Chaque question peut être traitée en admettant les résultats des questions précédentes.

Le sujet porte sur des problèmes de « recherche combinatoire » : dans un espace de recherche connu, on veut reconnaître un élément v fixé mais auquel on n'a accès qu'à travers un type limité de questions, supposées très coûteuses ; l'objectif est de minimiser le nombre de questions utilisées pour reconstruire v, quitte à faire par ailleurs plus de calculs.

On utilisera les notations $g_{n,d} = O(f_{n,d})$ ou $f_{n,d} = \Omega(g_{n,d})$ pour signifier qu'il existe une constante $c > 0$, telle que $g_{n,d} < c\, f_{n,d}$ pour tout n et d dans le domaine de définition. Enfin $\lceil x \rceil$ et $\lfloor x \rfloor$ désignent respectivement l'arrondi entier supérieur et inférieur de x.

Un problème de pesage

On considère un ensemble de n pièces réparties en deux groupes : les vraies de poids p et les fausses de poids $q \neq p$ (toutes les fausses pièces sont de même poids). Connaissant les poids p et q et sachant qu'il y a d fausses pièces, comment les trouver à l'aide d'une balance de précision (étalonnée, avec un seul plateau) en minimisant le nombre de pesées ? Formalisons ce problème : on numérote arbitrairement les pièces de 1 à n et on représente la configuration recherchée par un vecteur $v = (v_1, \ldots, v_n)$ avec $v_i = 1$ si la pièce i est fausse, 0 sinon.

Problème 1. Recherche de d fausses pièces parmi n.

- L'espace de recherche est l'ensemble des vecteurs $v = (v_1, \ldots, v_n)$ de $\{0,1\}^n$ de poids $|v| = \sum_{i=1}^{n} v_i = d$. On recherche un élément v fixé.
- Une question est un sous-ensemble Q de $\{1, \ldots, n\}$. La réponse $r_v(Q)$ à la question Q est $r_v(Q) = \sum_{i \in Q} v_i$. (Le nombre de fausses pièces dans le sous-ensemble pesé.)

Les algorithmes de recherche que nous considérons n'ont accès à l'inconnue v qu'au travers de ces questions. De plus on considère que le coût d'une question est très grand devant le coût du traitement des réponses, ce qui nous amène dans toute cette première partie à définir comme *complexité* d'un algorithme le nombre de questions posées dans le cas le pire.

Par exemple, si $d = 1$, on peut définir un algorithme de recherche binaire, dont l'appel `dicho(1,n ;v)` renvoie le numéro de la fausse pièce en posant au plus $\lceil \log_2 n \rceil$ questions :

---

algorithme `dicho(i,j ;v)`

- si $(i = j)$ renvoyer i.
- $k := \lceil (i+j)/2 \rceil$, $Q := \{i, \ldots, k\}$,
- si $(r_v(Q) = 0)$ renvoyer `dicho(k+1,j ;v)`,
  sinon renvoyer `dicho(i,k ;v)`.

Un algorithme est *adaptatif* si la $i$ème question Q_i dépend des réponses déjà obtenues aux questions Q_j, $j < i$. C'est le cas de l'algorithme `dicho`. Un algorithme est dit *non-adaptatif* si la suite des questions est fixée une fois pour toute.

Question 1. Donner une borne inférieure sur la complexité d'un algorithme qui résout le problème 1 dans le cas $d = 1$.

Question 2. Proposer un algorithme non-adaptatif optimal qui résout le problème 1 dans le cas $d = 1$. (Indication : on pourra utiliser l'écriture binaire des entiers)

Question 3. Proposer un algorithme adaptatif de complexité $O(d \log n)$ qui résout le problème 1 dans le cas général.

Recherche d'un vecteur de $\{0, \ldots, k-1\}^n$

On considère une variante du problème 1 avec k types de pièces en quantités inconnues.

Problème 2. Recherche d'un vecteur de $\{0, \ldots, k-1\}^n$.

