

TD n°6 - Recherche d'un élément majoritaire

Soit E une liste de n éléments rangés dans un tableau numéroté de 1 à n . On suppose que la seule opération qu'on sait effectuer sur les éléments est de vérifier si deux éléments sont égaux ou non. On dit qu'un élément $x \in E$ est *majoritaire* si l'ensemble $E_x = \{y \in E | y = x\}$ a strictement plus de $n/2$ éléments. On suppose que n est une puissance de 2, et on s'intéresse à la complexité dans le pire des cas.

Remarque : *On se restreint ici aux algorithmes où la seule opération qu'on sait effectuer sur les éléments est de tester si deux éléments sont égaux. Si la réponse est OUI, on dira qu'il s'agit d'une comparaison égale, et si la réponse est NON d'une comparaison inégale.*

Remarque : *s'il existe un élément majoritaire, il est nécessairement unique.*

1 Algorithme naïf

Écrire un algorithme calculant le cardinal de c_x de E_x pour un x donné. En déduire un algorithme pour vérifier si E possède un élément majoritaire. Quelle est la complexité de cet algorithme ?

Solution :

Algorithme 1 Naïf

```
pour  $i$  de 1 à  $n$  faire
   $c \leftarrow 0$ 
  pour  $j$  de 1 à  $n$  faire
    si  $E[j] = E[i]$  alors
       $c \leftarrow c + 1$ 
  si  $c > n/2$  alors
    retourner " $E[i]$  est majoritaire"
retourner "Pas d'élément majoritaire".
```

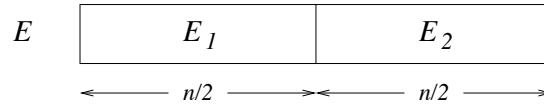
Complexité : *nombre total de comparaisons = n^2 .* □

2 Diviser pour régner

Question 2.1 Donner un autre algorithme récursif basé sur un découpage de E en deux listes de même taille. Quelle est sa complexité ?

Solution : 2.1 - Principe :

- Couper E en deux tableaux E_1 et E_2 de tailles $n/2$ (on suppose n pair).



- S'il existe un élément majoritaire x dans E , alors x est majoritaire dans au moins une des deux listes E_1 et E_2 (en effet si x non majoritaire dans E_1 , ni dans E_2 , alors dans E , $c_x \leq n/4 + n/4 = n/2$).
- Algorithme récursif : calculer les éléments majoritaires de E_1 et de E_2 (s'ils existent) avec le nombre total d'occurrences de chacun et en déduire si l'un des deux est majoritaire dans E .

L'algorithme suivant $Majoritaire(i, j)$ renvoie un couple qui vaut (x, c_x) si x est majoritaire dans le sous-tableau $E[i..j]$ avec c_x occurrences et qui vaut $(-, 0)$ s'il n'y a pas de majoritaire dans $E[i..j]$, l'appel initial étant $Majoritaire(1, n)$. La fonction $Occurrence(x, i, j)$ calcule le nombre d'occurrences de x dans le sous-tableau $E[i..j]$.

Algorithme 2 $Majoritaire(i, j)$

```

si  $i = j$  alors
    retourner  $(E[i], 1)$ 
sinon
     $(x, c_x) = Majoritaire(i, \lfloor (i + j)/2 \rfloor)$ 
     $(y, c_y) = Majoritaire(\lfloor (i + j)/2 \rfloor + 1, j)$ 
    si  $c_x \neq 0$  alors
         $c_x \leftarrow c_x + Occurrence(x, \lfloor (i + j)/2 \rfloor + 1, j)$ 
    si  $c_y \neq 0$  alors
         $c_y \leftarrow c_y + Occurrence(x, i, \lfloor (i + j)/2 \rfloor)$ 
    si  $c_x > \lfloor (j - i + 1)/2 \rfloor$  alors
        retourner  $(x, c_x)$ 
    sinon
        si  $c_y > \lfloor (j - i + 1)/2 \rfloor$  alors
            retourner  $(y, c_y)$ 
        sinon
            retourner  $(-, 0)$ 

```

Complexité : le nombre total de comparaisons dans le pire des cas $C(n)$ vérifie la relation (on suppose que n est une puissance de 2) :

$$C(n) = 2(C(\frac{n}{2}) + \frac{n}{2}) \text{ avec } C(1) = 0$$

On en déduit $C(n) = n \log_2(n)$. : avec le théorème général de résolution des récurrences du Cormen :

Soient $a \geq 1$ et $b > 1$ deux constantes, soit $f(n)$ une fonction et soit $T(n)$ définie pour les entiers non négatifs par la récurrence $T(n) = aT(n/b) + f(n)$. $T(n)$ peut alors être borné de la façon suivante.

