確率・統計（真貝）第２回　参考資料
02_PS_quicksort.pdf

代表的なソート法

n 個のデータを並べ替えるとき（例えば番号昇順），適したソート・探索アルゴリズムが必要になる [1].

1　代表的なソート法

1.1　バブルソート

最も単純で最も遅い.

- データをはじめから順に，隣り合う数を昇順に入れ替えていく（操作数 n）
- 以下くりかえし

はじめ	4	6	3	2	1	5
１回目１	4	6	3	2	1	5
１回目２	4	3	6	2	1	5
１回目３	4	3	2	6	1	5
１回目４	4	3	2	1	6	5
１回目５	4	3	2	1	5	6

全操作数は，$O(n^2)$

1.2　選択ソート

- n 個のデータから最小を探して取り出す（操作数 n）
- $n-1$ 個のデータから最小を探して取り出して先の列に加える（操作数 $n-1$）
- 以下くりかえし

はじめ	4	6	3	2	1	5		
１回目	4	6	3	2	5		1	
２回目	4	3	6	5			1	2
３回目	4	6	5				1	2	3

全操作数は,

$$n+(n-1)+\cdots+2+1=\sum_{k=1}^{n}k$$
$$=\frac{n(n+1)}{2}\sim O(n^2) \qquad (1)$$

1.3　挿入ソート

- はじめの２つのデータを昇順に（操作数 1）
- はじめの３つのデータを昇順に（操作数 2）
- 以下くりかえし

はじめ	4	6	3	2	1	5
１回目	{4	6}	3	2	1	5
２回目	{3	4	6}	2	1	5
３回目	{2	3	4	6}	1	5

全操作数は,

$$1+2+\cdots+(n-1)=\sum_{k=1}^{n-1}k$$
$$=\frac{n(n-1)}{2}\sim O(n^2) \qquad (2)$$

1.4　クィックソート

- データの先頭数を基準に上下２分
- 各グループの先頭数を基準に上下２分
- 以下くりかえし

はじめ	4	6	3	2	1	5
１回目	{3	2	1},	{4},	{6	5}
２回目	{2	1}{3}{},	{4},	{5}{6}		
３回目	{1}{2}{},	{3},	{4},	{5}{6}		

全操作数は，$O(n\log n)$．導出次ページ.
ただし，はじめから逆順にソートされているデータだと，最悪 n^2 ステップ.

2　クィックソートの平均操作数

教科書 [2]p37 コラム６で紹介した，クィックソートについて，n 個のものをクィックソートするときの平均操作数 Q_n を求めよう．

n 個のものに対して，初めの１つを基準に，残りの $n-1$ 個を２つに分けてゆく．残りが s 個と $n-1-s$ 個に分けられるとすると，

$$Q_n = n-1+Q_s+Q_{n-1-s} \tag{3}$$

と書くことができる．２つの分けかたは，$\{Q_0, Q_{n-1}\}$, $\{Q_1, Q_{n-2}\}, \cdots, \{Q_{n-1}, Q_0\}$ の n パターンあって，そのどれもが等確率 $1/n$ で得られるとすれば，平均操作数 Q_n は

$$\begin{aligned} Q_n &= n-1+\frac{1}{n}\sum_{s=0}^{n-1}(Q_s+Q_{n-1-s}) \\ &= n-1+\frac{2}{n}\sum_{s=0}^{n-1}Q_s \end{aligned} \tag{4}$$

添え字を１減らしたものについては

$$Q_{n-1} = n-2+\frac{2}{n-1}\sum_{s=0}^{n-2}Q_s \tag{5}$$

となるので，$n\times(4)$ $-(n-1)\times(5)$ より

$$\begin{aligned} nQ_n-(n-1)Q_{n-1} &= n(n-1)-(n-1)(n-2) \\ &\quad -2Q_{n-1} \\ nQ_n-(n+1)Q_{n-1} &= 2(n-1) \end{aligned}$$

となる．両辺を $n(n+1)$ で割ると，

$$\frac{Q_n}{n+1}-\frac{Q_{n-1}}{n} = \frac{2(n-1)}{n(n+1)} \tag{6}$$

となる．ここで，$P_n \equiv \dfrac{Q_n}{n+1}$ と新たに置くと，この式は

$$P_n - P_{n-1} = \frac{2(n-1)}{n(n+1)} \tag{7}$$

となって，P_n についての漸化式になる．なお，$P_0 = Q_0 = 0$ である．(7) の上限値を考えて，

$$P_n - P_{n-1} = \frac{2(n-1)}{n(n+1)} < \frac{2n}{n(n+1)} = \frac{2}{n+1} \tag{8}$$

として，１つずつ添え字を減らした式を並べ，

$$\begin{aligned} P_n - P_{n-1} &< \frac{2}{(n+1)} \\ P_{n-1} - P_{n-2} &< \frac{2}{n} \\ \vdots &\quad \vdots \\ P_1 - P_0 &< 2 \end{aligned}$$

これらの和をとると，

$$P_n - P_0 < 2\sum_{k=1}^{n+1}\frac{1}{k} \tag{9}$$

となる．右辺の和の式は，n を大きくすれば積分で書けて

$$\begin{aligned} \sum_{k=1}^{n+1}\frac{1}{k} &= \int_1^{n+1}\frac{1}{x}dx = [\log x]_{x=1}^{x=n+1} \\ &= \log(n+1) \end{aligned}$$

となることから，(9) は，

$$P_n - 0 < 2\log(n+1) \tag{10}$$

すなわち

$$\frac{Q_n}{n+1} < 2\log(n+1)$$

となる．したがって，

$$Q_n < 2(n+1)\log(n+1)$$

となって，

$$Q_n \sim n\log n \tag{11}$$

となる．

参考文献

[1] I. Ahmad 著，株式会社クイープ訳『プログラマーなら知っておきたい 40 のアルゴリズム』（インプレス，2021）

[2] 真貝寿明『徹底攻略 確率統計』（共立出版，2012）