TD 4. Tables de hachage

Espérance de X : $\mathbb{E}[X] = \sum_{v \in V} v \times \Pr[X = v]$
Linéarité de l'espérance : $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
Inégalité de Markov : $\Pr[X \geq \lambda \mathbb{E}[X]] \leq \frac{1}{\lambda}$.
Borne de l'union : $\Pr[E \vee F] \leq \Pr[E] + \Pr[F]$
Probabilité conditionnelle : $\Pr[E|F] = \Pr[E \wedge F] / \Pr[F]$

Exercice 1. *Tables sans collision*

Une fonction de hachage $h : \{0, \dots, N-1\} \to \{0, \dots, m-1\}$ est *sans collision* pour un ensemble $K \subset U$ si pour tout $x, y \in K$, $h(x) \neq h(y)$.

1. Donner une condition nécessaire et suffisante sur K pour qu'il existe une fonction de hachage sans collision pour K.
2. On s'intéresse à l'espérance du nombre de collisions, en fonction de m et $n = |K|$.
   i. Calculer l'espérance dans le modèle idéalisé où h est uniforme parmi les fonctions de $\{0, \dots, N-1\}$ dans $\{0, \dots, m-1\}$.
   ii. Calculer l'espérance dans le modèle universel où h est uniforme dans un ensemble universel.
3. On suppose maintenant que $m = n^2$. Montrer, dans les deux modèles idéalisé et universel, qu'avec probabilité $\geq \frac{1}{2}$, h est sans collision pour K.
4. On veut stocker n couples (clé,valeur) dans une table de hachage, sans avoir à gérer les collisions (pas de chaînage ni d'adressage ouvert). Pour cela, on utilise une table de taille $m = n^2$. On tire une première fonction de hachage et on insère tous les éléments dans la table. S'il y a une collision, on recommence avec une nouvelle fonction de hachage, etc.
   i. Quelle est la probabilité que la première fonction de hachage soit sans collision ?
   ii. Quelle est l'espérance du nombre de fonctions de hachage à tester pour en trouver une sans collision ?

Exercice 2. *Hachage parfait*

Le hachage parfait résout le problème des collisions en mettant, dans chaque case non vide de la table, une autre table de hachage. La structure est alors constituée d'une *table principale* $\mathcal{T} = (T, h)$, dont la case $T_{[i]}$ est soit vide, soit contient une *table secondaire* $\mathcal{T}^i = (T^i, h^i)$. Dans les tables secondaires, on n'accepte aucune collision (ni chaînage ni adressage ouvert) : pour cela, on utilise une table de taille $m_i = n_i^2$ pour stocker n_i éléments.

On se place dans un cadre *statique* en deux phases : 1. on initialise la table en insérant n couples (clé, valeur) ; 2. on effectue des appels à RECHERCHER uniquement. On utilise le modèle universel des fonctions de hachage.

1. Pendant l'initialisation, on crée chaque table secondaire $\mathcal{T}^i$ en tirant des fonctions de hachage h^i jusqu'à en avoir une sans collision pour les éléments de cette table. Justifier que l'espérance du nombre total de fonction de hachage à tirer pendant la phase d'initialisation est $O(n)$.
2. Justifier que lors de la phase de requêtes, chaque appel à RECHERCHER coûte $O(1)$ en pire cas.
3. On veut montrer que l'espérance de la taille *totale* de la table est $O(n)$. On compte le nombre total de cases : les cases de T et les cases de chaque T^i. On note K l'ensemble des clés insérées dans $\mathcal{T}$, et on définit les variables aléatoires $X_{k,i}$ pour $k \in K$ et $i \in \{0, \dots, n-1\}$ par $X_{k,i} = 1$ si $h(k) = i$ et 0 sinon.
   - **i.** Pour tout i, exprimer n_i en fonction des $X_{k,i}$, et exprimer la taille totale en fonction de n et des n_i.
   - **ii.** Montrer que $n_i^2 = \sum_{k \in K} X_{k,i} + \sum_{k_1 \neq k_2} X_{k_1,i} X_{k_2,i}$.
   - **iii.** Montrer que $\sum_{i=0}^{n-1} \sum_k X_{k,i} = n$.
   - **iv.** Montrer que pour $k_1 \neq k_2$, $\mathbb{E}\left[\sum_{i=0}^{n-1} X_{k_1,i} X_{k_2,i}\right] \leq \frac{1}{n}$.
   - **v.** En déduire que $\mathbb{E}\left[\sum_{i=0}^{n} \sum_{k_1 \neq k_2} X_{k_1,i} X_{k_2,i}\right] \leq n-1$.
   - **vi.** En déduire que l'espérance de la taille totale est $\leq 3n-1$.