1. Si $f(n) = O(n^{\log_b a - \epsilon})$ pour une certaine constante $\epsilon > 0$, alors $T(n) = \Theta(n^{\log_b a})$.
2. Si $f(n) = \Theta(n^{\log_b a})$, alors $T(n) = \Theta(n^{\log_b a} \log n)$.
3. Si $f(n) = \Omega(n^{\log_b a + \epsilon})$ pour une certaine constante $\epsilon > 0$, et si $af(n/b) \leq cf(n)$ pour une certaine constante $c < 1$ et pour tout n suffisamment grand, alors $T(n) = \Theta(f(n))$.

Ici $f(n) = n = \Theta(n)$, on est dans le deuxième cas, et $C(n) = \Theta(n^{\log_2 2} \log n)$. □

Pour améliorer l'algorithme précédent, on va se contenter dans un premier temps de mettre au point un algorithme possédant la propriété suivante :

- soit l'algorithme garantit que E ne possède pas d'élément majoritaire,
- soit l'algorithme fournit un entier $p > n/2$ et un élément x tels que x apparaisse au plus p fois dans E et tout élément autre que x apparaît au plus $n - p$ fois dans E .

Question 2.2 Donner un algorithme récursif possédant cette propriété. Quelle est sa complexité ?

Remarque : soit x vérifiant la propriété, alors aucun autre élément $z \neq x$ ne peut être majoritaire (car $c_z \leq n - p < n/2$), donc, dans ce cas, x est un candidat-majoritaire. Autrement dit un algorithme vérifiant les propriétés de l'exercice nous fournit :

- soit la garantie qu'il n'existe aucun élément majoritaire,
- soit un candidat-majoritaire x , dont on peut vérifier ensuite s'il est vraiment majoritaire en parcourant une fois le tableau (soit $n - 1$ comparaisons).

Remarque : un élément majoritaire x est bien un candidat-majoritaire avec $p = c_x$ (donc s'il existe un majoritaire, il sera bien proposé comme candidat-majoritaire par l'algorithme).

Solution : Algorithme récursif *Candidat – majoritaire*(E) renvoyant AUCUN s'il n'y a pas d'élément majoritaire dans E et sinon renvoyant (x, p) avec les bonnes propriétés.

Algorithme 3 *Candidat – majoritaire*(E)

```
si  $E$  n'a qu'un élément  $x$  alors
  retourner  $(x, 1)$ 
sinon
  couper  $E$  en deux tableaux  $E_1$  et  $E_2$  de taille  $n/2$ 
  appeler Candidat – majoritaire( $E_1$ ) qui renvoie AUCUN ou  $(x, p)$ 
  appeler Candidat – majoritaire( $E_2$ ) qui renvoie AUCUN ou  $(y, q)$ 
  suivant la réponse pour  $E_1$  et  $E_2$ , faire
  si AUCUN et AUCUN alors
    retourner AUCUN
  si AUCUN et  $(y, q)$  alors
    retourner  $(y, q + \frac{n}{4})$ 
  si  $(x, p)$  et AUCUN alors
    retourner  $(x, p + \frac{n}{4})$ 
  si  $(x, p)$  et  $(y, q)$  alors
    si  $x \neq y$  alors
      si  $p > q$  alors
        retourner  $(x, p + \frac{n}{2} - q)$ 
      si  $p < q$  alors
        retourner  $(y, q + \frac{n}{2} - p)$ 
      si  $p = q$  alors
        retourner AUCUN // (sinon ce serait  $x$  ou  $y$  mais  $c_x \leq \frac{n}{2}$  et  $c_y \leq \frac{n}{2}$ )
    sinon //  $x = y$ 
      retourner  $(x, p + q)$ 
```

Complexité : on a supposé n une puissance de 2, le nombre de comparaisons $C(n)$ pour un tableau à n éléments vérifie : $C(n) = 2C(\frac{n}{2}) + 1$ avec $C(1) = 0$, ce qui donne $C(n) = n - 1$. \square

Question 2.3 En déduire un algorithme efficace vérifiant si E possède un élément majoritaire.