- L'espace de recherche est l'ensemble des vecteurs $v = (v_1, \ldots, v_n)$ de $\{0, \ldots, k-1\}^n$. On recherche un élément v fixé.
- Une question est un sous-ensemble Q de $\{1, \ldots, n\}$. La réponse $r_v(Q)$ à la question Q est $r_v(Q) = \sum_{i \in Q} v_i$.

Comme le problème 1, ce problème peut se résoudre en temps linéaire à l'aide des questions $Q_i = \{i\}$. L'objectif des questions qui suivent est de montrer qu'on peut faire mieux.

Question 4. Montrer que l'existence d'un algorithme non-adaptatif de complexité p qui résout le problème 2 est équivalente à l'existence d'une matrice C d'entiers $\{0, 1\}$ de taille $p \times n$ qui soit *séparatrice*, *i.e.* telle que pour tous v et v' de $\{0, \ldots, k-1\}^n$,

$$C \cdot v = C \cdot v' \quad \Rightarrow \quad v = v'.$$

Sous-ensembles de $\{1, \ldots, m\}$ et ordre partiel d'inclusion La relation $x \subset y$ définit un ordre partiel sur les sous-ensembles de $\{1, \ldots, m\}$. La fonction de Moebius $\mu(x, y)$ associée à cet ordre partiel est donnée par

$$\mu(x, y) = \begin{cases} (-1)^{|y \setminus x|} & \text{si } x \subset y, \\ 0 & \text{sinon.} \end{cases}$$

Question 5. Soit y un sous-ensemble non vide de $\{1, \ldots, m\}$ et $h \leq 2^{|y|-1}$ un entier positif. Montrer qu'il existe une fonction $f_{y,h}$ de l'ensemble des sous-ensembles de y dans $\{0, 1\}$ telle que $f_{y,h}(\emptyset) = 0$ et

$$\sum_{x \subset y} f_{y,h}(x) \cdot (-1)^{|y \setminus x|} \; = \; (-1)^{|y|-1} h. \tag{1}$$

Illustrer votre construction pour $f_{y,h}$, $y \subset \{1, 2\}$.

Question 6. Soit y, y' tels que $y \not\subset y'$, et g une fonction définie sur les sous-ensembles de $y \cap y'$. Montrer que

$$\sum_{x \subset y} g(x \cap y') \cdot (-1)^{|y \setminus x|} \; = \; 0. \tag{2}$$

Soit h une fonction des sous-ensembles de $\{1, \ldots, m\}$ dans les entiers telle que $1 \leq h(y) \leq 2^{|y|-1}$ et soit $M = 2^m - 1$. Considérons la matrice $A = (f_{y,h(y)}(x \cap y))_{x,y}$ de taille $M \times M$ dont les lignes et colonnes sont indexées par des sous-ensembles x et y non vides de $\{1, \ldots, m\}$.

Question 7. Pour $m = 2$, construire les matrices A pour $h(\{1, 2\}) = 1$ et pour $h(\{1, 2\}) = 2$.

Question 8. Soit r et w deux vecteurs dont les coordonnées sont indexées par les sous-ensembles non vides de $\{1,\ldots,m\}$ et tels que $A \cdot w = r$. Supposons que y est un sous-ensemble tel que pour tout y' avec $y \subsetneq y'$ la coordonnée $w_{y'}$ de w est nulle. Montrer alors

$$\sum_{\substack{x \subset \{1,\ldots,m\} \\ x \neq \emptyset}} r_x \, \mu(x,y) \;=\; (-1)^{|y|-1} h(y)\, w_y.$$

En déduire que la matrice A est séparatrice au sens de la question 4.

Question 9. Déduire de la question 8 précédente un algorithme qui, étant donné r, détermine le vecteur w tel que $A \cdot w = r$ en utilisant un nombre d'opérations arithmétiques polynomial en M.

