Algorithmique

Pierre Aboulker, Paul Jeanmaire, Tatiana Starikovskaya

12 octobre 2023

Table des matières

Première partie

Lecture 1 - 28/09

Table des matières

1 Organisation

Mail Tatiana : `starikovskaya@di.ens.fr` Homeworks are 30% of the final grade, final (theory from lecture) Textbooks :

— *Introduction to Algorithms* - Cormen, Leiserson, Riverst, Stein

— *Algorithms on strings, trees, and sequences* - Gusfield

— *Approximation Algorithms* - Vazirani

— *Parametrized Algorithms* - Cygan, Fomin, Kowalik, Lokshtanov, Marx, Pilipczuk, Pilipczuk, Saurabh

2 Introduction

Algorithm take Inputs and give an output.

Open Problem 1 (Mersenne Prime). *Find a new prime of form* $2^n - 1$

Algorithms do not depend on the language. Algorithms should be simple, fast to write and efficient. Word RAM model : Two Parts : one with a constant number of registers of w bits with direct access, and one with any number of registers, only with indirect access (pointers). Allows for elementary operations : basic arithmetic and bitwise operations on registers, conditionals, goto, copying registers, halt and malloc. To index the memory storing input of size n with n words, we need register length to verify $w \geq \log n$ Algorithms can always be rewritten using only elementary operations. Complexity :

— $Space(n)$ is the maximum number of memory words used for input of size n

— $Time(n)$ is the maximum number of *elementary* operations used for input of size n

Complexity Notations :

— $f \in \mathcal{O}(g)$ if $\exists n_0 \in \mathbb{N}, c \in \mathbb{R}_+,\ f(n) \leq c \cdot g(n),\ \forall n \geq n_0$

— $f \in \Omega(g)$ if $\exists n_0 \in \mathbb{N}, c \in \mathbb{R}_+,\ f(n) \geq c \cdot g(n),\ \forall n \geq n_0$

— $f \in \Theta(g)$ if $\exists n_0 \in \mathbb{N}, c_1, c_2 \in \mathbb{R}_+,\ c_1 \cdot g(n) \leq f(n) \leq c_2 \cdot g(n),\ \forall n \geq n_0$

3 Data Structures

3.1 Introduction

Way to store elements of a data base that is created to answer frequently asked queries using pre-processing. We care about space used, construction, query and update time. Can be viewed as an algorithm, which analysed on basics. Containers are basic Data Structures, maintaining the following operations :

1. Random Access : given i, access e_i
2. Access first/last element
3. Insert an element anywher
4. Delete any element

3.2 Array

An array is a pre-allocated contiguous memory area of a *fixed* size. It has random access in $\mathcal{O}(1)$, but doesn't allow insertion nor deletion.

Linear Search : given an integer x return 1 if $e_i = x$ else 0.

Algorithme 1 Linear Search in an Array.
Complexity : Time = $\mathcal{O}(n)$ | Space = $\mathcal{O}(n)$

```
**Input**$x$
```

3.3 Doubly Linked List

Memory area that does not have to be contiguous and consists of registers containing a value and two pointers to the previous and next elements. It has random access in $\mathcal{O}(n)$, access/insertion/deletion at head/tail in $\mathcal{O}(1)$.

Algorithme 2 Insertion in a Doubly Linked List
Complexity : $\mathcal{O}(1)$

```
Input $L, x$
$x.next \leftarrow L.head$
if $L.head \neq NIL$ then
    $L.head.prev \leftarrow x$
end if
$L.head \leftarrow x$
$x.prev = Nil$
```

3.4 Stack and Queue

Stack : Last-In-First-Out data structure, abstract data structure. Access/insertion/deletion to top in $\mathcal{O}(1)$.

Open Problem 2 (Optimum Stack Generation)**.** *Given a finite alphabet Σ and $X \in \Sigma^n$. Find a shortest sequence of stack operations push, pop, emit that prints out X. You must start and finish with an empty stack. Current best solution is in $\tilde{\mathcal{O}}(n^{2.8603})$.*

Queue : First-In-First-Out abstract data structure. Access to front, back in $\mathcal{O}(1)$, deletion and insertion at front and back in $\mathcal{O}(1)$.

4 Approaches to algorithm design

Solve smalle sub-problems to solve a large one.

4.1 Dynamic Programming

Break the problem into many closely related sub-problems, memorize the result of the sub-problems to avoid repeated computation

Examples :

Levenshtein Distance between two strings can be computed in $\mathcal{O}(mn)$ instead of exponential time. Based on `https://arxiv.org/pdf/1412.0348.pdf`, this is the best one can do. RNA folding : retrieving the 3D shape of RNA based on their representation as strings. Currently, we know it is possible to find $\mathcal{O}(n^3)$, in $\tilde{\mathcal{O}}(n^{2.8606})$ and if *SETH* is true, there is no $\mathcal{O}(n^{\omega-\varepsilon})$. We know $\omega \in [2, 2.3703)$

Open Problem 3. *Is there a better Complexity for RNA folding ? What is the true value of ω ?*

Knapsack problem : An optimization problem with bruteforce complexity $\mathcal{O}(2^n)$.

Algorithme 3 Recursive Fibonacci Numbers
Complexity : Exponential

```
RFibo($n$) :
**Input**$n$
**if** $n \leq 1$ **then**
    **return** $n$
**end if**
**return** RFibo($n-1$) + RFibo($n-2$)
```

Algorithme 4 Dynamic Programming Fibonacci Numbers
Time = $\mathcal{O}(n)$ | Space = $\mathcal{O}(n)$

```
**Input**$n$
$Tab \leftarrow zeros(n)$                    ▷ $zeros(n)$ returns a $n$-array of zeros.
$Tab[0] \leftarrow 0$
$Tab[1] \leftarrow 1$
**for** $i \leftarrow 2$ to $n$ **do**
    $Tab[i] = Tab[i-1] + Tab[i-2]$
**end for**
**return** Tab[n]
```

Algorithme 5 Knapsack : Dynamic Programming
Time = $\mathcal{O}(nW)$ | Space = $\mathcal{O}(nW)$

```
**Input** $W, w, v$                          ▷ Capacity, weight and values vectors.
$KP = zeros(n, W)$
**for** $i \leftarrow 0$ to $n$ **do**
    $KP[i, 0] = 0$
**end for**
**for** $w \leftarrow 0$ to $W$ **do**
    $KP[0, w] = 0$
**end for**
**for** $i \leftarrow 0$ to $n$ **do**
    **for** $w \leftarrow 0$ to $W$ **do**
        **if** $w < w_i$ **then**
            $KP[i, w] \leftarrow KP[i-1, w]$
        **else**
            $KP[i, w] = \max \begin{cases} KP[i-1, w] \\ KP[i-1, w-w_i] + v_i \end{cases}.$
        **end if**
    **end for**
**end for**
**return** $KP[n, W]$
```

4.2 Greedy Techniques

Problems solvable with the greedy technique form a subset of those solvable with DP. Problems must have the optimal substrcture property. Principle : choosing the best at the moment.

