

## Algorithmes Probabilistes – Devoir à la maison 2008

Responsable : Alexandre Zvonkine

9 novembre 2009

Rendre les copies avant la fin de l'année

## Recherche de la médiane d'un ensemble

Soit  $R$  un ensemble dont les éléments appartiennent à un univers totalement ordonné : cela veut dire que les éléments de  $R$  sont deux à deux comparables, et pour deux éléments distincts on peut toujours dire lequel de deux est plus petit et lequel est plus grand. Sauf que les éléments de  $R$  sont donnés sous la forme d'une liste désordonnée. . .

Pour simplifier les considérations qui suivent nous supposons que :

- tous les éléments de  $R$  sont distincts ;
- le nombre d'éléments  $|R| = n$  est impair.

**Définition (Médiane)** La *médiane* de l'ensemble  $R$  est un élément de  $R$ , noté  $M(R)$ , tel que dans  $R$  il y ait exactement  $(n-1)/2$  éléments plus petits que  $M(R)$  et  $(n-1)/2$  éléments plus grands que  $M(R)$ .

**Exercice 1** Proposer un algorithme de complexité  $O(n \log n)$  qui calcule la médiane de l'ensemble  $R$  de taille  $|R| = n$ .

Il existe un algorithme de complexité linéaire qui résout ce problème mais il est compliqué. Dans ce devoir, nous proposons un algorithme probabiliste, de complexité linéaire, beaucoup plus simple que l'algorithme déterministe et qui possède, en plus, plus petite constante multiplicative.

L'algorithme en question est décrit sur la page 2. Sa complexité, ainsi que la probabilité d'ÉCHEC, dépendent de trois paramètres :  $m$  (utilisé à l'étape 1),  $k$  (utilisé à l'étape 2), et  $t$  (utilisé à l'étape 4). Les valeurs de ces paramètres seront précisées plus tard, en exercice 12.

**Exercice 2** Expliquer le fonctionnement de l'algorithme. Entre autres choses, expliquer pourquoi cet algorithme, s'il arrive jusqu'au bout et ne s'arrête pas au milieu en renvoyant ÉCHEC, donne toujours un bon résultat (c'est-à-dire, la vraie médiane de l'ensemble  $R$ ).

---

### Algorithme : recherche de la médiane d'un ensemble

1. Choisir au hasard un *multi-ensemble*  $S$  de taille  $|S| = m$  d'éléments de  $R$ . Les éléments sont choisis uniformément, indépendamment l'un de l'autre, et sont remis dans  $R$  (ils peuvent donc se répéter dans  $S$ ).
  2. Trier  $S$ . Poser  $a$  égal à l'élément numéro  $m/2 - k$  en ordre croissant dans  $S$ , et  $b$  égal à l'élément numéro  $m/2 + k$  en ordre croissant dans  $S$ .
  3. Passer en revue tous les éléments de  $R$  et choisir le sous-ensemble  $T = \{x \in R \mid a \leq x \leq b\}$ . En même temps, calculer le nombre  $t_1$  d'éléments  $x < a$  dans  $R$ , le nombre  $t_2 = |T|$  d'éléments  $a \leq x \leq b$  dans  $R$ , et le nombre  $t_3$  d'éléments  $x > b$  dans  $R$ .
  4. Si  $t_1 > n/2$ , ou  $t_3 > n/2$ , ou  $t_2 > t$ , alors retourner ÉCHEC et s'arrêter. Sinon, continuer : passer à l'étape 5.
  5. Trier  $T$ . Retourner un élément de  $T$  numéro  $(n + 1)/2 - t_1$  en ordre croissant : c'est la médiane  $M(R)$ .
- 

**Exercice 3** Pourquoi à l'étape 1 on fait le tirage au sort avec la remise des éléments choisis dans l'ensemble? Ne serait-il pas plus raisonnable de faire le tirage au sort sans remise?

**Exercice 4** Estimer la complexité de cet algorithme en fonction des paramètres  $n, m, k, t$ . Expliquer pourquoi, si l'on veut obtenir une complexité linéaire, il faut prendre  $m \ll n$  et  $t \ll n$ .

Il nous faut maintenant estimer la probabilité d'ÉCHEC.

**Exercice 5** Montrer que la condition  $t_1 > n/2$  est équivalente à la condition  $|\{x \in S \mid x \leq M(R)\}| < m/2 - k$  et, d'une manière symétrique, la condition  $t_3 > n/2$  est équivalente à la condition  $|\{x \in S \mid x \geq M(R)\}| < m/2 - k$ .