Application aux vecteurs à coordonnées bornées La matrice précédente A est carrée et conduit donc toujours à un algorithme linéaire pour le problème 2. On veut améliorer la complexité, en utilisant l'idée d'écriture en base k pour tirer profit du fait que les coordonnées du vecteur recherché sont dans $\{0,\ldots,k-1\}$.

On considère pour cela une variante $B = (b_{x,(y,i)})$ de la matrice A, dans laquelle, pour tout y sous-ensemble non vide de $\{1,\ldots,m\}$, la colonne indexée par y de A est remplacée dans B par $\ell(y) = \lfloor \log_k 2^{|y|-1} \rfloor$ colonnes indexées (y,i) pour $i = 0,\ldots,\ell(y)-1$ avec $b_{x,(y,i)} = f_{y,k^i}(x,y)$. Soit $N(m)$ le nombre de colonnes de B.

Question 10. Construire B pour $m = 3$ et $k = 2$.

Question 11. Montrer que B est toujours séparatrice et adapter l'algorithme de la question 9 à la détermination d'un vecteur w de taille $\{0,\ldots,k-1\}^{N(m)}$ connaissant $r = B \cdot w$.

Question 12. Montrer que le nombre de colonnes $N(m)$ est minoré par $2^m(m/(2\log_2 k) - 2)$.

Question 13. Déduire des questions précédentes un algorithme de complexité $O(n \log k / \log n)$ qui résolve le problème 2. (On supposera que $k = O(\log n)$.) Expliciter l'algorithme pour $k = 2$, $m = 3$.

Recherche d'un graphe de degré borné

Un graphe $G = (V,E)$ est la donnée d'un ensemble de sommets V et d'un ensemble d'arêtes $E \subset \mathcal{P}_2(V)$ (paires d'éléments distincts de V). Le degré d'un sommet est le nombre d'arêtes qui lui sont incidentes, c'est-à-dire qui le contiennent. Un graphe de degré au plus d est un graphe dont tous les sommets sont de degré inférieur ou égal à d.

Problème 3. Recherche d'un graphe.

- L'espace de recherche est l'ensemble des graphes de degré au plus d sur un ensemble de sommets V. On recherche un graphe $G = (V,E)$ fixé.
- Une question est un sous-ensemble Q de V. La réponse $\mu_G(Q)$ à la question Q est le nombre d'arêtes du sous-graphe induit par Q, $\mu_G(Q) = |E \cap \mathcal{P}_2(Q)|$.

Un graphe biparti $G' = (V_1, V_2, E')$ est la donnée de deux ensembles de sommets V_1 et V_2 et d'un ensemble d'arêtes $E' \subset V_1 \times V_2$. Le problème 3 admet la variante suivante.

Problème 3'. Recherche d'un graphe biparti.

- L'espace de recherche est l'ensemble des graphes bipartis de degré au plus d sur les ensembles de sommets V_1 et V_2. On recherche un graphe $G' = (V_1, V_2, E')$ fixé.

– Une question est un couple (X, Y) avec $X \subset V_1$ et $Y \subset V_2$. La réponse $\mu_{G'}(X, Y)$ à la question (X, Y) est le nombre d'arêtes entre X et Y, $\mu_{G'}(X, Y) = |E' \cap (X \times Y)|$.

Ces deux problèmes admettent naturellement une solution triviale de complexité quadratique en testant chaque arête séparément.

Question 14. Donner un algorithme de complexité au plus $d|V_1| \log_2 |V_2|$ pour le problème 3'.

Question 15. Montrer que le problème 3 de la recherche de graphes de degré borné se ramène au problème 3' pour les graphes bipartis et donner un lien entre les complexités de ces deux problèmes.

(indication : utiliser un graphe biparti avec le double de sommets du graphe initial.)

Question 16. Dans le cas $d = 1$ du problème 3', où le graphe recherché est un couplage, proposer un algorithme de complexité $O(|V_1| \log |V_2| / \log |V_1|)$.