Example : The Fractional Knapsack Problem
Algorithm : Iteratively select the greatest value-per-weight ratio.

Théorème 4.2.1. *This algorithm returns the best solution, in time $\mathcal{O}(n \log n)$*

By contradiction. Suppose we have $\frac{v_1}{w_1} \geq \ldots \geq \frac{v_n}{w_n}$. Let $ALG = p = (p_1, \ldots, p_n)$ be the output by the algorithm and $OPT = q = (q_1, \ldots, q_n)$ be optimal.
Assume that $OPT \neq ALG$, let i be the smallesst index such $p_i \neq q_i$. There is $p_i > q_i$ by construct. Thus, there exists $j > i$ such that $p_j < q_j$. We set $q' = (q_1^{'}, \ldots, q_n^{'}) = (q_1, \ldots, q_{i-1}, q_i + \varepsilon, q_{i+1}, \ldots, q_j - \varepsilon\frac{w_i}{w_j}, \ldots, q_n)$
$q^{'}$ is a feasible solution : $\sum_{i=1}^{n} q_i^{'} \cdot w_i = \sum_{i=1}^{n} q_i w_i \leq W$
Yet, $\sum_{i=1}^{n} q_i^{'} \cdot v_i > \sum_{i=1}^{n} q_i \cdot v_i$, ce qui contredit la ■

Deuxième partie

TD 1 - 29/09

Table des matières

5 Mathematical Complexity

5.1 Exercice 1

5.1.1 Question 1

Non : prendre $f = 1$ et $g = exp$.

5.2 Question 2

Non, si g = h = f.

5.2.1 Question 3

Non : Si on a $f = n, g = n^2 \in \Omega(f(n)), h = f \in \Theta(f(n))$ alors $g + h \neq \mathcal{O}(f(n))$

5.3 Exercice 2

On rappelle la formule de Stirling :

$$n! \sim \left(\frac{n^n}{e^n}\right)\sqrt{2\pi n}$$

Immédiatement, on en déduit la première relation.
On a par ailleurs la seconde égalité en passant au logarithme, fonction continue en $+\infty$

5.4 Exercice 3

On rappelle les formules suivantes :

$$\begin{cases} (n+a)^b & = n^b(1+\frac{a}{n})^b \\ (1+\frac{a}{n})^b & = 1 + b\frac{a}{n} + o(\frac{a}{n}) \in [1; 1+ba] \end{cases}$$

Immédiatement, on a la relation souhaité.

6 Data Structures

6.1 Exercice 4

Il suffit de diviser l'array en deux sous arrays de taille $n/2$, une array commençant en $i = 0$, une commençant en $j = -1$ et on stocke les deux indices de fin de la pile courant.

6.2 Exercice 5

6.2.1 Question 1

On définit un algorithme de *reverse* de list en temps linéaire en ajoutant tous les éléments dans une pile puis en dépilant dans une liste. On effectue bien $2n = \mathcal{O}(n)$ opérations.
Il suffit alors de comparer les deux listes en temps linéaire.

6.2.2 Question 2

Pour une liste vide, ou d'un seul élément on renvoie True. On reverse en place la première moitié de la liste et on la compare à la seconde et normalement ça marche.

6.3 Exercice 6

On utilise deux piles : On push dans la première, et on pop de la seconde. Lorsque la seconde pile est vide, on pop de la première et on push dans la seconde, ce qui permet bien de former une pile.

6.4 Exercice 7

6.4.1 Question 1

On utilise les arrays standards et lorsqu'on dépasse la capacité, on double le nombre de cases, qu'on initialise à -1, en stockant l'indice du dernier élément de la liste. On a alors toujours une complexité en espace en $\mathcal{O}(n)$ puisqu'on a toujours au plus $2n$ cases.

6.4.2 Question 2

On effectue la suite suivante d'instructions, pour $n \in \mathbb{N}$:

1. On ajoute $2n$ éléments
2. On retire $n+1$ éléments
3. On ajoute 1 élément
4. On recommence en modifiant n

6.4.3 Question 3

Il suffit alors d'attendre de passer en dessous de la barre de 25% du tableau rempli. On a bien tout de même une complexité en $\mathcal{O}(n)$.

7 Greedy Algorithms

7.1 Exercice 8

7.1.1 Question 1

Algorithme 6 Greedy Algorithm for Scheduling Problem

```
Input $a$                                    ▷ Vecteur de tuples correspondant aux activités
$E \leftarrow ListeVide()$
$Sort(a, (fun : x, y \mapsto x[1] \leq y[1]))$    ▷ On trie les activités par date de fin croissante
$s \leftarrow PileVide()$
$Push(a, s)$                                 ▷ On ajoute une à une les activités de $a$ dans une pile.
while ( do$s$)
    $ac \leftarrow Pop(s)$
    if ( then$ac$ est compatible)
        $Ajouter(E, ac)$
    end if
end while
return $E$
```

Correction. On introduit une solution optimale, la plus proche possible de l'algo. ■

7.1.2 Question 2

On prend $T1 = [1, 2], T2 = [3, 4], T3 = [1.5, 2.5]$

7.1.3 Question 3

Bon on fait de la Programmation Dynamique. Relation de récurrence $\forall i DP(i)$ est le max des poids sur $\{T_1, \ldots, T_i\}$

$$\begin{cases} DP(0) & = 0 \\ DP(i+1) & = \max(DP(i), w_{i+1} + DP(p(i+1))) \text{ où } p(i) \text{ est le dernier indice de la dernière tâche compatible} \end{cases}$$

Troisième partie
Lecture 2 - 5/10

Table des matières

8 Divide and Conquer

Divide a problem into smaller ones to solve those, then combine the solutions to get a solution of the bigger problem.
Example : *Merge Sort* : Its complexity verifies $T(n) = T(\lceil n/2 \rceil) + T(\lfloor n/2 \rfloor) + \mathcal{O}(n)$. From that we will derive that $T(n) = \mathcal{O}(n \log n)$

9 Analysis of Recursive Algorithms

We have recurrences we want to solve. We have three methods :

9.1 Substitution Method

The method :

1. Guess the asymptotic of $T(n)$
2. Show the answer via induction

For *Merge Sort* : we guess $T(n) \leq c \cdot n \log_2 n, \forall n \geq 2$. We choose c that verifies that property until $n = 6$.
Substituting in the recurrence equation :

$$T(n) \leq cn \log_2 \frac{n}{2} + c \log_2 \frac{n}{2} + c\frac{n+2}{2} + a \cdot n = c\dot{n} \log_2 n + a \cdot n + c \cdot \log_2 n - c\frac{n}{2}$$

If we then choose c so that it is bigger than $20a$ we get :

$$T(n) \leq cn \log_2 n + a \cdot n - c \cdot n/20 \leq cn \log_2 n$$

9.2 Recursion-tree Method

1. Simplify the equation :
   — Delete floors and ceils
   — Suppose n is of a good form
2. Draw a tree, rooted with the added term and the recursive calls
3. Start again, and recursively fill the tree

We get a tree of depth $\log_k n$ if n is divided by k in recursive calls. We can now sum the values of the nodes, to get an approximation, and start verifying.

