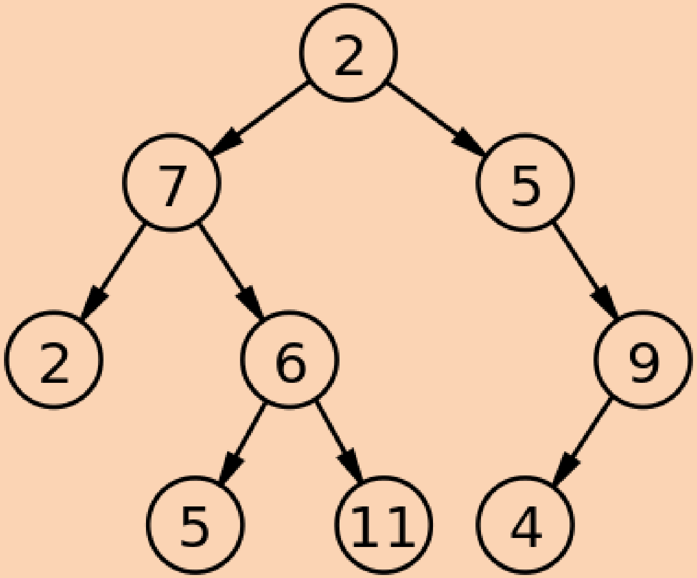
ساختمان‌های داده

محاسبه تصادفی آماره ترتیبی

Randomized Computation of Order Statistic



مدرس:

سیدکمال‌الدین غیاثی شیرازی

مساله

i 1 2 3 4 5 6

1, 2, 5, 6, 7, 12

$A = \underbrace{[1, 12, 6, 7, 5, 2]}_{n}$

- **ورودی:** آرایه ای با n مقدار مجزا و یک عدد i بین 1 تا n
- **خروجی:** i امین آماره‌ی ترتیبی (i امین عدد کوچک آرایه)

- **توجه:** فرض مجزا بودن مقادیر آرایه فقط برای سادگی تحلیل است و الگوریتم برای داده‌های تکراری نیز به درستی عمل می‌کند.

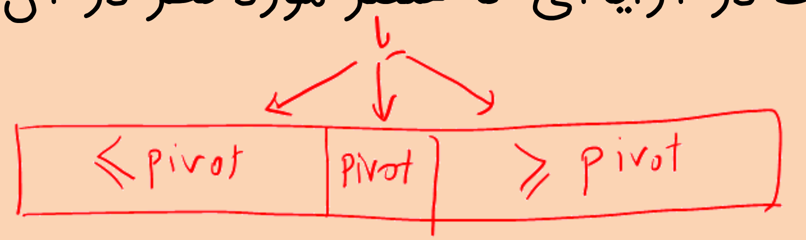
راه حل مبتنی بر مرتب سازی

- الگوریتم:
  - آرایه را مرتب کن
  - i امین عنصر آرایه مرتب شده را برگردان

- زمان اجرا $\Theta(n \log n)$

راه حل با ایده‌ی افراز حول محور

- عنصری از آرایه به عنوان محور انتخاب می شود.
- مانند الگوریتم مرتب‌سازی سریع ، آرایه به دو قسمت افراز می‌شود و محور در محل خود قرار می گیرد.
- برخلاف الگوریتم مرتب‌سازی سریع که عملیات را در هر دو آرایه چپ و راست ادامه می دهد ، در اینجا عملیات فقط در آرایه‌ای که عنصر مورد نظر در آن است ادامه پیدا می‌کند.

