TD 4 : Hachage

Exercice 1. *Hachage sans collision*

Une fonction de hachage $h : U \to \{0,\dots,m-1\}$ est *sans collision* pour un ensemble $X \subset U$ si pour tout $x, y \in X$, $h(x) \neq h(y)$. Dans cet exercice, on suppose X fixé.

1. Donner une condition nécessaire et suffisante sur X pour qu'il existe une fonction de hachage sans collision pour X.
2. Supposons qu'on ait choisi une fonction h aléatoirement et uniformément. Exprimer l'espérance du nombre de collisions pour X en fonction de m et $n = |X|$.
3. Quelle est la probabilité qu'une fonction aléatoire h soit sans collision pour X.
4. Suposons qu'on cherche une fonction sans collision pour X en tirant des fonctions aléatoires tant qu'on en a pas trouvé une qui convienne. On note T la variable aléatoire correspondant au nombre d'essais nécessaire avant de trouver une fonction sans collision. Et on note E l'espérance de T.
   - **a.** En utilisant la formule de l'espérance totale conditionnée au fait de trouver une fonction sans collision au premier tirage ou non, montrez que $E = 1 + (1 - m!/(m-n)!m^n)E$.
   - **b.** En déduire que $E = (m-n)!m^n/m!$
   - **c.** Retrouver le résultat précédent en considérant une variable aléatoire suivant une loi géométrique que l'on précisera.

Exercice 2. *La case la plus remplie*

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

1. Soient j un indice entre 0 et $n-1$ et X_j la variable aléatoire qui compte le nombre d'éléments en case $T_{[j]}$.
   - **a.** Quelle est l'espérance du nombre d'éléments en case j, c'est-à-dire, que vaut $\mathrm{E}[X_j]$?
   - **b.** Pourquoi on ne peut pas conclure directement ?
2. Afin de majorer $\mathrm{E}[\max_i X_i]$, on commence par établir les relations suivantes.
   - **a.** Montrer que $\Pr[X_j \geq \ell] \leq \binom{n}{\ell}\frac{1}{n^\ell}$.
   - **b.** Montrer que $\binom{n}{\ell} \leq \frac{n^\ell}{\ell!}$.
   - **c.** En admettant que $(\ell!)^2 \geq \ell^\ell$ pour tout $\ell \geq 1$, déduire des questions précédentes que $\Pr[X_j \geq \ell] \leq \frac{1}{\ell^{\ell/2}}$.
3. On pose $\ell = \frac{c\log n}{\log\log n}$, pour une certaine constante c.
   - **a.** Justifier que $\frac{c\log n}{\log\log n} \geq \sqrt{\log n}$ pour n suffisamment grand.
   - **b.** En déduire que pour n suffisamment grand, $\frac{1}{\ell^{\ell/2}} \leq \frac{1}{n^{c/4}}$, puis que $\Pr[X_j \geq \ell] \leq \frac{1}{n^{c/4}}$.
4. Pour la fin de l'exercice, on note M le nombre d'élément dans la case la plus remplie, c'est-à-dire $M = \max_i X_i$.
   - **a.** Montrer que $\Pr[M \geq \ell] \leq n.\Pr[X_j \geq \ell]$.
   - **b.** En déduire que la probabilité que la case la plus remplie possède plus de $c\log n/\log\log n$ éléments est $\leq 1/n^d$ pour une constante d à déterminer.
5. On va pouvoir maintenant borner $\mathrm{E}[M]$.
   - **a.** Montrer que pour tout ℓ, $\mathrm{E}[M] \leq \ell.\Pr[M \leq \ell] + n.\Pr[M > \ell]$.
   - **b.** À l'aide de la question 4.b et en majorant $\Pr[M \leq \ell]$ par 1, en déduire que $\mathrm{E}[M] = O(\log n/\log\log n)$.

Exercice 3. *Filtres de Bloom*

On s'intéresse dans cet exercice à une structure de données qui permet 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 : la structure de données répond parfois que x appartient à l'ensemble alors que ça n'est pas le cas. Son utilisation en pratique vient en appui d'une *vraie* structure de donnée, pour fournir un pré-test d'appartenance très rapide [1].