10 Master Theorem

10.1 The Theorem

Théorème 10.1.1 (Master Theorem). *If we have recurrence equation $T(n) = aT(n/b) + f(n)$ where $a \geq 1, b > 1$ are integers, $f(n)$ is asymptotically positive. Let $r = \log_b a$, we have :*

1. *If $f(n) = \mathcal{O}(n^{r-\varepsilon})$ for some $\varepsilon > 0$, then $T(n) = \Theta(n^r)$*
2. *If $f(n) = \Theta(n^r)$ then $T(n) = \Theta(n^r \log n)$*
3. *If $f(n) = \Omega(n^{r+\varepsilon})$ for some $\varepsilon > 0$, and $af(n/b) \leq cf(n)$ for some constant $c < 1$ and all sufficiently large n (regularity condition) then $T(n) = \Theta(f(n))$.*

Remarque 10.1.1.1. *The Master Theorem 10.1.1 does not cover all possible cases for $f(n)$. Example : $f(h) = h^r / \log h$*

Remarque 10.1.1.2. *The Master Theorem 10.1.1 is sometimes called Théorème sur les Récurrences de Partition*

11 The Proof

Plan :
— Analyse the recurrence as if T is defined over reals (continuous version)
— Prove the discrete version

11.1 Continuous Master Theorem

Démonstration.

Lemme 11.1.1. *Define* $T(n) = \begin{cases} \Theta(1) & \text{if } n \leq \hat{n} \\ aT(n/b) + f(n) & \text{if } n > \hat{n} \end{cases}$ *Then*

$$T(n) = \Theta(n^r) + \sum_{k=0}^{\lceil \log_b(n/\hat{n}) \rceil - 1} a^k f(n/b^k)$$

Démonstration. In the Recursion-Tree, stopped when the argument of T is smaller than $\hat{n}$ which is when depth is $\lceil \log_b(n/\hat{n}) \rceil - 1$, we get :

$$\begin{aligned} T(n) &\leq \sum_{k=0}^{\lceil \log_b(n/\hat{n}) \rceil - 1} a^k f(n/b^k) + \Theta(a^{log_b(n/\hat{n})}) \\ &= \sum_{k=0}^{\lceil \log_b(n/\hat{n}) \rceil - 1} a^k f(n/b^k) + \Theta(a^{\log_b(n)}) \\ &= \sum_{k=0}^{\lceil \log_b(n/\hat{n}) \rceil - 1} a^k f(n/b^k) + \Theta(n^{\log_b(a)}) \end{aligned}$$

■

Back to the proof :

Lemme 11.1.2. *Define* $g(n) = \Theta(n^r) + \sum_{k=0}^{q} a^k f(n/b^k)$ *Then :*

1. *If* $f(n) = \mathcal{O}(n^{r-\varepsilon})$ *then* $g(n) = \Theta(n^r)$
2. *If* $f(n) = \Theta(n^r)$ *then* $g(n) = \Theta(n^r \log n)$
3. *If* $f(n) = \Omega(n^{r+\varepsilon})$ *and we have the regularity condition then* $g = \Theta(f)$

Démonstration. 1. Case 1 :

$$g(n) \begin{aligned} &= \Theta(n^r) + \sum_{k=0}^{q} a^k f(n/b^k) \\ &= \Theta(n^r) + \mathcal{O}\left(\sum_{k=0}^{q} a^k (n/b^k)^{r-\varepsilon}\right) \end{aligned}$$

However :

$$\begin{aligned} \sum_{k=0}^{q} a^k (n/b^k)^{r-\varepsilon} &= n^{r-\varepsilon} \sum_{k=0}^{q} (ab^{\varepsilon}/b^r)^k \\ &= n^{r-\varepsilon} \sum_{k=0}^{\lceil \log_b(n/\hat{n}) \rceil - 1} (b^{\varepsilon})^k = \mathcal{O}(n^{r-\varepsilon}(n/\hat{n})^{epsilon}) \end{aligned}$$

Thus : $g(n) = \Theta(n^r)$

2. Case 2 : We have :

$$\begin{aligned} g(n) &= \Theta(n^r) + \sum_{k=0}^{q} a^k f(n/b^k) \\ &= \Theta(n^r) + \Theta\left(\sum_{k=0}^{q} a^k (n/b^k)^r\right) \end{aligned}$$

However :

$$\begin{aligned} \sum_{k=0}^{q} a^k (n/b^k)^r &= n^r \sum_{k=0}^{q} (a/b^r)^k \\ &= n^r \sum_{k=0}^{\lceil \log_b(n/\hat{n}) \rceil - 1} 1 = n^r \Theta(\log_b n/\hat{n}) \end{aligned}$$

3. Case 3 : By induction on k : $a^k f(n/b^k) \leq c^k f(n)$. Thus :

$$\sum_{k=0}^{q} a^k f(n/b^k) \leq \sum_{k=0}^{q} c^k f(n) = f(n) \sum_{k=0}^{q} c^k = \Theta(f(n))$$

■

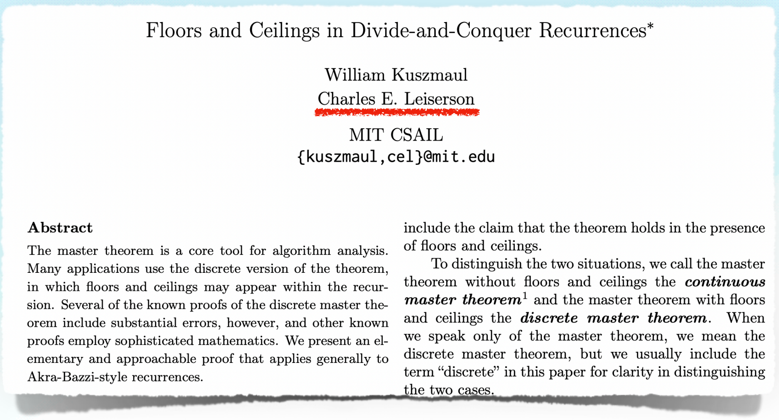
We thus have proved the continuous Master Theorem. ■

11.2 Discrete Master Theorem

We have now showed the continuous Master Theorem, following William Kuszmaul, Charles E. Leiserson, *Floors and Ceilings in Divide-and-Conquer Recurrences*, Symposium on Simplicity in Algorithms 2021.

Démonstration. See slides below

Why not to follow CLRS textbook?



Floors and Ceilings in Divide-and-Conquer Recurrences*

William Kuszmaul
Charles E. Leiserson

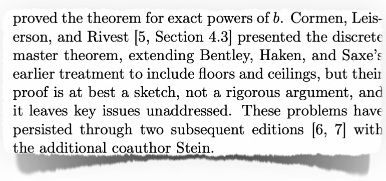
MIT CSAIL
{kuszmaul,cel}@mit.edu

Abstract

The master theorem is a core tool for algorithm analysis. Many applications use the discrete version of the theorem, in which floors and ceilings may appear within the recursion. Several of the known proofs of the discrete master theorem include substantial errors, however, and other known proofs employ sophisticated mathematics. We present an elementary and approachable proof that applies generally to Akra-Bazzi-style recurrences.

include the claim that the theorem holds in the presence of floors and ceilings.