Solution : Algorithme complet pour rechercher un majoritaire :

- Recherche d'un candidat-majoritaire avec l'algorithme précédent, soit $n - 1$ comparaisons.
- Si l'algorithme renvoie un candidat, vérification que le candidat est bien majoritaire en parcourant le tableau, soit $n - 1$ comparaisons.

Complexité au total = $2n - 2$ comparaisons. \square

3 Algorithme des balles

On change la manière de voir les choses. On a un ensemble de n balles et on cherche le cas échéant s'il y a une couleur majoritaire parmi les balles.

Question 3.1 Supposons que les balles soient rangées en file sur une étagère, de manière à n'avoir jamais deux balles de la même couleur à côté. Que peut-on en déduire sur le nombre maximal de balles de la même couleur ?

Solution : Soit k le nombre de balles rangées en file sur l'étagère telles que deux balles consécutives n'aient jamais la même couleur, alors le nombre de balles d'un même couleur est toujours $\leq \lceil \frac{k}{2} \rceil$.

Preuve précise : utiliser le lemme des tiroirs, c'est à dire "si M chaussettes sont rangées parmi m tiroirs avec $M > m$, alors il existe un tiroir avec au moins deux chaussettes". Ici découper l'étagère en $\lceil \frac{k}{2} \rceil$ tiroirs disjoints qui sont des paires d'emplacements consécutifs sur l'étagère. S'il y a plus de $\lceil \frac{k}{2} \rceil$ balles d'une même couleur, il en existe deux dans un même tiroir, c'est à dire ici adjacentes sur l'étagère. \square

On a un ensemble de n balles, une étagère vide où on peut les ranger en file et une corbeille vide. Considérons l'algorithme suivant :

- **Phase 1** - Prendre les balles une par une pour les ranger sur l'étagère ou dans la corbeille. **SI** la balle n'est pas de la même couleur que la dernière balle sur l'étagère, la ranger à côté, et si de plus la corbeille n'est pas vide, prendre une balle dans la corbeille et la ranger à côté sur l'étagère. **SINON**, c'est à dire si la balle est de la même couleur que la dernière balle sur l'étagère, la mettre dans la corbeille.

- **Phase 2** - Soit C la couleur de la dernière balle sur l'étagère à la fin de la phase 1. On compare successivement la couleur de la dernière balle sur l'étagère avec C . **SI** la couleur est la même on jette les deux dernières balles posées sur l'étagère, sauf s'il n'en reste qu'une, auquel cas on la met dans la corbeille. **SINON** on la jette et on jette une des balles de la corbeille, sauf si la corbeille est déjà vide auquel cas on s'arrête en décrétant qu'il n'y a pas de couleur majoritaire. Quand on a épuisé toutes les balles sur l'étagère, on regarde le contenu de la corbeille. Si elle est vide alors il n'y a pas de couleur majoritaire, et si elle contient au moins une balle alors C est la couleur majoritaire.

Question 3.2 (*Correction de l'algorithme*) Montrer qu'à tout moment de la phase 1, toutes les balles éventuellement présentes dans la corbeille ont la couleur de la dernière balle de l'étagère. En déduire que s'il y a une couleur dominante alors c'est C . Prouver la correction de l'algorithme.

Solution : *Correction de l'algorithme : on montre des invariants (propriétés restants vraies pendant une phase de l'algorithme).*

- **Phase 1** -

Invariant 1 : *si la corbeille n'est pas vide, toutes les balles dans la corbeille ont la même couleur que la dernière balle sur l'étagère.*

Preuve : *montrer que si la prop. est vraie à un instant donné, elle reste vraie à l'étape suivante. Plusieurs cas peuvent se produire ici, vérifier la conservation de la propriété dans chaque cas (à faire) :*

- ajout d'une balle différente de la dernière et corbeille pleine ...
- ajout d'une balle différente de la dernière et corbeille vide ...
- ajout d'une balle identique à la dernière ...

