Lecture 10 Hashing

Thursday, September 28, 2017 2:11 PM

- Monte Carlo alg for area of a circle

Let $X_i = \begin{cases} 0 & i\text{-th point is not in circle} \\ 1 & i\text{-th point is in circle} \end{cases}$

$$\mathbb{E}[X_i] = \Pr[X_i = 1] = \frac{\text{area of circle}}{\text{area of square}} = p \qquad \left(\frac{\pi}{4}\right)$$

Final output: Let $X = \sum_{i=1}^{n} X_i$ (X = count in alg)

output $\frac{4 \cdot X}{n}$

$$\mathbb{E}[\text{output}] = \frac{4\,\mathbb{E}[X]}{n} = \frac{4 \sum_{i=1}^{n} \mathbb{E}[X_i]}{n} = 4 \cdot p = \text{area of circle}$$

Variance

$$\text{Var}[X_i] = \mathbb{E}\left[(X_i - \underbrace{\mathbb{E}[X_i]}_{p})^2\right]$$

$$= \mathbb{E}\left[(X_i - p)^2\right] \qquad X_i = \begin{cases} 0 & \text{w.p. } 1-p \\ 1 & \text{w.p. } p \end{cases}$$

$$= (1-p)(0-p)^2 + p(1-p)^2$$

$$= p(1-p)\left[p + (1-p)\right]$$

$$= p(1-p)$$

Recall: $X = \sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n$

$$\text{Var}[X] = \sum_{i=1}^{n} \text{Var}[X_i] = n\,p(1-p)$$

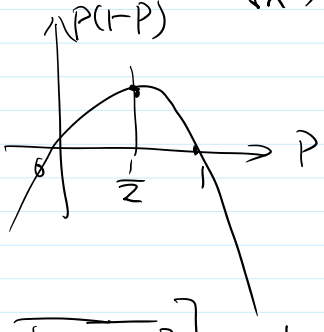
output $= \frac{4X}{n}$ $\qquad \text{Var}[c \cdot X] = \boxed{c^2}\,\text{Var}[X]$

$$\text{Var}[\text{output}] = \left(\frac{4}{n}\right)^2 \cdot \text{Var}[X] = \frac{16\,p(1-p)}{n}$$

Claim: $\Pr\left[\left|\text{output} - \text{area of circle}\right| > \frac{4}{\sqrt{n}}\right] \le \frac{1}{4}$

(in other words, with probability $\frac{3}{4}$, output is within $\frac{4}{\sqrt{n}}$)

$$\text{Var}[\text{output}] = \frac{16\,p(1-p)}{n} \le \frac{4}{n}$$



$$\sqrt{\text{Var}[\text{output}]} \le \frac{2}{\sqrt{n}}$$

$$2\sqrt{\text{Var}[\text{output}]} \le \frac{4}{\sqrt{n}}$$

$$2\sqrt{\mathrm{Var}[\text{output}]} \le \frac{4}{\sqrt{n}}$$

By chebysher $\Pr\left[\left|\text{output} - \mathbb{E}[\text{output}]\right| > 2\sqrt{\mathrm{Var}[\text{output}]}\right] \le \frac{1}{4}$

(choosing $\lambda = 2$ in the inequality)

$$\Pr\left[\left|\text{output} - \text{area}\right| > \frac{4}{\sqrt{n}}\right] \le \Pr\left[\left|\text{output} - \mathbb{E}[\text{output}]\right| > 2\sqrt{\mathrm{Var}[\text{output}]}\right]$$

$$\le \frac{1}{4}$$

□

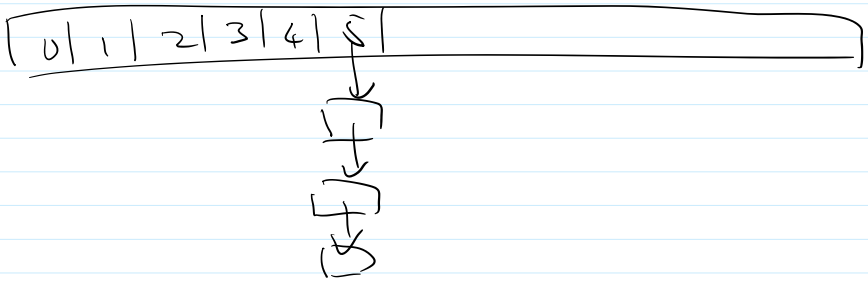
- Hashing:
  - fixed hash function

    $a[m]$

    $f(i) = i \bmod m$

    if the set is $S = \{s, m+s, 2m+s, \cdots, 2nm+s\}$

    for all $i \in S$ $f(i) = s$



  - random function

    how to choose a random hash function

    for every integer i, choose $f(i)$ independently at random inside $\{0, 1, 2, \cdots, m-1\}$

    (function is chosen at the beginning, and fixed for later use)

  - hash function by modular arithmetic

    $f_{a,b}(x) = ax + b \pmod p$

    want for $x \neq y \bmod p$ u, v

    $$\Pr_{a,b}\left[f_{a,b}(x) = u, f_{a,b}(y) = v\right] = \frac{1}{p^2}$$

Proof: in order to have $f_{a,b}(x)=u$ $f_{a,b}(y)=v$

$$\begin{cases} a\boxed{x}+b \equiv \boxed{u} \pmod p \\ a\boxed{y}+b \equiv \boxed{v} \pmod p \end{cases}$$

system of linear equation has a unique solution

$$a=\frac{(u-v)}{(x-y)} \qquad b=u-ax=v-ay$$

$$Pr_{a,b}[f(x)=u, f(y)=v]=\frac{1}{p^2}$$

← unique solution of equation

← p^2 choices for a, b

Pseudo-code for hashing

```
initialize()
    choose a prime number p
    choose a,b ∈ {0,1,...,p-1} randomly.
    create an array a[0...p-1]
    initialize each cell a[i] to be
        the head of an empty linked list.
f(x)
    return (ax+b) mod p;
insert(x)
    insert x to the linked list a[f(x)]
find(x)
    check if x is in the linked list a[f(x)]
```

← the size of array can also be $m < p$

← if the size of array is m, return ((ax+b) mod p) mod m