To distinguish the two situations, we call the master theorem without floors and ceilings the ***continuous master theorem***[1] and the master theorem with floors and ceilings the ***discrete master theorem***. When we speak only of the master theorem, we mean the discrete master theorem, but we usually include the term "discrete" in this paper for clarity in distinguishing the two cases.



proved the theorem for exact powers of b. Cormen, Leiserson, and Rivest [5, Section 4.3] presented the discrete master theorem, extending Bentley, Haken, and Saxe's earlier treatment to include floors and ceilings, but their proof is at best a sketch, not a rigorous argument, and it leaves key issues unaddressed. These problems have persisted through two subsequent editions [6, 7] with the additional coauthor Stein.

Why not to follow CLRS textbook?

- Aho, Hopcroft, Ullman offered one of the first treatments of divide-and-conquer recurrences, giving three cases for recurrences of the form T(n) = aT (n/b) + cn (1974)
- Bentley, Haken, and Saxe introduced the master theorem in modern form, but proved it for $n = b^k$ only (1980)
- CLRS extended the proof to the discrete version, but gave only a sketch of the proof (1990)
- Akra and Bazzi considered $T(n) = \sum_{i=1}^{t} a_i T(n/b_i) + f(n)$ (1998)
- Leighton simplifies the proof of Akra and Bazzi and extends is to the discrete version (1996)
- Campbell spots several flows in the proof of Leighton and devotes more than 300 pages to carefully correct the issues (2020)
- More generalizations by Drmota and Szpankowski(2013), Roura (2001), Yap (2011)

Definitions

Discrete recurrences

$$T(n) = f(n) + \sum_{i \in S} a_i T(\lfloor n/b_i \rfloor) + \sum_{i \notin S} a_i T(\lceil n/b_i \rceil)$$

$a_i \in \mathbb{R}^+, b_i \in \mathbb{R}^+, n \geq \hat{n}$

For $1 \leq n < \hat{n}$, there exist c_1, c_2: $c_1 \leq T(n) \leq c_2$

Polynomial-growth condition

$\exists \hat{n} > 0$ such that $\forall \Phi \geq 1 \, \exists d > 1 : d^{-1} f(n) \leq f(\varphi n) \leq d f(n)$

for all $1 \leq \varphi \leq \Phi$ and $n \geq \hat{n}$

6 technical slides ahead!

KEEP CALM AND CARRY ON

Lemma 1. For $\beta > 1, n \in \mathbb{N}$ let $L = \Pi_{i=1}^{n}(1 - \frac{1}{\beta^i + 1})^2$, $U = \Pi_{i=1}^{n}(1 + \frac{1}{\beta^i - 1})^2$

We have $L = \Omega(1)$ and $U = O(1)$.

Proof.

$$\beta > 1 \Rightarrow \frac{1}{\beta^i} < \frac{1}{\beta^i - 1} \Rightarrow 1/L = \Pi_{1=1}^{n}(1 + \frac{1}{\beta^i})^2 < \Pi_{i=1}^{n}(1 + \frac{1}{\beta^i - 1})^2 = U$$

$$U = \Pi_{i=1}^{n}(1 + \frac{1}{\beta^i - 1})^2 \leq \Pi_{i=1}^{\infty}(1 + \frac{1}{\beta^i - 1})^2 \leq \Pi_{i=1}^{\infty}(e^{1/(\beta^i - 1)})^2 =$$

(Here we use $1 + 1/x \leq e^{1/x}$ for $x \neq 0$)

$$= \exp(\sum_{i=1}^{\infty} \frac{2}{\beta^i - 1}) \leq \exp(\sum_{i=1}^{\infty} \frac{4}{\beta^i}) + O(1) = O(1)$$

Lemma 2. Let $\beta > 1; \beta_i \geq \beta, 1 \leq i \leq k; B := \Pi_{i=1}^{k}\beta_i$

There exists $c = c(\beta) > 0$ such that for all $n_1, n_2, \ldots, n_k$ where $n_i > \max(\beta, 1 + 1/(\sqrt{\beta} - 1))$ and $\lfloor n_{i-1}/\beta_i \rfloor \leq n_i \leq \lceil n_{i-1}/\beta_i \rceil$, we have $c^{-1/4}(n_0/B) \leq n_k \leq c^{1/4}(n_0/B)$.

Proof. Let $\rho_i := \dfrac{n_i}{n_{i-1}/\beta_i}$.

$$(n_0/B)\Pi_{i=1}^{k}\rho_i = \frac{n_0\Pi_{i=1}^{k}\rho_i}{\Pi_{i=1}^{k}\beta_i} = n_0\Pi_{i=1}^{k}\frac{\rho_i}{\beta_i} = n_0\Pi_{i=1}^{k}\frac{n_i}{n_{i-1}} = n_k$$

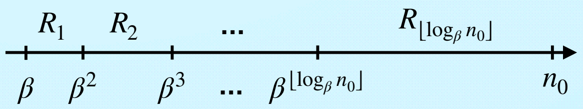
It is enough to show that $c^{-1/4} \leq \Pi_{i=1}^{k}\rho_i \leq c^{1/4}$ for some $c = c(\beta)$

$$n_{i-1}/\beta_i - 1 \leq \lfloor n_{i-1}/\beta_i \rfloor \leq n_i \leq \lceil n_{i-1}/\beta_i \rceil \leq n_{i-1}/\beta_i + 1 \Rightarrow$$

$$n_i - 1 \leq n_{i-1}/\beta_i \leq n_i + 1 \Rightarrow \boxed{\underbrace{\frac{n_i}{n_i + 1}}_{1 - \frac{1}{n_i + 1}} \leq \rho_i \leq \underbrace{\frac{n_i}{n_i - 1}}_{1 + \frac{1}{n_i - 1}} \; (*)}$$

Proof of Lemma 2 (continued).

$$\boxed{\underbrace{\frac{n_i}{n_i+1}}_{1-\frac{1}{n_i+1}} \le \rho_i \le \underbrace{\frac{n_i}{n_i-1}}_{1+\frac{1}{n_i-1}} \; (*)}$$


R_1
R_2
...
$R_{\lfloor \log_\beta n_0 \rfloor}$
β
β^2
β^3
...
$\beta^{\lfloor \log_\beta n_0 \rfloor}$
n_0

From $(*)$: $\rho_i \le 1 + \dfrac{1}{n_i - 1} \le 1 + \dfrac{1}{1/(\sqrt{\beta}-1)} = \sqrt{\beta}$

$n_{i+2} = \dfrac{n_i \rho_{i+1} \rho_{i+2}}{\beta_{i+1}\beta_{i+2}} \le n_i/\beta \Rightarrow$ every range R_j contains at most two n_i's

From $(*)$ again: $n_i \in R_j \Rightarrow 1 - \dfrac{1}{\beta^j + 1} \le \rho_i \le 1 + \dfrac{1}{\beta^j - 1} (n_i > \beta^j)$

Therefore, $\Pi_{i=1}^k \rho_i = \Pi_{j=1}^{\lfloor \log_\beta n_0 \rfloor} (\Pi_{n_i \in R_j} \rho_i) \le \Pi_{j=1}^{\lfloor \log_\beta n_0 \rfloor} (1 + \dfrac{1}{\beta^j - 1})^2 \le c^{1/4}$ **(Lemma 1)**

$$\Pi_{i=1}^k \rho_i \ge \Pi_{j=1}^{\lfloor \log_\beta n_0 \rfloor} (1 - \frac{1}{\beta^j + 1})^2 \ge c^{-1/4} \textbf{ (Lemma 1)}$$

Lemma 3. $\beta_{\min}, \beta_{\max} > 1$. Assume that for all $1 \le i \le k$, $\beta_{\min} \le \beta_i \le \beta_{\max}$, and let $B = \Pi_i \beta_i$.