**Exercice 6** Soit  $X_i$  une variable aléatoire égale à 1 si le  $i$ -ème élément de  $R$  choisi au hasard est  $\leq M(R)$ . Soit  $X = \sum_{i=1}^m X_i$ , la somme de  $m$  variables indépendantes  $X_i$ .

1. Montrer que la condition  $t_1 > n/2$  est équivalente à la condition  $X < m/2 - k$ , et la condition  $t_3 > n/2$  est équivalente à la condition  $X > m/2 + k$ . (Utiliser l'exercice 5.)

- Calculer l'espérance et la variance de  $X_i$ , puis celles de  $X$ . (Pour la variance, ce n'est pas une expression exacte mais un majorant qui nous intéresse.)

**Remarque** Dans l'exercice précédent, les probabilités que  $X_i$  soient égales à 1 ou à 0 peuvent être calculées exactement, mais après avoir fait ce calcul on peut utiliser leurs valeurs approchées.

**Exercice 7 (Inégalité de Chebyshev)** Soit une variable aléatoire  $Z$ . Notons  $E(Z) = \bar{z}$ ,  $\text{Var}(Z) = \sigma^2$ . Montrer que pour toute constante  $c > 0$  on a

$$P(|Z - \bar{z}| > c \cdot \sigma) \leq 1/c^2.$$

**Exercice 8** En utilisant l'inégalité de Chebyshev, montrer que la probabilité de l'événement " $t_1 > n/2$  ou  $t_3 > n/2$ " est majorée par  $m/4k^2$ .

Considérons maintenant la troisième condition d'ÉCHEC. Si l'ensemble  $T$  contient plus de  $t$  éléments, alors au moins un des deux événements suivants se réalise :

- $A$  : il y a au moins  $t/2$  éléments dans  $T$  qui sont supérieurs à la médiane ;
- $B$  : il y a au moins  $t/2$  éléments dans  $T$  qui sont inférieurs à la médiane.

Nous procédons à l'estimation de la probabilité de  $A$  ; il est clair que la probabilité de  $B$  sera la même par symétrie.

**Exercice 9** Montrer que la condition  $A$  est équivalente à la condition suivante : au moins  $m/2 - k$  éléments de l'échantillon  $S$  appartiennent à l'ensemble de  $(n - t)/2$  plus grands éléments de  $R$ .

**Exercice 10** Soit  $Y_i$  une variable aléatoire égale à 1 si le  $i$ -ème élément de  $R$  choisi au hasard appartient à l'ensemble de  $(n - t)/2$  plus grands éléments de  $R$ . Soit  $Y = \sum_{i=1}^m Y_i$ , la somme de  $m$  variables indépendantes  $Y_i$ .

- Montrer que la condition  $A$  est équivalente à la condition  $Y \geq m/2 - k$ . (Utiliser l'exercice 9.)
- Calculer l'espérance et la variance de  $Y_i$ , puis celles de  $Y$ . (Pour la variance, ce n'est pas une expression exacte mais un majorant qui nous intéresse.)

**Exercice 11** En utilisant l'inégalité de Chebyshev, montrer que la probabilité de l'événement  $A$  est majorée par  $mn^2/(mt - 2nk)^2$ .

Un choix judicieux des paramètres  $m, k, t$  est un choix pour lequel les probabilités trouvées dans les exercices 8 et 11 seraient du même ordre de grandeur.

**Exercice 12** Montrer que si  $m = n^{3/4}$ ,  $k = n^{1/2}$ ,  $t = 4n^{3/4}$ , alors la probabilité de l'ÉCHEC est majorée par  $n^{-1/4}$ . (Ignorez le problème d'arrondi des nombres non entiers.)

**Exercice 13** Expliquer pourquoi la complexité de l'algorithme, pour les valeurs de paramètres choisies dans l'exercice 12, est linéaire.

La possibilité d'obtenir un ÉCHEC comme une seule réponse, bien que peu probable, reste gênante. Pour remédier à cette situation, on peut répéter l'algorithme plusieurs fois jusqu'au premier succès.

**Exercice 14 (Loi géométrique)** On fait des essais aléatoires avec la probabilité de succès  $p$  et la probabilité d'échec  $q = 1 - p$ , et soit  $H$  le nombre d'essais jusqu'au premier succès.

1. Trouver les probabilités  $P(H = k)$  pour  $k = 1, 2, 3, \dots$
2. Trouver l'espérance  $E(H)$ .

**Exercice 15** Montrer que pour notre algorithme, si  $n \geq 16$  alors le nombre moyen de répétitions de l'algorithme jusqu'au premier succès est majoré par 2.

FIN DE L'ÉNONCÉ