Exercice 3. *Adressage ouvert*

Soit $\mathcal{T}$ une table de hachage T de taille m contenant n éléments dans laquelle les conflits sont résolus par *adressage ouvert*. On se place dans le cadre idéalisé suivant : on dispose de m fonctions de hachage h_0, …, h_{m-1} telles que pour tout k, $(h_0(k), h_1(k), \dots, h_{m-1}(k))$ est une permutation aléatoire de $\{0, \dots, m-1\}$, et pour $k_1 \neq k_2$ et $0 \leq i, j < m$, $h_i(k_1)$ est indépendant de $h_j(k_2)$.
On effectue une recherche *infructueuse* : on cherche une clé k dans la table qui n'y est pas. Soit X la variable aléatoire qui compte le nombre de cases visitées lors de cette recherche. On a démontré en cours que $\mathbb{E}[X] \leq m/(m-n)$.

1. On souhaite borner $\Pr[X \geq t]$ pour un t fixé. Pour cela, on définit pour tout j l'évènement E_j : « les j premières cases visitées sont occupées ».
   - **i.** Exprimer l'événement « $X \geq t$ » en fonction d'un E_i.
   - **ii.** Montrer que pour $j \geq 0$, $\Pr[E_{j+1} | E_j] = (n-j)/(m-j)$.
   - **iii.** En déduire que pour $1 \leq t \leq m$, $\Pr[X \geq t] \leq (n/m)^{t-1}$.

On étudie maintenant le scénario suivant : on part de la table vide (de taille m) et on insère successivement n clés, avec $n \leq m/2$. On rappelle qu'une insertion doit trouver une première case vide : c'est l'équivalent d'une recherche infructueuse.

2. On note X_i le nombre de cases visitées lors de la $i^{\text{ème}}$ insertion, et $X = \max_{1\le i\le n} X_i$.

i. Montrer que pour tout i, $\Pr[X_i > \lfloor 2\log n\rfloor] \le 1/n^2$.

ii. Montrer que $\Pr[X > \lfloor 2\log n\rfloor] < 1/n$.

iii. En déduire que l'espérance de X est $O(\log n)$. *Couper* $\mathbb{E}[X] = \sum_{t=0}^{n} t\Pr[X = t]$ *en deux sommes* ($t \le \lfloor 2\log n\rfloor$ *et* $t > \lfloor 2\log n\rfloor$) *et borner indépendamment chacune des deux.*

Exercice 4. *Une famille quasi-universelle*

Pour $w > 0$, soit I_w l'ensemble des entiers impairs entre 1 et $2^w - 1$. Pour $\ell < w$ et $a \in I_w$, on définit $h_a(x) = (ax \bmod 2^w) \operatorname{quo} 2^{w-\ell}$ où quo et mod désignent le quotient et le reste dans la division euclidienne. On s'intéresse à la famille $\mathcal{H}_{w,\ell} = \{h_a : a \in I_w\}$.

1. **i.** On écrit ax en binaire, avec les bits $b_0, b_1, \ldots$ Exprimer $h_a(x)$ en fonction des b_i. *Faire un dessin !*

ii. Montrer que $h_a(x) \in \{0, \ldots, 2^\ell - 1\}$ pour tout $x \ge 0$.

2. **i.** Montrer que pour tout x, $y \in I_w$, il existe un unique $a \in I_w$ tel que $ax \bmod 2^w = y$.

ii. En déduire que pour tout x, $y \in I_w$, $\Pr_a[ax \bmod 2^w = y] = 1/2^{w-1}$.

3. Soit $x < y$. Montrer que $h_a(x) = h_a(y)$ si et seulement si $h_a(y-x) = 0$ ou $h_a(y-x) = 2^\ell - 1$.

4. On va montrer que la famille $\mathcal{H}_{w,\ell}$ est quasi-universelle. On fixe pour cela deux entiers positifs $x < y < 2^w$, on pose $z = y - x$ et on écrit $z \bmod 2^w = q2^r$ où $q \in I_{w-r}$ et $r < w$.

i. Montrer que pour tout $v \in I_{w-r}$, $\Pr_a[az \bmod 2^w = v2^r] = 1/2^{w-r+1}$.

ii. En déduire que

$$\begin{cases} \Pr[h_a(z) = 0] = \Pr[h_a(z) = 2^\ell - 1] = 0 & \text{si } r > w-\ell, \\ \Pr[h_a(z) = 0] = 0 \text{ et } \Pr[h_a(z) = 2^\ell - 1] = 1/2^{\ell-1} & \text{si } r = w-\ell, \text{ et} \\ \Pr[h_a(z) = 0] = \Pr[h_a(z) = 2^\ell - 1] = 1/2^\ell & \text{si } r < w-\ell. \end{cases}$$

iii. En déduire que pour tout $x \ne y$, $\Pr[h_a(x) = h_a(y)] \le 1/2^{\ell-1}$.