(Indication : on pourra utiliser les résultats des questions 2 et 13.)

Recherche d'une partition

Soit $X = \{x_1, \ldots, x_n\}$ un ensemble fixé. On considère une relation d'équivalence $\sim$ sur cet ensemble ou, de manière équivalente, une partition $(X, \sim)$ de X en r classes d'équivalence $X_1, \ldots, X_r$: x et y deux éléments de X sont dans la même classe si et seulement si $x \sim y$, et on dit alors que x et y représentent la même classe.

Problème 4. Recherche d'une partition.

– L'espace de recherche est l'ensemble des partitions en r classes de l'ensemble $X = \{x_1, \ldots, x_n\}$. On recherche une partition $(X, \sim)$ fixée.
– Une question est un sous-ensemble Q de l'ensemble des objets X. La réponse $r_\sim(Q)$ à la question Q est le nombre de classes représentées par les objets de Q.

Question 17. Donner un algorithme incrémental (qui reconstruit successivement les partitions $\{x_1, \ldots, x_i\}$, $i = 1, \ldots, n$) de complexité $O(n \log r)$.

Question 18. Soit $X_1 \subset X$ et $X_2 = X \setminus X_1$. Supposons les partitions $(X_1, \sim)$ et $(X_2, \sim)$ reconstruites. Montrer que la reconstruction de $(X, \sim)$ se ramène alors à un problème considéré à la section précédente. Quelle est la complexité de cette étape ?

Question 19. Supposons maintenant que $X = X_1 \cup \ldots \cup X_k$ avec $|X_i| = k$ pour tout k et $|X| = k^2$ (les X_i sont disjoints). Quelle est la complexité de reconstruction de $(X, \sim)$, les partitions $(X_i, \sim)$ étant reconstruites ?

Question 20. Montrer que l'utilisation récursive de la stratégie de la question précédente donne donne un algorithme de complexité $O(n \log \log n)$ pour le problème 4.

Question 21. Proposer un algorithme de complexité $O(n \log \log r)$ pour le problème 4.

(Indication : on pourra différencier les cas $n > r$ et $n \leq r$ dans la stratégie récursive.)

Deux applications à l'analyse de données génomiques

Les développements précédents ont été motivés originellement par des questions d'analyse de données génomiques. On en donne une présentation quelque peu simplifiée.

Un généticien étudie un ensemble de chromosomes distincts mais pour lui inséparables et indistinguables (les chromosomes d'une espèce). Pour cela il a constitué un ensemble de petits

morceaux d'ADN (les « témoins ») tels que chaque témoin, mis en présence des chromosomes, réagit avec un et seul de ces chromosomes. L'opérateur ne peut faire qu'un type d'expérience : copier un sous-ensemble des témoins et les mélanger aux chromosomes, puis observer combien de chromosomes ont réagi. Son objectif est de déterminer les témoins qui réagissent avec un même chromosome. Chaque groupe de témoins ainsi identifié caractérisera un chromosome.

Question 22. En établissant un lien avec les problèmes précédents, estimer le nombre le nombre d'expériences nécessaires à la caractérisation des k chromosomes d'une espèce quelconque à l'aide d'un ensemble de n témoins.

Les chromosomes de certaines bactéries sont circulaires. Le généticien précédent considère maintenant un chromosome de ce type qu'il a pu caractériser à l'aide d'un ensemble de n témoins et cherche à déterminer l'ordre cyclique des témoins sur le chromosome. Il se livre à un nouveau type d'expérience : il copie un sous-ensemble des témoins et par une autre réaction en présence du chromosome, détermine le nombre de paires de témoins du sous-ensemble qui sont voisins sur le cycle.

Question 23. En établissant un lien avec les problèmes précédents, donner une borne sur le nombre d'expériences nécessaires à la détermination de l'ordre cyclique des n témoins sur un chromosome cyclique.