Option Informatique en Spé MP et MP*
Calcul de la médiane en temps linéaire : le corrigé

Question 1 • Un tri par insertion répond à la question.

Question 2 • Si $\mathcal{A}$ effectue moins de treize comparaisons, l'arbre de décision qui lui est associé comporte au plus $2^{12} = 4096$ feuilles. Or il existe $7! = 5040$ permutations d'une liste de longueur 7.

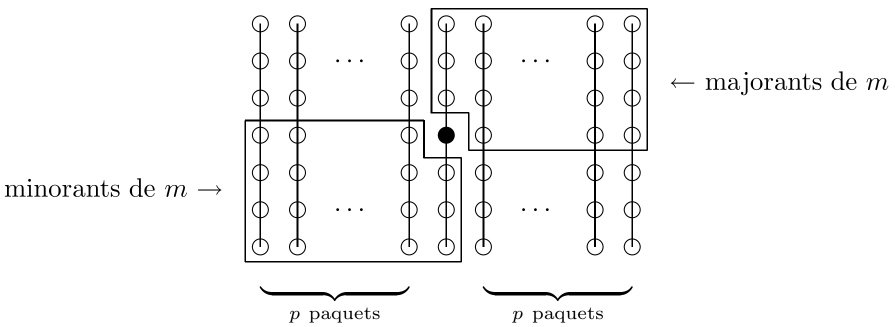
Question 3 • $14p - 7 < n \leqslant 14p + 7$, soit $14(p-1) < n - 7 \leqslant 14p$, ou encore $p = \left\lceil \dfrac{n-7}{14} \right\rceil$.

Question 4 • Le découpage en paquets de sept ne coûte rien, le tri coûte $13(2p+1)$ d'après la remarque de l'énoncé. Notons M le maximum de la liste ; nous calculerons M une fois pour toutes, pour un coût $n-1$; ultérieurement, s'il faut compléter la liste, nous ajouterons des éléments égaux à M ; ceci ne changera pas le résultat.

Question 5 • Par définition, le coût de cette étape est $C(2p+1)$.

Question 6 • Il existe p paquets dont la médiane est inférieure à m ; dans chacun de ces paquets, quatre éléments sont inférieurs à m (la médiane du paquet et ses deux minorants) ; ceci nous fournit $4p$ minorants stricts de m. En ajoutant les trois éléments du paquet médian inférieurs à m, nous obtenons $4p+3$ minorants stricts de m. Symétriquement, nous connaissons $4p+3$ majorants stricts de m.

Dans le dessin, chaque paquet est rangé en ordre croissant vers le haut ; la médiane m est le cercle noir.



Question 7 • Comptons les éléments dont la place par rapport à m n'est pas encore connue : il y en a $3p$ dans les paquets dont la médiane est inférieure à m, et $3p$ dans les paquets dont la médiane est supérieure à m. Ceci nous donne un total de $6p$. Le découpage peut être réalisé au prix de $4p$ comparaisons, puisque chaque paquet de trois est trié : on compare m à la médiane du paquet, puis en fonction du résultat à un seul des deux autres éléments.

Question 8 • $4p + 3 \leqslant r \leqslant 10p + 3$.

Question 9 • Si $k \leqslant r$, la recherche se poursuit dans l'ensemble (connu maintenant) des minorants stricts de m. Si $k = r+1$, le résultat cherché est m. Enfin, si $k > r+1$, on cherche le $(k-r-1)$-ième élément de l'ensemble (connu maintenant) des majorants stricts de m.

