

Les documents distribués dans le cadre du cours et les notes personnelles sont autorisés.

1 Famille de fonctions de hashage 2-universelle (3 points)

Soit $\mathcal{H}_k$ une famille de fonctions de hashage de $[M]$ vers $[M]$, où M est 2^k pour un entier k , et $[M]$ représente l'ensemble $\{0, 1, \dots, M-1\}$. $\mathcal{H}_k$ est constituée de toutes les fonctions h de la forme

$$h(x) = x \oplus r,$$

où r est un entier de $[M]$ est $\oplus$ est le ou exclusif bit par bit. Par exemple pour $k = 2$, $\mathcal{H}_k$ est composé des 2^k fonctions suivantes

$$\begin{aligned} x &\mapsto x \\ x &\mapsto x \oplus 1 \\ x &\mapsto x \oplus 2 \\ x &\mapsto x \oplus 3. \end{aligned}$$

La première fonction est l'identité, la deuxième change le bit de poids faible, la troisième le bit de poids fort, et la dernière inverse tous les bits.

Est-ce que cette famille est 2-universelle ?

2 Échantillonner dans k-means++ (3 points)

Considérons un flux σ de n points distincts dans $\mathbb{R}^2$. La distance euclidienne est notée par d et définie comme

$$d((x_1, y_1), (x_2, y_2)) := \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}.$$

Par extension $d(a, S)$ pour un point a et ensemble de points S est définie comme $d(a, S) := \min_{b \in S} d(a, b)$.

On veut constituer un ensemble de k points du flux de la manière suivante.

Le premier point p_1 est choisi uniformément au hasard de σ . Puis pour chaque $i = 2, \dots, k$ on choisit comme p_i un point $a \in \sigma$ avec probabilité

$$\frac{d(a, \{p_1, \dots, p_{i-1}\})^2}{\sum_{b \in \sigma} d(b, \{p_1, \dots, p_{i-1}\})^2}.$$

Décrivez un algorithme qui réalise cette tâche en k lectures du flux et prouvez qu'il produit la distribution probabiliste demandée.

3 Deux éléments manquants (3 points)

On vous donne un flot de $n - 2$ entiers distincts tous entre 0 et $n - 1$. Il y a donc exactement deux entiers $p, q \in [n]$ qui ne sont pas présents dans le flot. Concevez un algorithme de flux qui détermine les deux entiers manquants. Quel est sa complexité en mémoire ? Si vous n'avez pas d'idée, commencez par trouver une solution pour le cas $n = 4$ puis $n = 8$. En général l'entier n peut ne pas être une puissance de 2.

A Corrections

A.1 Famille de fonctions de hashage 2-universelle

Non. La famille n'est pas assez large. En effet pour une fonction $h \in \mathcal{H}_k$ on a $h(x) = y$ et $h(x') = y'$ si et seulement si $x \oplus x' = y \oplus y'$. Donc pour $x' \neq x$ la probabilité $\mathbb{P}_h[h(x) = y \wedge h(x') = y']$ est 0 si $x \oplus x' \neq y \oplus y'$ et 1 sinon. Mais pour être 2-universelle cette probabilité devrait être $1/M^2$ dans tous les cas.

Par contre la famille est 1-universelle, chaque valeur y a la même probabilité $1/2^k$ d'être l'image $h(x)$ pour $h \in \mathcal{H}_k$. Pour s'en convaincre il suffit de le montrer pour $k = 1$, et d'observer que h agit sur chaque bit de manière indépendante.

A.2 Échantillonner dans k-means++

Pour le premier point on maintient un compteur r et on garde le r -ème point avec probabilité $1/r$. La distribution du choix est uniforme ce qu'on peut prouver par induction sur r .

Pour le i -ème point, on utilise la même technique, mais pondérée. On maintient un compteur D de toutes les distances entre $\mathcal{P} = \{p_1, \dots, p_{i-1}\}$ et les points observés. Alors lors du traitement du r -ème point a on ajoute $d(p_1, \mathcal{P})$ à D et on sélectionne $p_i = a$ avec probabilité $d(p_1, \mathcal{P})/D$.

Comme les points sont distincts, D ne sera pas zéro, et le premier point est sélectionné avec probabilité 1. Puis il faut montrer qu'à tout moment la probabilité qu'un point b soit p_i est $d(b, \mathcal{P})/D$. C'est le cas par l'algorithme au moment du traitement de b . Puis plus tard quand un point a est traité, D sera augmenté à $D' = D + d(a, \mathcal{P})$, et a sera choisi avec probabilité $d(a, \mathcal{P})/D'$ tandis que b restera le choix de p_i avec probabilité $d(b, \mathcal{P})/D \cdot D/D' = d(b, \mathcal{P})/D'$.

A.3 Deux éléments manquants

Essayons d'abord d'appliquer la méthode pour trouver un unique élément manquant. On fait la somme S de tous les éléments du flux et on pose $x = n(n-1)/2 - S$. On a alors $x = p + q$, où p, q sont les deux éléments manquants. Il nous manque donc une information, car plusieurs couples d'éléments manquants ont la même somme. En d'autres mots il n'est pas possible de résoudre un système d'équation à une seule équation et deux inconnus.

Algorithme à deux passes et mémoire $O(\log n)$ Certains étudiants ont alors eu l'idée suivante. Si on suppose $p < q$, alors on sait que $p < x/2 < q$. Il suffit alors de restreindre le flux aux éléments inférieurs à $x/2$ et dans une deuxième passe de déterminer l'unique élément p manquant dans ce flux restreint.

Algorithme à une passe et mémoire $O(\log^2 n)$ Pour trouver un algorithme à une seule passe, il faudrait donc extraire plusieurs informations du flux. La complexité en mémoire de cet algorithme est $O(\log^2 n)$ bits, correspondant à $2 \log n$ compteurs. Dans x_i^b on cumule tous les entiers y du flux dont le i -ème bit dans l'expansion binaire

est b . Soit z_i^b la somme de tous les entiers $y \in [n]$ dont le i -ème bit dans l'expansion binaire est b . Alors $x_i^b \leq z_i^b$ et il y a égalité si et seulement si à la fois p et q ont le i -ème bit dans l'expansion binaire est $1 - b$. Comme $p \neq q$ il existe forcément une position i où les expansions binaires de p et q diffèrent. Pour cette position nous avons les inégalités strictes $x_i^0 < z_i^0$ et $x_i^1 < z_i^1$. Et parmi les différences $z_i^0 - x_i^0$ et $z_i^1 - x_i^1$ une est forcément p et l'autre q .

Algorithme à une passe et mémoire $O(\log n)$ On calcule la somme de toutes les valeurs du flux et la somme de tous les carrés des valeurs du flux. Ainsi on a un système d'équation à résoudre de la forme $x = p + q, y = p^2 + q^2$, pour deux valeurs x, y ainsi extraite du flux. En posant $q = x - p$ on trouve

$$\begin{aligned}
 p^2 + (x - p)^2 &= y \\
 p^2 + x^2 - 2xp + p^2 &= y \\
 2p^2 - 2xp + x^2 - y &= 0 \\
 p &= \frac{2x \pm \sqrt{4x^2 - 8(x^2 - y)}}{4} \\
 p &= \frac{x - \sqrt{2y - x^2}}{2} \\
 q &= x - p.
 \end{aligned}$$