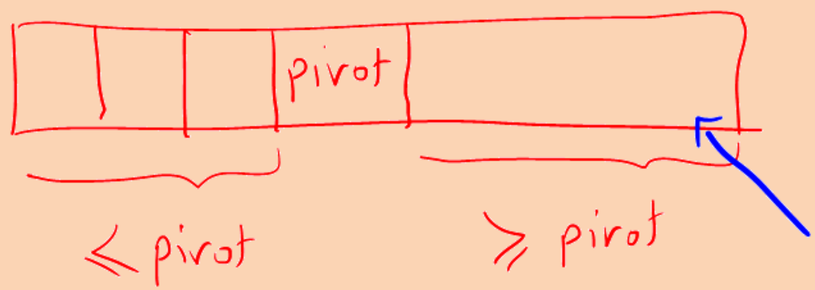


سوال کلاسی

$i = 7$ $n = 10$

- فرض کنید که در آرایه ای به طول ۱۰ به دنبال ۷ امین آماره ترتیبی می‌گردیم.
- آرایه را افراز می‌کنیم و محور در ۴ امین خانه آرایه افراز شده قرار می‌گیرد.
- مقدار موردنظر در کدام سمت محور قرار دارد و باید به دنبال چندمین آماره ترتیبی باشیم؟

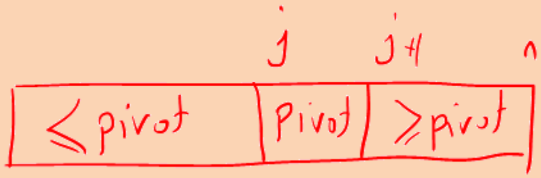
۴ عنصر ابتدای آرایه یکی از ترتیب ج



3 امین آماره ترتیب

Randomized Selection الگوریتم

```
Rselect(Array A, length n, order statistic i)
   if i < 1 Or i > n:
      Error
   if n == 1:
pivot return A[1]
   p = ChoosePivot(A, n)
   j = Partition(A, p)
   if j == i:
      return p یا A[j] یا A[i]
   elif j > i:     i < j
      return Rselect(1st part of A, j-1 , i )
   else:   i > j
      return Rselect(2nd part of A, n-j , i-j )
```



Randomized Selection الگوریتم

```
Rselect(Array A, length n, order statistic i)
   if i < 1 Or i > n:
      Error
   if n == 1:
      return A[1]
   p = ChoosePivot(A, n)
   j = Partition(A, p)
   if j == i:
      return A[i]
   elif j > i:
      return Rselect(1st part of A, j-1 , i )
   else:
      return Rselect(2nd part of A, n-j , i-j)
```

تحلیل زمان اجرا در بدترین حالت

- اگر به دنبال n امین آماره‌ی ترتیبی باشیم و در هر فراخوانی $Rselect$ کوچک‌ترین عنصر به عنوان محور انتخاب شود، آنگاه اندازه‌ی آرایه در فراخوانی بعدی تنها یک واحد از فراخوانی فعلی کمتر است.

- n فراخوانی لازم است که آماره مورد نظر به دست آید.

$n + (n-1) + (n-2) + \dots + 1$

- در فراخوانی i ام طول آرایه $n - i + 1$ است.

$= \frac{n(n+1)}{2} \in \theta(n^2)$

- بنابراین زمان اجرای الگوریتم در بدترین حالت از مرتبه $\Theta(n^2)$ است.

زمان اجرای متوسط

- برای هر آرایه دلخواه ورودی به اندازه n، متوسط زمان اجرای الگوریتم $Rselect$ برابر $\Theta(n)$ است.

گام اول اثبات: ردیابی پیشرفت با استفاده از فازبندی

- **مشاهده اول:** عملیاتی که درون تابع RSelect پیش از فراخوانی بازگشتی انجام می شود از مرتبه n است و بنابراین از cn کوچک تر است (برای یک مقدار ثابت $c > 0$).

- **فازبندی:** می گوییم RSelect در فاز j ام است هرگاه طول کنونی آرایه بین $n \left(\frac{3}{4}\right)^{(j+1)}$ و $\left(\frac{3}{4}\right)^{j} n$ باشد.