Question 10 • $C(n)$ est la somme de quatre coûts :

- découpage en paquets de sept, tri de ces paquets : coût $13(2p+1)$ d'après la question 4 ;
- calcul de la médiane des médianes : coût $C(2p+1)$ d'après la question 5 ; or $n > 1024$ implique $p \geqslant \lceil \frac{1017}{14} \rceil = 73$; mais alors $n - (2p+1) > 12p - 8 > 868$, à plus forte raison $2p + 1 < n$;
- découpage de la liste des éléments non placés par rapport à m : coût $4p$ d'après la question 7 ;
- recherche du k-ième (ou du $(k-r-1)$-ième) élément dans une liste d'au plus $10p+3$: coût $C(10p+3)$ par définition ; or $n > 1024$ implique $p \geqslant 73$, puis $n - (10p+3) > 4p - 10 > 282$, à plus forte raison $10p + 3 < n$.

Question 11 • Pour $32 < n \leqslant 1024$, le tri par insertion d'une liste de longueure n coûte au plus $10n$; mais $n > 32$ implique $n \geqslant 33$ puis $15n - 163 - 10n = 5n - 163 \geqslant 0$, donc $C(n) \leqslant 15n - 163$. Supposons la majoration $C(n) \leqslant 15n - 163$ acquise jusqu'au rang $n_0 > 1024$ inclus. Soit $n = n_0 + 1$; définissons p comme à la question 3 : nous avons vu à la question précédente que $2p + 1$ et $10p + 3$ sont strictement majorés par n, ce qui permet d'appliquer l'hypothèse de récurrence. Nous aurons :

$$C(n) \leqslant 13(2p+1) + 15(2p+1) - 163 + 4p + 15(10p+3) - 163 = 210p - 253 = 15(14p-6) - 163$$

Mais $n \geqslant 14p - 6$; nous en déduisons $C(n) \leqslant 15n - 163$, ce qui termine la preuve.

Références bibliographiques

▶ Le calcul de la médiane, ou plus généralement du k-ième élément d'une liste non triée, constitue le *problème de sélection*. La tradition fait remonter ce problème à Charles DODGSON (plus connu sous son nom de plume de Lewis CARROLL), qui avait observé que, dans un tournoi de tennis, le joueur classé deuxième peut fort bien ne pas être le deuxième meilleur joueur de la compétition.

▶ En 1929, Hugo STEINHAUS proposa au cours d'un séminaire le problème suivant : quel est le plus petit nombre de rencontres que doit compter un tournoi de tennis entre n participants pour être certain que la première et la deuxième place ont été correctement attribuées. SCHREIER proposa en 1932 une méthode nécessitant $n - 2 + \lfloor \lg(n) \rfloor$ rencontres, et affirma que cette méthode était optimale ; hélas, sa preuve était incorrecte. Ce n'est qu'en 1964 qu'une preuve correcte (mais compliquée) fut donnée par KISLITSYN.

▶ Il est clair qu'au prix d'un tri de la liste, on peut déterminer son k-ième élément, pour un coût $\mathcal{O}\big(n \lg(n)\big)$. Les progrès suivants se produisirent en 1971, avec la publication par Manuel BLUM d'un algorithme de coût $\mathcal{O}\big(n \lg(lg(n))\big)$; puis, en 1973, un article du *Journal of Computer and Systems Science*, cosigné par Manuel BLUM, Richard FLOYD, Vaughan PRATT, Ronald RIVEST et Robert TARJAN, exposa le premier algorithme de sélection, de coût linéaire : c'est celui qui est exposé dans ce texte. Une optimisation complémentaire menait les auteurs à un coût majoré par $5{,}43n$.

▶ Le même article donnait une borne *inférieure* du coût du problème de sélection : la détermination du k-ième élément d'une liste de longueur n nécessite au moins $n + \min(k, n-k) - \mathcal{O}(1)$ comparaisons. En 1985, BENT et JOHN améliorèrent ce résultat avec le minorant $2n - o(n)$. Par ailleurs, SCHÖNHAGE, PATERSON et PIPPENGER avaient en 1976 proposé un algorithme de sélection de coût $3n + o(n)$.

▶ Parmi les travaux les plus récents sur ce sujet, la thèse de Dorit DOR (1995) propose un algorithme de coût $2{,}9423n$ et donne une borne inférieure légèrement meilleure que celle de BENT et JOHN.

FIN