Tri rapide aléatoire

Léo Gayral

2017-2018

ref : Beauquier – Éléments d'algorithmique – p.33

Lemme 1. On considère l'algorithme de tri rapide en place :

```
def TRI(L) :
    si |L| ≥ 2 :
        i := PIVOT(L)
        L[0], L[i] = L[i], L[0]
        i = 0
        j = |L|
        tant que j > i + 1 :
            si L[i + 1] < L[i] :
                L[i], L[i + 1] = L[i + 1], L[i]
                i+ = 1
            sinon :
                j− = 1
                L[j], L[i + 1] = L[i + 1], L[j]
        TRI(L[: i])
        TRI(L[j :])
```

Pour tout choix de $PIVOT$, qui à L associe l'indice d'un de ses éléments, l'algorithme termine en $O\left(|L|^2\right)$ dans le pire des cas.

Théorème 1. On considère le cas du tri rapide randomisé, où $PIVOT(L) \sim \mathcal{U}\left(\{0, \dots, |L| - 1\}\right)$ renvoie un indice choisi uniformément au hasard, de sorte que les appels de $PIVOT$ soient indépendants dans leur ensemble.

On définit la variable aléatoire $C(L)$ qui désigne le nombre de comparaisons entre deux éléments de L effectuées au cours de l'exécution du tri rapide. On a alors :

$$\mathbb{E}\left[C(L)\right] \sim 2n \ln(n)\,.$$

Démonstration.
L'hypothèse d'indépendance des pivots garantit que la loi de $C(L)$ est entièrement déterminée par $n = |L|$, et on la note alors C_n. Le résultat est clair

pour $|L| \leq 1$, où aucune comparaison n'est faite. Supposons le résultat vrai jusqu'au rang $n-1$.

Notons alors $G = L[:i]$ et $D = L[j:]$ les listes *aléatoires*, après l'exécution du *tant que* mais avant les appels récursifs à TRI. On a $C(L) = n - 1 + C(G) + C(D)$ par définition de l'algorithme TRI.

Considérons de plus $P = |G| \in [\![0, n-1]\!]$, qui correspond à la position de l'élément choisi par $PIVOT(L)$ dans la liste après le *tant que* ; on a donc $P \sim \mathcal{U}$.

Conditionnellement à l'évènement $P = k$, les listes G et D sont constantes. Les variables $C(G)$ et $C(D)$, qui comptent le nombre de comparaisons pour effectuer $TRI(G)$ et $TRI(D)$, sont mesurables selon deux ensembles de pivots indépendants par hypothèse de modélisation ; ces variables sont donc indépendantes, de lois C_k et C_{n-k-1} par hypothèse de récurrence.

On en déduit $C(L) \overset{d}{=} n - 1 + C_P + C_{n-P-1}$, entièrement déterminé par $|L| = n$. En particulier :

$$c_n := \mathbb{E}[C_n] = (n-1) + \mathbb{E}\left[C_P + C_{n-P-1}\right] = n - 1 + \frac{2}{n}\sum_{k=0}^{n-1} c_k \,.$$

Quitte à multiplier par n, on a :

$$\begin{aligned}(n+1) \times c_{n+1} &= n(n+1) + 2\sum_{k=0}^{n} c_k \,, \\ &= 2n + 2c_n + n(n-1) + 2\sum_{k=0}^{n-1} c_k \,, \\ &= 2n + (n+2)c_n \,.\end{aligned}$$

En divisant par $(n+1)(n+2)$ on obtient $\frac{c_{n+1}}{n+2} - \frac{c_n}{n+1} = 2\frac{n}{(n+1)(n+2)} \sim \frac{2}{n}$ qui est le terme général d'une série positive divergente. Les sommes partielles sont équivalentes, donc par télescopage :

$$\frac{c_n}{n} \sim \frac{c_n}{n+1} \sim 2\sum_{k=1}^{n} \frac{1}{k} \sim 2\ln(n)$$

d'où le résultat voulu, $c_n \sim 2n\ln(n)$.

□

Remarque 1. Par des arguments d'arbres de décision, on sait que la limite théorique inférieure – en termes de complexité moyenne d'un algorithme de

tri déterministe – est $n \log_2(n)$, autrement dit la constante optimale devant $n \ln(n)$ est $\frac{1}{\ln(2)} \approx 1,44$, donc l'algorithme précédent est assez proche de la limite théorique en pratique.

On peut également montrer que $\mathrm{Var}\left(C_n\right) = \theta\left(n^2\right)$, ce qui garantit non seulement des bonnes performances en moyenne, mais également un faible écart-type, donc des performances relativement fiables en moyenne.