On se donne un ensemble X de taille n sous-ensemble d'un ensemble V. Un *filtre de Bloom* pour l'ensemble X est donné par un entier m (la taille de la représentation) et k fonctions de hachage $h_1, \ldots, h_k : X \to \{0, \ldots, m-1\}$ indépendantes. L'ensemble X est représenté par un mot booléen w de taille m. L'ensemble vide est représenté par le mot $0 \cdots 0$. Pour insérer un nouvel élément x de 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. Maintenant, pour tester si un élément y de V 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 de X. On se place dans le modèle aléatoire uniforme pour le choix des 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 ?
4. On fait maintenant l'hypothèse qu'une fraction p des bits de w sont à 0. 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 $k = m/n$, la probabilité de faux positifs est au plus $(3/4)^{m/n}$, c'est-à-dire, exponentiellement petite. *On pourra utiliser, entre autres, que* $1 - x \geq 2^{-2x}$ *pour* $0 \leq x \leq 1/2$.

Exercice 4. *Adressage Ouvert*

On suppose qu'on dispose d'une table de hachage T de taille m, contenant n éléments. Les conflits sont résolus par *adressage ouvert* : on dispose de m fonctions de hachages $h_0, \ldots, h_{m-1}$ et un élément x est inséré en case $T[h_0(x)]$ si elle est libre, sinon en case $T[h_1(x)]$ si elle est libre, et ainsi de suite. *On suppose l'hypothèse forte de hachage uniforme : pour tout x, $(h_0(x), h_1(x), \ldots, h_{m-1}(x))$ est une permutation aléatoire de $\{0, \ldots, m-1\}$, et si $x \neq y$, $h_i(x)$ est indépendant de $h_j(y)$ pour tout i et tout j.*

On effectue une recherche *infructueuse* : on cherche un élément x dans la table mais il n'y est pas. On souhaite borner l'espérance $E_{m,n}$ du nombre de cases visitées lors de cette recherche.

1. Montrer que pour tout nouvel élément x, la probabilité que $T[h_0(x)]$ soit libre est $1 - n/m$.
2. Montrer que $E_{m,n} = 1 + \frac{n}{m} E_{m-1,n-1}$.
3. En déduire que $E_{m,n} \leq m/(m-n)$.
4. On note X la variable aléatoire qui compte le nombre de cases visitées lors d'une recherche infructueuse. On vient de montrer que $\mathrm{E}[X] = E_{m,n} \leq m/(m-n)$. On souhaite maintenant borner $\Pr[X \geq k]$ pour un k fixé. Pour cela, on définit pour tout j l'évènement E_j : « les j premières cases visitées sont occupées ».
   - **a.** Exprimer l'évènement « $X \geq k$ » en fonction de $E_1, \ldots, E_{k-1}$, pour $k \geq 2$.
   - **b.** En déduire que $\Pr[X \geq k] = \Pr[E_{k-1} | E_1 \wedge E_2 \wedge \cdots \wedge E_{k-2}] \Pr[X \geq k-1]$, pour $k \geq 2$.
   - **c.** Montrer que pour tout $j > 1$, $\Pr[E_j | E_1 \wedge \cdots \wedge E_{j-1}] = \frac{n-j+1}{m-j+1}$.
   - **d.** En déduire que $\Pr[X \geq k] \leq (n/m)^{k-1}$ pour $1 \leq k \leq m$.
5. On imagine maintenant qu'on part de la table vide (de taille m) et qu'on insère successivement n valeurs, avec $n \leq m/2$. On rappelle qu'une insertion doit trouver la première case vide parmi les cases d'indices $h_0(x), \ldots, h_{m-1}(x)$: cette recherche est l'équivalent d'une recherche infructueuse. On note X_i le nombre de cases visitées lors de la $i^{\text{ème}}$ insertion, et $X = \max_{1 \leq i \leq n} X_i$.

1. Voir `https://en.wikipedia.org/wiki/Bloom_filter#Examples` pour de nombreux exemples d'utilisation de ces objets en pratique.

a. Montrer que pour tout i, $\Pr[X_i > k] < 1/2^k$.

b. En déduire que pour tout i, $\Pr[X_i > 2\log n] < 1/n^2$.

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

d. En déduire que l'espérance de X est $O(\log n)$.
Écrire $E[X] = \sum_{k\leq 2\log n} kPr[X = k] + \sum_{k>2\log n} kPr[X = k]$ *et borner chacune des deux sommes.*