



جلسه چهارم

خلاصه سازی برای مهداده

مروزی بر الگوریتم موریس++

● صورت مساله: شمارنده

● الگوریتم موریس:

● با احتمال 2^{-X} عدد X را اضافه کن،

● تخمین ما: $2^X - 1$

● تحلیل

$$E[2^X] = n + 1$$

● دوست داشتیم: $Pr[|\tilde{n} - n| < \epsilon n] \leq \delta$

$$Var[2^X]$$

● رسیدیم: $Pr[|\tilde{n} - n| < \epsilon n] \leq \frac{1}{2\epsilon^2}$

مروزی بر الگوریتم موریس++

میانگین خوب

$$Pr[|\tilde{n} - n| < \epsilon n] \leq \frac{1}{2\epsilon^2}$$

واریانس خطی

الگوریتم موریس++:

s تا موریس، پاسخ: میانگین

$$Pr[|\tilde{n} - n| < \epsilon n] \leq \frac{1}{2s\epsilon^2} < \delta$$

واریانس نمایی

الگوریتم موریس++:

t تا موریس+ (با $\delta = 1/3$)، پاسخ: میانه

$$Pr[|\tilde{n} - n| < \epsilon n] \leq 2e^{\Theta(t)}$$

$$O(\epsilon^{-2} \lg(1/\delta) (\lg \lg(n/(\epsilon\delta))))$$

حافظه:



شمارش اعداد متفاوت

(F0)

مسئله تعداد متفاوت‌ها (F0)

• ورودی:

$$i_1, \dots, i_m \in [n] \quad \bullet$$

• خروجی:

• تعداد متفاوت‌ها

- **Solution 1:** keep a bit array of length n , initialized to all zeroes.
Set the i th bit to 1 whenever i is seen in the stream (n bits of memory).
- **Solution 2:** Store the whole stream in memory explicitly
 - جلسه ۴ کوئیزک
???? : ۱ bits of memory
- **Solution 1+2:** $\min\{n, m \lceil \log_2 n \rceil\}$

مساله F0: هدف

جواب واقعی

$$P[|t - \tilde{t}| > \varepsilon t] < \delta$$

تخمین ما

الگوریتم :FM

Algorithm FM:

1. Pick random hash function $h : [n] \rightarrow [0, 1]$.
2. Maintain in memory the smallest hash we've seen so far: $X = \min_{i \in \text{stream}} h(i)$.
3. `query()`: output $1/X - 1$.

ایده: ◎

◎ غیرقابل پیادهسازی

الگوریتم :FM

Algorithm FM:

1. Pick random hash function $h : [n] \rightarrow [0, 1]$.
2. Maintain in memory the smallest hash we've seen so far: $X = \min_{i \in \text{stream}} h(i)$.
3. `query()`: output $1/X - 1$.

ایده: ◎

◎ غیرقابل پیادهسازی

تحليل الگوریتم :FM

Claim 7. $\mathbb{E}[X] = \frac{1}{t+1}$.

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > \lambda) d\lambda$$

اثبات -

تحليل الگوریتم :FM

Claim 7. $\mathbb{E}[X] = \frac{1}{t+1}$.

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > \lambda) d\lambda$$

اثبات -

Claim 8. $\mathbb{E}[X^2] = \frac{2}{(t+1)(t+2)}.$

$$\mathbb{E}[X^2] = \int_0^\infty \mathbb{P}(X^2 > \lambda) d\lambda$$

اثبات -

Claim 8. $\mathbb{E}[X^2] = \frac{2}{(t+1)(t+2)}.$

$$\mathbb{E}[X^2] = \int_0^\infty \mathbb{P}(X^2 > \lambda) d\lambda$$

اثبات -

Claim 8. $\mathbb{E}[X^2] = \frac{2}{(t+1)(t+2)}.$

$$\text{Var}[X] =$$

$$\frac{2}{(t+1)(t+2)} - \frac{1}{(t+1)^2} = \frac{t}{(t+1)^2(t+2)} < (\mathbb{E}[X])^2.$$

الگوریتم : +FM

1. Instantiate $s = \lceil 1/(\epsilon^2\eta) \rceil$ FMs independently, $\text{FM}_1, \dots, \text{FM}_s$.
2. Let X_i be the output of FM_i .
3. Upon a query, output $1/Z - 1$, where $Z = \frac{1}{s} \sum_i X_i$.

کوئیزک:

؟؟؟؟

$$\text{Var}[X_i] = \frac{t}{(t+1)^2(t+2)}$$

چرا جواب‌های الگوریتم‌های FM^+
مستقل‌اند؟

Claim 2.2.3. $\mathbb{P}(|Z - \frac{1}{t+1}| > \frac{\epsilon}{t+1}) < \eta$

اثبات

-

Claim 2.2.4. $\mathbb{P}\left(\left|\left(\frac{1}{Z} - 1\right) - t\right| > O(\varepsilon)t\right) < \eta$

اثبات

-

Claim 2.2.3. $\mathbb{P}(|Z - \frac{1}{t+1}| > \frac{\epsilon}{t+1}) < \eta$

اثبات

-

Claim 2.2.3. $\mathbb{P}(|Z - \frac{1}{t+1}| > \frac{\epsilon}{t+1}) < \eta$

اثبات

-

Claim 2.2.4. $\mathbb{P}\left(\left|\left(\frac{1}{Z} - 1\right) - t\right| > O(\varepsilon)t\right) < \eta$

اثبات

-

Claim 2.2.4. $\mathbb{P}\left(\left|\left(\frac{1}{Z} - 1\right) - t\right| > O(\varepsilon)t\right) < \eta$

- Instantiate $q = \lceil 18 \ln(1/\delta) \rceil$ independent copies of FM+ with $\eta = 1/3$.
- Output the median \hat{t} of $\{1/Z_j - 1\}_{j=1}^q$ where Z_j is the output of the j th copy of FM+.

Claim 2.2.5. $\mathbb{P}(|\hat{t} - t| > \epsilon t) < \delta$

تحليل الگوریتم

$$\begin{aligned}
 Y_i &: \text{копия FM+} & Y &= \sum_{i=1}^q Y_i \\
 EY_i &\geq \frac{1}{4} & EY &= \frac{q}{4} \\
 \mathbb{P}[Y \leq \frac{q}{4}] && \text{برهان: نظر آنچه که (برهان تجزیه و ترکیب)} \\
 &= \mathbb{P}[Y - EY \leq \frac{q}{4} - EY] & (\text{جدا از } EY) \\
 &= \mathbb{P}[EY - Y \geq EY - \frac{q}{4}] & (\text{دسته بندی } (-1)) \\
 &\leq \mathbb{P}[|EY - Y| \geq EY - \frac{q}{4}] & (\text{فرمولیکی لاریزای (فرمولیکی لاریزای)}) \\
 &\leq \mathbb{P}[|EY - Y| \geq \frac{1}{4} EY] & \\
 &\leq e^{-(\frac{1}{4})^2 \frac{q}{4}} & (EY \geq \frac{q}{4}) \\
 &\leq e^{-(\frac{1}{4})^2 \frac{q}{4}} &
 \end{aligned}$$

الگوريتم FM_{++} : حافظه:

- به جز تابع تصادفی h ,
- تعداد اعداد حقیقی:
- چند تا $:FM_+$
- چند تا $:FM$
- هر $:FM$

الگوريتم FM++: حافظه:

- به جز تابع تصادفی h ,
- تعداد اعداد حقیقی:
- چند تا $\ln(1/\delta)$:FM₊
- چند تا $1/(\epsilon^2)$:FM
- هر میک عدد