- **متغیر تصادفی X_j:** تعداد فراخوانی های بازگشتی در فاز j

$$Running\ time\ of\ RSelect \leq \sum_{phases\ j} X_j . c. \left(\frac{3}{4}\right)^{j} n$$

گام دوم اثبات: ساده سازی به پرتاب سکه

- **مشاهده:** اگر RSelect محوری انتخاب کند که داده ها را به دو دسته 25% و 75% و یا بهتر (یعنی نزدیک تر به پنجاه پنجاه) تقسیم کند، آنگاه فاز جاری تمام می شود زیرا طول آرایه جدید حداکثر ¾ طول آرایه فعلی خواهد بود.
- **مشاهده:** احتمال یک افراز 25% و 75% و یا بهتر 0.5 است.
- بنابراین امید ریاضی X_j برابر است با امید ریاضی تعداد دفعات لازم برای پرتاب یک سکه سالم تا اینکه بالاخره یک بار شیر بیاید.

گام سوم اثبات: تحلیل پرتاب سکه

- فرض کنیم Y متغیری تصادفی باشد که تعداد دفعات لازم برای انجام آزمایشی با امید موفقیت p برای یافتن اولین موفقیت را نشان دهد. این متغیر تصادفی دارای توزیع هندسی است.

$$P(Y=y)=(1-p)^{y-1}p$$

- همانطور که در انتهای بحث نشان خواهیم داد، میانگین توزیع هندسی با پارامتر p برابر $\frac{1}{p}$ است.

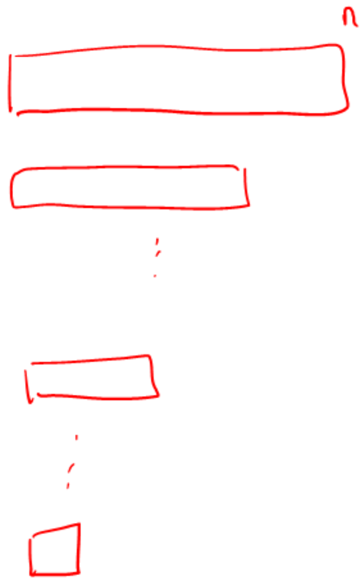
$$E(Y)=\sum_{y=1}^{\infty} y(1-p)^{y-1}p=\frac{1}{p}$$

گام چهارم اثبات: کنار هم گذاشتن نتایج قبلی

$$E(Runtime) \leq E\left[\sum_{phases\ j} X_j.\,c.\left(\frac{3}{4}\right)^j n\right]$$

$$= cn \sum_{phases\ j} E(X_j)\left(\frac{3}{4}\right)^j = 2cn \sum_{phases\ j}\left(\frac{3}{4}\right)^j$$

$$= 2cn \sum_{phases\ j}\left(\frac{3}{4}\right)^j \leq \frac{2cn}{1-\frac{3}{4}} = 8cn \in O(n)$$



تناهی گام

n — فاز صفرم

$\frac{3}{4} n$ — فاز اول

$(3/4)^2 n$

$(3/4)^j n$ — فاز jام

$(3/4)^{j+1} n$

X_j : تعداد دفعات باقی ماندن در فاز jام

در فاز jام حداکثر طول آرایه $(3/4)^j n$ $\Leftarrow$ حداکثر زمان اجرا برای یک بار باقی ماندن در فاز jام برابر $\Theta(n)$

زمان اجرای الگوریتم $\le \sum_{j=0}^{\infty} X_j \, c \left(\frac{3}{4}\right)^j n$ $\Leftarrow$ $\le cn$

در هر فاز j اندازه حداکثر آرایه $(3/4)^j n$ ⇐

زمان اجرای الگوریتم $< \sum_{j=0}^{\infty} X_j \, c \left(\frac{3}{4}\right)^j n$

توزیع هندسی

متوسط زمان اجرا $\leq cn \sum_{j=0}^{\infty} E[X_j] \left(\frac{3}{4}\right)^j \leq 2cn \sum_{j=0}^{\infty} \left(\frac{3}{4}\right)^j = 8cn \in \Theta(n)$