Invariant 2 : *sur l'étagère, il n'y a jamais deux balles de même couleur côte à côte.*

Preuve : *idem au cas par cas.*

Fin de Phase 1 : *soit C la couleur de la dernière balle sur l'étagère. Les balles d'autres couleurs sont dans les $k-1$ premières balles de l'étagère, où k est le nombre de balles sur l'étagère et $k \leq n$. Comme il n'y a pas deux balles adjacentes de même couleur, d'après la question 1, le nombre de balles d'une couleur différente de C est $\leq \lceil \frac{n-1}{2} \rceil = \lfloor \frac{n}{2} \rfloor$. Donc la seule couleur candidate pour être majoritaire est C .*

- **Phase 2** - *vérifie si C est bien majoritaire :*

- *Quand on retire une paire de balles, on en a toujours une de couleur C et l'autre différente, donc C est majoritaire si et seulement si une majorité de balles à la fin (étagère + corbeille) est de couleur C .*
- *Deux cas de fin se présentent :*
 - *La phase 2 s'arrête car on a besoin d'une balle de la corbeille mais la corbeille est vide. Dans ce cas, la dernière balle sur l'étagère n'est pas de couleur C et d'après la question 1, au plus la moitié des balles restant sur l'étagère sont de couleur C , il n'y a pas de majoritaire. Ok, c'est bien ce que répond l'algorithme.*
 - *L'algorithme va jusqu'au bout de l'étagère, les balles restantes (s'il en reste) sont dans la corbeille, elles sont toutes de couleur C . Donc C est majoritaire si et seulement si la corbeille est non vide, c'est bien le test effectué par l'algorithme.*

□

Question 3.3 (*Complexité*) Donner la complexité dans le pire des cas en nombre de comparaisons de couleurs des balles.

Solution : Complexité :

Phase 1 : $(n-1)$ comparaisons (-1 car la première balle est directement posée sur l'étagère).

Phase 2 : une comparaison par paire de balles éliminées, plus éventuellement une comparaison avec C pour la dernière balle de l'étagère s'il en reste une unique, soit $\leq \lfloor \frac{n}{2} \rfloor$ comparaisons (-1 car au début de la phase 2, on sait que la dernière balle sur l'étagère est de couleur C).

Complexité totale $\leq \lceil \frac{3n}{2} \rceil$ comparaisons. □

4 Optimalité

On considère un algorithme de majorité pour les couleurs de n balles. Le but est de regarder faire l'algorithme, en choisissant au fur et à mesure les couleurs des balles pour que l'algorithme ait le maximum de travail (tout en restant cohérent dans le choix des couleurs). On obtiendra ainsi une borne inférieure de complexité pour un algorithme de majorité. C'est la technique de l'*adversaire*.

À tout moment de l'algorithme, on aura une partition des balles en deux ensembles : l'*arène* et les *gradins*. L'arène contient un certain nombre de composantes connexes de deux sortes : les *binômes* et les *troupeaux*. Un binôme est un ensemble de deux balles pour lesquelles l'algorithme a déjà testé si elles étaient de la même couleur et a répondu non. Un troupeau est un ensemble non vide de balles de la même couleur, connectées par des tests de l'algorithme. Ainsi, un troupeau avec k éléments a subi au moins $k - 1$ comparaisons de couleurs entre ses membres. Soient B le nombre de binômes et T le nombre de troupeaux. Soient g le nombre d'éléments dans les gradins et t le nombre total d'éléments dans tous les troupeaux. Enfin, soit $m = \lfloor n/2 \rfloor + 1$ le "*seuil de majorité*".

Au début de l'algorithme toutes les balles sont des troupeaux à un élément. La stratégie de l'adversaire est la suivante. L'algorithme effectue un test $\text{couleur}(x) = \text{couleur}(y)$?

1. Si x ou y sont dans les gradins, la réponse est non.
2. Si x (resp. y) est dans un binôme, la réponse est non et x (resp. y) est envoyé dans les gradins alors que l'élément restant devient un troupeau singleton.
3. Si x et y sont dans le même troupeau la réponse est oui.
4. Si x et y sont dans des troupeaux différents alors cela dépend de $d = B + t$.
 - $d > m$: cela signifie que les balles sont dans des troupeaux singletons. La réponse est non et les balles deviennent un nouveau binôme.
 - $d = m$: la réponse est oui et les troupeaux de x et y fusionnent.

Question 4.1 Vérifier que les quatre cas précédents traitent tous les cas possibles. Montrer qu'à tout moment $d \geq m$ et que si $d > m$ alors tous les troupeaux sont des singletons.