Exercice 5. *Case la plus remplie*

Soit $h : U \to \{0, \ldots, n-1\}$ une fonction de hachage aléatoire. On insère n clefs dans une table $\mathcal{T}$ de taille n à l'aide de h, en utilisant une résolution par chaînage. On souhaite connaître l'espérance de la case de $\mathcal{T}$ la plus remplie.

1. **i.** Soit j un indice entre 0 et $n-1$. Quelle est l'espérance du nombre d'élément en case j ?

ii. Pourquoi on ne peut pas conclure directement ?

2. Soit X_j la variable aléatoire qui compte le nombre d'éléments en case $T_{[j]}$.
   i. Montrer que $\Pr[X_j \geq k] \leq \binom{n}{k}\frac{1}{n^k} \leq 1/k!$.
   ii. En déduire que $\Pr[X_j \geq k] < \frac{1}{(k/2)^{k/2}}$.
3. On pose $k = \frac{4c\log n}{\log\log n}$, pour une certaine constante c.
   i. Justifier que $\frac{2c\log n}{\log\log n} \geq \sqrt{\log n}$ pour n suffisamment grand.
   ii. En déduire que pour n suffisamment grand, $\Pr[X_j \geq k] \leq \frac{1}{n^c}$.
   iii. En déduire que la probabilité que la case la plus remplie possède plus de $\frac{4c\log n}{\log\log n}$ éléments est $\leq 1/n^d$ pour une constante d à déterminer.
4. On note M le nombre d'éléments dans la case la plus remplie, et on veut borner $\mathbb{E}[M]$.
   i. Montrer que pour tout k, $\mathbb{E}[M] \leq k\Pr[M \leq k] + n\Pr[M > k]$.
   ii. En déduire que $\mathbb{E}[M] = O(\log n/\log\log n)$.

Exercice 6. *Filtres de Bloom*

Les filtres de Bloom permettent de stocker de manière très compressée un ensemble (statique, c'est-à-dire duquel on ne supprime jamais d'élément). La contrepartie est la présence de faux-positifs : le filtre répond parfois que x appartient à l'ensemble alors que ça n'est pas le cas.

Un filtre de Bloom pour un ensemble de taille n est donné par un entier m (la taille de la représentation) et k fonctions de hachage $h_1, \ldots, h_k$ indépendantes. Un ensemble X est représenté par un mot booléen w de taille m. L'ensemble vide est représenté par le mot $0\cdots0$. Pour insérer un nouvel élément x, on passe à 1 les k bits de w d'indices $h_1(x), \ldots, h_k(x)$. Un bit peut être mis plusieurs fois à 1. Pour tester si un élément y appartient à x, on vérifie si $w_{h_j(y)}$ vaut 1 pour $1 \leq j \leq k$: si c'est le cas, on répond « oui » et sinon on répond « non ».

Dans la suite, on suppose qu'on a construit la représentation w d'un ensemble X de taille n. On se place dans le modèle aléatoire pour les fonctions de hachage.

1. Laquelle des deux réponses de l'algorithme de recherche est toujours exacte ?
2. Montrer que le i-ème bit w_i de w vaut 1 si et seulement s'il existe $x \in X$ et j tels que $h_j(x) = i$.
3. Quelle est la probabilité p que le i-ème bit de w soit égal à 0 ? *On rappelle qu'on se place dans le modèle aléatoire, et que la probabilité dépend du choix des fonctions de hachage.*

On fait maintenant l'hypothèse qu'une fraction p des bits de w sont à 0.

4. Pourquoi cette hypothèse ne découle pas de la question précédente ?
5. Soit $y \notin X$. Quelle est la probabilité d'obtenir un faux-positif, c'est-à-dire que l'algorithme de recherche réponde « oui » sur l'entrée y ?
6. Montrer qu'en prenant $k = m\ln 2/n$, cette probabilité est exponentiellement petite. *On pourra utiliser, entre autres, que* $1 - x \geq e^{-2x}$ *pour* $x \leq 1/2$.