By the prime number theorem, $\pi(M) \geq M/(2 \ln M)$ for sufficiently large M.