There exists $c = c(\beta_{\min}, \beta_{\max})$ such that for any $n_1, n_2, \dots, n_k$ with $n_0 \ge cB$ and $\lfloor n_{i-1}/\beta_i \rfloor \le n_i \le \lceil n_{i-1}/\beta_i \rceil$, we have $c^{-1}(n_0/B) \le n_k \le c(n_0/B)$.

Proof.

Let $c = c(\beta_{\min})$ be the constant from Lemma 3. W.l.o.g. $\sqrt{c} > \max\{\dfrac{1}{\sqrt{\beta_{\min}} - 1} + 1, \beta_{\min}\}$ $(*)$ and $c^{1/4} > 2\beta_i$

If $n_j \ge \sqrt{c}$ for all j, then **Lemma 3** follows from **Lemma 2** and $(*)$. Let j be the smallest value such that $n_j \le \sqrt{c}$. We have $j \ge 1$ as $n_0 \ge cB \ge \sqrt{c}$.

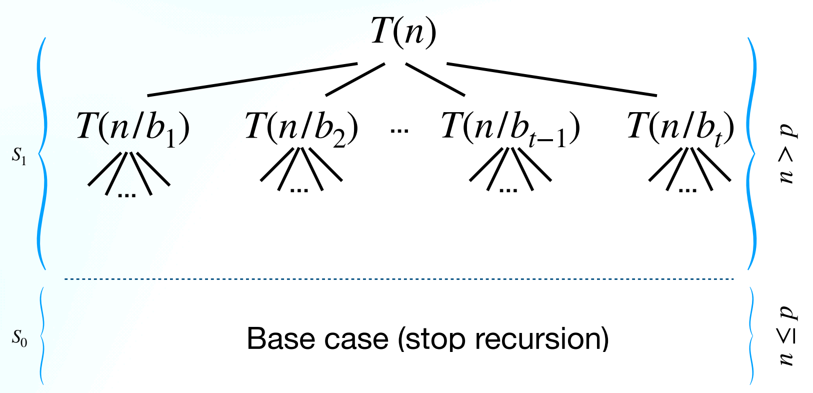
- If $j = 1$, then $n_{j-1} = n_0 \ge c^{-1/4}(n_0/B)$ (trivial).
- If $j > 1$, we apply **Lemma 2** to $\beta_1, \beta_2, \dots, \beta_{j-1}$ and $n_0, n_1, \dots, n_{j-1}$ and $\beta = \beta_{\min}$ (all conditions are satisfied) to obtain that $n_{j-1} \ge c^{-1/4}(n_0/B)$

In both cases, $n_{j-1} \ge c^{-1/4}(n_0/B) \ge c^{3/4}$. Therefore, $n_j \ge \lfloor \underbrace{n_{j-1}}_{\ge c^{1/4} > 2\beta_j} / \beta_j \rfloor \ge n_{j-1}/(2\beta_j) > n_{j-1}/c^{1/4} \ge \sqrt{c}$

Lemma 4. Let $a_1, a_2, \ldots, a_t > 0$ and $b_1, b_2, \ldots, b_t > 1$ be constants, $f(n) : \mathbb{R}^+ \to \mathbb{R}^+$ which satisfies the polynomial-growth condition. Consider $T(n) = f(n) + \sum_{i=1}^{t} a_i T(n/b_i)$ defined for $n \in \mathbb{R}^+$ ($*$). Assume that $T'(n)$ defined on $\mathbb{N}$ also satisfies ($*$), but each n/b_i is replaced with $\lfloor n/b_i \rfloor$ or $\lceil n/b_i \rceil$. Then $T'(n) = \Theta(T(n))$.

Proof.

Let c be the constant from Lemma 3 for $\beta_{\min} = \min_i b_i$ and $\beta_{\max} = \max_i b_i$. Let $\hat{n}$ be a sufficiently large constant. Define $p := \max\{\hat{n}, c \cdot \max_i b_i\}$. For $T(n)$, the base case is $n \le p$.



Subproblem q at level j:
$q = < q(1), q(2), \ldots, q(j) >$
$q(i) \in \{1,2,\ldots,t\}, |q| = j$
$q(i)$ - choice of the recursion brunch at level i
$n_q = n/\Pi_{i=1}^{|q|} b_{q(i)}$

Proof.

$$T(n) = \sum_{q \in S_1} f(n_q)\Pi_{i=1}^{|q|} a_{q(i)} + \sum_{q \in S_0} T(n_q)\Pi_{i=1}^{|q|} a_{q(i)} + f(n) = \sum_{q \in S_1} f(n_q)\Pi_{i=1}^{|q|} a_{q(i)} + \Theta(\sum_{q \in S_0} \Pi_{i=1}^{|q|} a_{q(i)}) + f(n)$$

When computing $T'(n)$ for a subproblem q:

$$\lfloor \frac{n'_{<q(1),q(2),\ldots,q(j-1)>}}{q(j)} \rfloor \le n'_q \le \lceil \frac{n'_{<q(1),q(2),\ldots,q(j-1)>}}{q(j)} \rceil$$

$$T'(n) = \sum_{q \in S_1} f(n'_q)\Pi_{i=1}^{|q|} a_{q(i)} + \sum_{q \in S_0} T'(n'_q)\Pi_{i=1}^{|q|} a_{q(i)} + f(n) \text{ (*)}$$

As $n_q > p$ for $q \in S_1$, $n_q > p/\max_i b_i \ge c$ for all $q \in S$. By Lemma 3 with $\beta_i = b_{q(i)}$, for all q we have $n'_q = \Theta(n_q)$. It follows that $\exists \Phi > 1$ such that $n'_q \in [\Phi^{-1} n_q, \Phi n_q]$. Therefore, $n'_q \ge n_q/\Phi \ge \hat{n}/\Phi$ and we can choose $\hat{n}$ so that ($*$) is defined correctly.