Solution : Tous les cas possibles sont bien représentés. Au départ $d = n$ et tous les troupeaux sont des singletons. Quels cas modifient d ?

Cas 1 : non.

Cas 2 : non (car -1 binôme, +1 dans les troupeaux).

Cas 3 : non.

Cas 4(a) : oui, d diminue de 1 (mais les troupeaux restent des singletons).

Cas 4(b) : non.

Conclusion :

- Invariant $d \geq m$ (si d change, c'est que $d > m$ avec le cas 4(a) et alors $d - 1 \geq m$).
 - $d > m$ implique que tous les troupeaux sont des singletons.
-

Question 4.2 Montrer qu'à tout moment les deux coloriage suivants sont consistants avec les réponses de l'adversaire :

- (i) Toutes les balles sont de couleurs différentes sauf celles qui sont dans un même troupeau.

(ii) Une même couleur est attribuée à toutes les balles de tous les troupeaux et à une balle de chaque binôme. Les balles restantes ont chacune une couleur distincte.

Solution : *Consistence des réponses de l'adversaire à tout moment.*

Coloration (i) : correspond bien aux réponses données, car respecte les binômes (comparaisons inégales) et les troupeaux (comparaisons égales). De plus les balles envoyées dans les gradins ont toujours engendré la réponse NON aux comparaisons, ce qui fonctionne si leurs couleurs sont toutes différentes et différentes des couleurs des balles dans l'arène.

Coloration (ii) : idem avec de plus le fait qu'une balle dans un binôme n'a jamais été comparée à quelqu'un d'autre que son binôme (donc pas de risque de contradiction avec les choix des couleurs ailleurs). □

Question 4.3 Montrer que si un algorithme correct s'arrête alors l'arène ne contient qu'une seule composante connexe qui est un troupeau de taille m .

Solution : *Un algorithme correct ne peut s'arrêter que si l'arène ne contient qu'une composante connexe qui sera un troupeau de taille m .*

Preuve par l'absurde : *supposons que l'arène contienne au moins deux composantes (ce qui implique $n \geq 2$ et $m \geq 2$). Par définition de d , chaque troupeau contient $< d$ balles. Chaque troupeau contient donc $< m$ balles, car soit $d = m$, soit $d > m$ et d'après la question 1 chaque troupeau est un singleton. La coloration (i) n'a pas d'élément majoritaire, mais la coloration (ii) a un élément majoritaire (car $d \geq m$). Donc aucun algorithme correct ne peut s'arrêter à cet instant.*

Conclusion : *à la fin, il y a nécessairement une unique composante, nécessairement de taille $d = m$ (par définition de d et d'après la question 1 qui implique que $d \not\geq m$).* □

Question 4.4 À tout moment le nombre de comparaisons inégales effectuées par l'algorithme est au moins $2g + B$ et le nombre de comparaisons égales est au moins $t - T$.

Solution : *A tout moment, on a eu jusqu'à présent :*

- Nombre de comparaisons inégales $\geq 2g + B$, car il faut 2 comparaisons inégales pour mettre une balle dans les gradins (entrée dans un binôme puis sortie du binôme) et il y a B comparaisons inégales qui ont construit les binômes en place.
- Nombre de comparaisons égales $\geq t - T$, car un troupeau i avec t_i balles est connexe donc il a $\geq t_i - 1$ arêtes, soit ici $t_i - 1$ comparaisons égales.

□

Question 4.5 Considérons un algorithme qui résout la majorité. Montrer qu'il existe une donnée pour laquelle l'algorithme effectue au moins $2(n - m) = 2\lceil n/2 \rceil - 1$ comparaisons d'inégalité et au moins $\lfloor n/2 \rfloor$ comparaisons d'égalité, et donc au moins au total $3\lceil n/2 \rceil - 2$ comparaisons.

Solution : *En fin d'algorithme, d'après la question 3, $g = n - m$, $B = 0$, $t = m$ et $T = 1$. Donc, avec la question 4, il y a eu $\geq 2(n - m) = 2\lceil \frac{n}{2} \rceil - 2$ comparaisons inégales et $\geq m - 1 = \lfloor \frac{n}{2} \rfloor$ comparaisons égales, soit au total $\geq \lceil \frac{3n}{2} \rceil$ comparaisons.*

□