$p = 1/2$

$(3/4)^j n$

داده‌ها به صورت فرضی مرتب هستند

×	✓	✓	×

هر ✓ طول آرایه‌ای کوچکتر از $\underbrace{(3/4) \times (3/4)^j \times n}_{(3/4)^{j+1} n}$ فراهم می‌کند

$$P(X_j = k) = (1-p)^{k-1} p$$

$$E[X_j] = \frac{1}{p} = 2$$

محاسبه E(Y) روش اول: تبدیل به سیگمای دو گانه

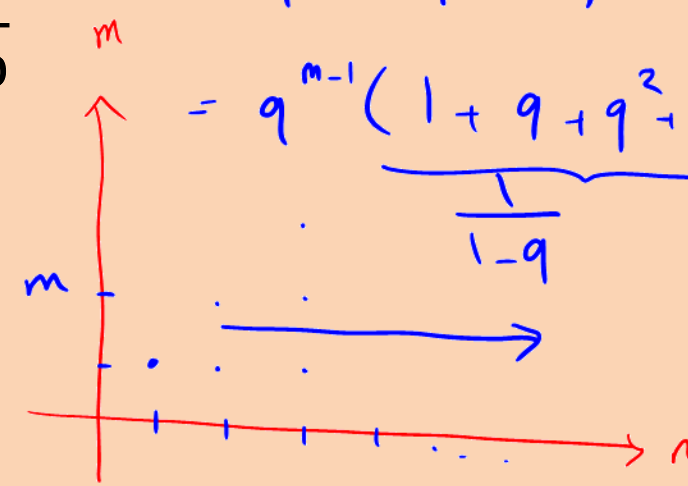
$$E(Y) = \sum_{n=1}^{\infty} \overbrace{n(1-p)^{n-1}p}^{P(Y=n)} = p\sum_{n=1}^{\infty} nq^{n-1}$$

$$= p\sum_{n=1}^{\infty}\sum_{m=1}^{n} q^{n-1} = p\sum_{m=1}^{\infty}\sum_{n=m}^{\infty} q^{n-1}$$

$$= \cancel{p}\sum_{m=1}^{\infty} \frac{q^{m-1}}{\cancel{1-q}} = \frac{1}{1-q} = \frac{1}{p}$$

$$q^{m-1} + q^{m} + q^{m+1} \cdots$$

$$= q^{m-1}\Big(\underbrace{1 + q + q^2 + \cdots}_{\frac{1}{1-q}}\Big)$$



$$P(Y=k) = q^{k-1}p$$

محاسبه E(Y) روش دوم: استفاده از سری تیلور

$$\frac{dq^n}{dq}$$

$$E(Y) = p\sum_{n=1}^{\infty} nq^{n-1} = p\frac{d}{dq}\sum_{n=1}^{\infty} q^n$$

$$= p\frac{d}{dq}\left(\frac{q}{1-q}\right) = p\left(\frac{1}{1-q}\right)^2 = \frac{1}{p}$$

محاسبه E(Y) روش سوم: استفاده از تابع مولد گشتاور

$p(y)$

$$M_Y(t) = E(e^{tY}) = \sum_{y=1}^{\infty} \overbrace{q^{y-1}p}^{p(y)} e^{ty}$$

$q^y (e^t)^y$

$(qe^t)^y$

$$= \frac{p}{q}\sum_{y=1}^{\infty}(qe^t)^y = \frac{p}{\cancel{q}}\frac{\cancel{q}e^t}{1-qe^t} = \frac{pe^t}{1-qe^t}$$

$$E(Y) = \frac{d}{dt}M_Y(t)|_{t=0} = \frac{pe^t}{(1-qe^t)^2}|_{t=0} = \frac{1}{p}$$

محاسبه E(Y) روش چهارم: مشروط کردن به اولین پرتاب

- رخداد Z را موفقیت در اولین پرتاب در نظر بگیرید. داریم:

$$E(Y)=\sum_{y=1}^{\infty}P(y)y=\sum_{y=1}^{\infty}[\overset{p(z)}{p}P(y/Z)+\overset{p(z^c)}{q}P(y\,/\,Z^c)]y$$

$$=p\sum_{y=1}^{\infty}P(y/Z)y+q\sum_{y=1}^{\infty}P(y\,/\,Z^c)y \qquad q\,E[Y\,|\,Z^c]$$

$$=p+q\big(1+E(Y)\big)=1+qE(Y)$$

$$\Rightarrow E(Y)=\frac{1}{p}$$