By the polynomial-growth condition, $f(n'_q) = \Theta(f(n_q))$ for all $q \in S$. For $q \in S_0$, $n'_q = \Theta(1)$ and therefore $T'(n'_q) = \Theta(1)$. It follows:

$$T'(n') = \sum_{q \in S_1} \Theta(f(n_q))\Pi_{i=1}^{|q|} a_{q(i)} + \Theta(\sum_{q \in S_0} \Pi_{i=1}^{|q|} a_{q(i)}) + f(n) = \Theta(T(n))$$

Proof.

$$T(n) = \sum_{q \in S_1} f(n_q)\Pi_{i=1}^{|q|} a_{q(i)} + \sum_{q \in S_0} T(n_q)\Pi_{i=1}^{|q|} a_{q(i)} + f(n) = \sum_{q \in S_1} f(n_q)\Pi_{i=1}^{|q|} a_{q(i)} + \Theta(\sum_{q \in S_0} \Pi_{i=1}^{|q|} a_{q(i)}) + f(n)$$

When computing $T'(n)$ for a subproblem q:

$$\lfloor \frac{n'_{<q(1),q(2),\ldots,q(j-1)>}}{q(j)} \rfloor \le n'_q \le \lceil \frac{n'_{<q(1),q(2),\ldots,q(j-1)>}}{q(j)} \rceil$$

$$T'(n) = \sum_{q \in S_1} f(n'_q)\Pi_{i=1}^{|q|} a_{q(i)} + \sum_{q \in S_0} T'(n'_q)\Pi_{i=1}^{|q|} a_{q(i)} + f(n) \; (*)$$

As $n_q > p$ for $q \in S_1$, $n_q > p/\max_i b_i \ge c$ for all $q \in S$. By Lemma 3 with $\beta_i = b_{q(i)}$, for all q we have $n'_q = \Theta(n_q)$, and hence $\exists \Phi > 1$ such that $n'_q \in [\Phi^{-1} n_q, \Phi n_q]$. Therefore, $n'_q \ge n_q/\Phi \ge \hat{n}/\Phi$ and we can choose $\hat{n}$ so that $(*)$ is defined correctly.

By the polynomial-growth condition, $f(n'_q) = \Theta(f(n_q))$ for all $q \in S$. For $q \in S_0$, $n'_q = \Theta(1)$ and therefore $T'(n'_q) = \Theta(1)$. It follows:

$$T'(n') = \sum_{q \in S_1} \Theta(f(n_q))\Pi_{i=1}^{|q|} a_{q(i)} + \Theta(\sum_{q \in S_0} \Pi_{i=1}^{|q|} a_{q(i)}) + f(n) = \Theta(T(n))$$

Discrete Master theorem

$T(n) = a_1 T(\lfloor n/b \rfloor) + a_2 T(\lceil n/b \rceil) + f(n)$, where
$a := a_1 + a_2 \ge 1$, $b > 1$, $f(n)$ - asymptotically positive.

Define $r := \log_b a$.

Case 1. If $f(n) = O(n^{r-\varepsilon})$ for some $\varepsilon > 0$, then $T(n) = \Theta(n^r)$.

Case 2. If $f(n) = \Theta(n^r)$, then $T(n) = \Theta(n^r \log n)$.

Case 3. If $f(n) = \Omega(n^{r+\varepsilon})$ for some $\varepsilon > 0$, and if $a_1 f(\lfloor n/b \rfloor) + a_2 f(\lceil n/b \rceil) \le cf(n)$ for some constant $c < 1$ and all sufficiently large n, then $T(n) = \Theta(f(n))$.

Discrete Master theorem

Case 1.

Fact. Replacing $f(n)$ with a function $f'(n)$ satisfying $f'(n) \leq f(n)$ (resp. $f'(n) \geq f(n)$) for all n in the domain of f does not increase (resp. decrease) $T(n)$.

Let $f(n) = O(n^c)$ for $c < \log_b a$. Then as a "bigger" function consider $f'(n) = r(n^c + 1)$ for r big enough. By Lemma 4 and the continuos Master theorem, $T(n) = O(n^{\log_b a})$.

As a "smaller" function, consider $f'(n) = 0$. By Lemma 4 and the continuos Master theorem, $T(n) = \Omega(n^{\log_b a})$.

Exercise. Both bigger and smaller functions satisfy the polynomial growth condition.

Case 2. Analogous.

Discrete Master theorem

Case 3.

$T(n) \geq f(n)$ and hence $T(n) = \Omega(f(n))$. It remains to show that $T(n) = O(f(n))$.

Regularity condition: $a_1 f(\lfloor n/b \rfloor) + a_2 f(\lceil n/b \rceil) \leq cf(n)$ for some $c < 1$ and all $n \geq p$.

For all $n < p$, there exists $s \geq 1$: $T(n) \leq sf(n)$. We show by induction that for all $n \in \mathbb{N}$, $T(n) \leq qf(n)$ for $q = s/(1-c)$.

- Base case: $n < p$ - by the choice of s
- Suppose that $n \geq p$ and the claim holds for all smaller n

$$T(n) = a_1 T(\lfloor n/b \rfloor) + a_2 T(\lceil n/b \rceil) + f(n) \leq a_1 q f(\lfloor n/b \rfloor) + a_2 q f(\lceil n/b \rceil) + f(n) \leq$$

$$\leq qcf(n) + f(n) = (\frac{sc}{1-c} + 1)f(n) = \frac{s - \overbrace{(1-c)s + 1 - c}^{\leq 0}}{1-c} f(n) \leq qf(n)$$

■

Remarque 11.2.0.1 (Remarks on the Proof). — *Lemmas 1 to 3 serve to show that the argument does not go too far when it is rounded up or down.*

— *Slide 36 Last Line :* $\frac{2}{\beta^i-1} < \frac{4}{\beta^i}$ *for* $i \geq i_0$. *Thus :* $\sum_{i=1}^{\infty} \frac{2}{\beta^i-1} < \sum_{i=1}^{\infty} \frac{4}{\beta^i} + \sum_{i=0}^{i_0} \frac{2}{\beta^i-1}$ *and that last sum is* $\mathcal{O}(1)$

— *Slide 37 Line 3 : The first inequalities comes from the Recursion-Tree, so that we can ensure the argument does not deviate to much, by the second inequalities.*

11.3 Use Cases

Using the Master Theorem we can show the complexity of many algorithms :

1. Merge Sort Complexity : $T(n) = T(\lceil n/2 \rceil) + T(\lfloor n/2 \rfloor) + \mathcal{O}(n) = \Theta(n \log n)$
2. Strassen's Algorithm for Matrix Multiplication : $T(n) = 7T(n/2) + \Theta(n^2) \Rightarrow T(n) = \mathcal{O}(n^{\log_2 7}) = \mathcal{O}(n^{2.8074})$

12 Fast Multiplication of Polynomials

The sum of two degree n polynomials can be done in $\mathcal{O}(n)$, Horner's rule for evaluation produces $\mathcal{O}(n)$ complexity. The naïve product can be done in $\mathcal{O}(n^2)$
Remembering Lagrange's Theorem on Polynomials (or Vandermonde's Determinant, or anything really), degree n polynomials are entirely represented by their point-value reprensentation over n distinct points (x_i, y_i). Then, by converting the coefficient reprensentation to point-value representation, then by point-wise multiplicating the polynomials, then by going back to the coefficient representation, we can have a better algorithm.

12.1 Point-Value Multiplication

It is easily done in $\mathcal{O}(n)$ if both polynomials are represented over the same axis.

12.2 Coefficient to Point-Value Conversion - Fast Fourier Transform

For $P = \sum_{i=0}^{n-1} a_i x^i$, we define :

$$\begin{aligned} P_{odd}(x) &= a_{n-1}x^{n/2-1} + a_{n-3}x^{n/2-3} + \ldots + a_1 x \\ P_{even}(x) &= a_{n-2}x^{n/2-2} + a_{n-4}x^{n/2-4} + \ldots + a_2 x^{2/2} + a_0 \end{aligned}$$

1. We have : $P = xP_{odd}(x^2) + P_{even}(x^2)$
2. We evaluate P_{odd}, P_{even} at $(\omega_n^i)^2$ recursively by the halving property.
3. We combine the result.

12.3 Point-Value to Coefficient Conversion - Inverse Fast Fourier Transform

Théorème 12.3.1. $V_n^{-1}[i,j] = \omega_n^{-ij}/n$

Quatrième partie
Hashing

Table des matières

Answering queries of appartenance on a set, maintaining a dictionary. Python dictionaries do not have upper bound guarantees, we shouldn't use them.
Suppose we have a set S over a universe U. We dnote by n the number of keys, m the size of the hash table

13 Naïve Array-based implementation

Here we assume that no two objects have equal keys. We store 1 in a bitvector of length $|U|$ at each position i such that i is in S.
This takes $\mathcal{O}(|U|)$ space, for $\mathcal{O}(1)$ search time, and $\mathcal{O}(1)$ modification time.

14 Chained Hash Tables

We give ourselves a function $h : U \to [1, m]$. We send the keys to their image by h in a table. However, since we have no guarantee that $|U| \leq m$, there might be collisions.
To deal with that, instead of storing in the table a boolean, we store a list of all the keys corresponding to $h(k)$.
Then we insert a key in constant time $\mathcal{O}(1)$, search a key in time $\mathcal{O}(|h[key]|)$ and delete a key in time $\mathcal{O}(|h[key]|)$ since in the worst case the key is at the end of the list.

Further analysis leads to the following theorem.

Théorème 14.0.1 (Simple Uniform Hashing Assumption)**.** *Assuming SUHA : « h equally distributes the keys into the table slots », and assuming $h(x)$ can be computed in $\mathcal{O}(1)$, $E\left[T_{search}(n)\right] = \mathcal{O}\left(1 + \frac{n}{m}\right)$, and same for deletion time.*
Formally, SUHA is :

$$\forall y \in [1, |T|] \mathbb{P}\left(h(x) = y\right) = \frac{1}{|T|}$$
$$\forall y_1, y_2 \in [1, |T|]^2 \mathbb{P}\left(h(x_1) = y_1,\ h(x_2) = y_2\right) = \frac{1}{|t|^2}$$

Démonstration.

1. Unsuccessful Search : Suppose that $k_0, \ldots, k_{n-1}$ are keys in the dictionary and we perform an unsuccessful search for a key k. The number of comparisons is : $\sum_{i=0}^{n-1} \mathbb{1}_{h(k)=h(k_i)}$. Then, the expected time is : $\mathbb{E}[T_{search}] = \frac{n}{|T|}$ by SUHA.
2. Successful Search : Suppose keys were introduced in order $k_0, \ldots, k_{n-1}$. k_i appears before any of $k_0, \ldots, k_{i-1}$ and after any of $k_{i+1}, \ldots, k_{n-1}$ that are in the same linked list. Then, to search for k_i, we need $\sum_{j=i+1}^{n-1} \mathbb{1}_{h(k_j)=h(k_i)}$. Under SUHA, the expectation of each of these variables is $\frac{1}{|T|}$. Then, the average expected search time is : $\frac{1}{n}\sum_{i=0}^{n-1} = 1 + \frac{1}{n}\sum_{i=0}^{n-1} \frac{(n-1-i)}{|T|} = \mathcal{O}(1 + \frac{n}{|T|})$

This concludes the proof of the theorem. ■

Good hash functions are functions that distribute the keys evenly. Yet, we do not know what the keys are, and thus will need various heuristics to answer this question. At least, without loss of generality, we can assume that keys are integers.

14.1 Heuristic Hash Functions

— Division Method : $h(k) = k \mod m$. It is better to choose m to be a prime number, and avoid $m = 2^p$.

— Multiplication Method : $h(k) = \lfloor m \{k \cdot A\} \rfloor$. $A \in (0, 1)$ and $m = 2^p$.

Yet, fixing the function can allow anyone to construct a probability distribution for which the function will be "bad".

14.2 Universal Family of Hash Functions

$H = \{h : U \to [0, |T| - 1]\}$ is Universal if :

$$\forall k_1 \neq k_2 \in U,\ |\{h \in H \mid h(k_1) = h(k_2)\}| \leq \frac{|H|}{m}$$

Théorème 14.2.1. *If h is a hash function chosen uniformly at random from a universal family of hash functions. Suppose that $h(k)$ can be computed in constant time and there are at most n keys. Then the expected search time is $\mathcal{O}(1 + \frac{n}{|T|})$*

Démonstration. Same as the case when h satisfies SUHA. Observe that the probability comes this time from choosing the function. ■

Théorème 14.2.2. *Let $p \in \mathcal{P}$ such that $U \subseteq [0, p-1]$. Then $H = \{h_{a,b}(k) = ((ak + b) \mod p) \mod |T| \mid a \in \mathbb{Z}_p^*,\ b$ is a universal family.*

Démonstration. Let $k_1 \neq k_2 \in U$. Let $l_i = (ak_i + b) \mod p$. We have $l_1 \neq l_2$ and $a = ((l_1 - l_2)((k_1 - k_2)^{-1} \mod p) \mod p)$ and $b = (l_1 - ak_1) \mod p$. There is then one-to-one mapping between (a, b) and (l_1, l_2). The number of $h \in H$ such that $h(k_1) = h(k_2)$ is :

$$|\{(l_1, l_2) \mid l_1 \neq l_2 \in \mathbb{Z}_p,\ l_1 = l_2 \mod m\}| \leq \frac{p(p-1)}{|T|} \leq \frac{|H|}{|T|}$$

■

15 Open Addressing

Elements are stored in the table. To insert, we probe the hash table until we find x or an empty slot. If we find an empty slot, insert x here. To define which slots to probe, we use a hash function that depends on the key and the probe number. To search, we probe the hash table until we either find x (return YES) or an empty slot (return NO). In the analysis, we will assume h to be uniform.

Théorème 15.0.1 (Analysis). *Given an open-address hash-table with load factor $\alpha = \frac{n}{|T|} < 1$, the expected number of probes in an unsuccessfulsearch is at most $\frac{1}{1-\alpha}$, assuming uniform hashing.*

Démonstration. An unsuccessful search on x means that every probed slot except the last one is occupied and does not contain x, and the last one is empty. We define A_i the event « The i-th probe occurs and is occupied. ». By Bayes's Theorem, we must estimate :

$$\mathbb{P}[\# \text{ of probes} \geq i] = \prod_{k=1}^{i-1} \mathbb{P}[A_k \mid \bigcap_{j=1}^{k-1} A_j]$$

But we have : $\mathbb{P}[A_j \mid A_1 \cap \ldots \cap A_{j-1}] = \frac{n-j+2}{|T|-j+2}$ So we have : $\mathbb{P}[\# \text{ of probes} \geq i] \leq \frac{n}{|T|}^{i-1} = \alpha^{i-1}$
Then, the expected number of probes is :

$$\sum_{i=1}^{+\infty} \mathbb{P}[\# \text{ of probes} \geq i] \leq \sum_{i=1}^{+\infty} \alpha^{i-1} = \frac{1}{1-\alpha}$$

■

Corollaire 15.0.1.1. *The expected number of probes during insertion is at most $\frac{1}{1-\alpha}$.*

Démonstration. If we insert x we first ran an unsuccessful search for it ■

Théorème 15.0.2. *The expected number of probes during a Successful search is at most $\frac{1}{\alpha}\ln\frac{1}{1-\alpha}$.*

Démonstration. A successful search for x probes the same sequence of slots as insertion.
If x is the i-th element inserted into the table, insertion probes less than $\frac{1}{1-\frac{i}{|T|}}$ slots in expectation. Therefore, the expected time of a successful search is at most :

$$\frac{1}{n}\sum_{i=0}^{n-1}\frac{|T|}{|T|-i} = \frac{|T|}{n}\sum_{i=0}^{n-1}\frac{1}{|T|-i} = \sum_{i=0}^{|T|}\frac{1}{i} - \sum_{i=0}^{|T|-n}\frac{1}{i} \leq \frac{|T|}{n}\ln\frac{|T|}{|T|-n} = \frac{1}{\alpha}\ln\frac{1}{1-\alpha}$$

■

This is hard to implement, so we will use heuristics.

15.1 Heuristic hash functions

Let $h^{'}, h^{''}$ be two auxiliary hash functions.

- Linear Probing : $h(k,i) = (h^{'}(k) + i) \mod |T|$ This is easy to implement but it suffers from clustering.
- Quadratic Probing : $h(k,i) = (h^{'}(k) + c_1 i + c_2 i^2) \mod |T|$ We must choose the constants c_1 and c_2 carefully, and this still suffers from clustering.
- Double Hashing : $h(k,i) = (h^{'}(k) + ih^{''}(k)) \mod |T|$ To use the whole table, $h^{''}(k)$ must be relatively prime to m, e.g. $h^{''}(k)$ is always odd, $m = 2^i$.

16 Cuckoo Hashing

Hashing scheme with search time constant in the worst case, as it maintains a hash function without collisions to achieve perfect hashing. This is possible if the set of keys is static. Assume that we have two hash functions h_1, h_2 that satisfy SUHA. We store x in either $T[h_1(x)]$ or $T[h_2(x)]$. Search for x is done in constant time.

```
function INSERT(x)
    if $x = T[h_1(x)]$ or $x = T[h_2(x)]$ then return
    end if
    $pos \leftarrow h_1(x)$
    for $i \leftarrow 0$ to $n$ do
        if  then$T[pos] = Null$
            $T[pos] = x$ ; return
        end if
        $x \longleftrightarrow T[pos]$
        if $pos = h_1(x)$ then
            $pos \leftarrow h_2(x)$
        else
            $pos \leftarrow h_1(x)$
        end if
    end for
    REHASH
    INSERT(x)
end function
```

Théorème 16.0.1. *This insertion is done in constant time.*

Lemme 16.0.2. *Suppose that $|T| \geq c \cdot n$ for some $c > 1$. For any i, j, the probability that there exists a path from i to j of length $l \geq 1$ which is a shortest path from i to j is at most $\frac{1}{c^l \cdot |T|}$*

Démonstration. By induction on l :

- Initialization : By SUHA, $\mathbb{P}[h_{1,2}(x) = y] = \frac{1}{|T|}$. Thus $\mathbb{P}[\text{there is an edge from } i \text{ to } j] = \frac{n}{|T|^2} \leq \frac{1}{c|T|}$
- Heredity : For $l \geq 1$ there must exist k such that there is a path of length $l-1$ from i to k and an edge from k to j.

■

Démonstration. We define the bucket of x as all the cells that can be reached either from $h_1(x)$ or $h_2(x)$. x, y are in the same bucket if and only if there is a path from $\{h_1(x), h_2(x)\}$ to $\{h_1(y), h_2(y)\}$. Then :

$$\mathbb{P}[x, y \text{ are in the same bucket}] \leq 4 \sum_{l=1}^{\infty} \frac{1}{c^l\,|T|} = \frac{4}{(c-1)\,|T|}$$

So :

$$\mathbb{E}[|\text{bucket of } x|] = \mathbb{E} \sum X_{x,y} = \sum \mathbb{E}[X_{x,y}] = \sum \mathbb{P}[x, y \text{ are in the same bucket}] \leq \frac{4}{c-1}$$

Hence, in the absence of rehash, expected insertion time in constant.

The probability that we need a rehash is at most probability that there is a cycle, i.e. a path from i to i : $\frac{4}{c-1} \leq \frac{1}{2}$. The probability that we will need d rehashes is then at most $\frac{1}{2^d}$. Thus, the expected time per insertion is :

$$\frac{1}{n} \cdot \mathcal{O}(n) \sum_{d=1}^{+\infty} \frac{1}{2^d} = \mathcal{O}(1)$$

■

16.1 Rolling Hash Functions

Définition 16.1.1. *The Karp-Rabin fingerprint of a string $S = s_1 \dots s_m$ is :*

$$\varphi(s_1 \dots s_m) = \sum_{i=1}^{m} s_i r^{m-1} \mod p$$

where p is a big enough prime and $r \in \mathbb{F}_p$.

Proposition 16.1.1. — *If $S = T$, then $\varphi(S) = \varphi(T)$*

— *Else, $\varphi(S) \neq \varphi(T)$ with high probability.*

Démonstration. Let σ be the size of the alphabet, $p \geq max\,\{\sigma, n^c\}$ where $c > 1$ is a constant. We have :

$$\varphi(S) = \varphi(T) \Leftrightarrow \sum_{i=1}^{m} (s_i - t_i) \cdot r^{m-i} \mod p = 0$$

Hence, r is a root of $P(x) = \sum_{i=1}^{m} (s_i - t_i) \cdot x^{m-i}$ a polynomial over $\mathbb{F}_p$. The probability of such event is at most $\frac{m}{p} \leq \frac{1}{n^{c-1}}$. ■

The Karp-Rabin algorithm is as follows :

- Compute the fingerprint of the pattern
- Compare it with the fingerprint of each m-length substring of the text. If the fingerprint is equal to the fingerprint of a substring, report it as an occurrence

Proposition 16.1.2. *We have :*

$$\varphi(s_1 \dots s_{j+1}) = \varphi(s_1 \dots s_j) \cdot r + s_{j+1} \mod p$$

Démonstration. Observe that :

$$\varphi(s_1 \dots s_{j+1}) = \sum_{i=1}^{j+1} s_i r^{j+1-i} \mod p$$

■