Streaming Algorithms

Meng-Tsung Tsai

02/27/2018

A Reference Book

- "The Probabilistic Method", Alon and Spencer (2004)

You may find an e-copy of this book on www.lib.nctu.edu.tw

Reschedule

I have to travel abroad for some business in the week of midterm exams, Apr 15 - 20. Here are 2 possible plans. [Voted]

(A) Let the second written assignment be a take-home midterm exam. The number of problem sets will be a little more than an ordinary written assignement, and

the weight of midterm exam 10% → 20%
the weight of final project 40% → 30%

~~(B) Find some free time in common to have a 3-hour make-up class.~~

More on Expectation

Inequalities on Expectation

Let X be a random variable. Then we have:

(1) $\Pr[X \geq E[X]] > 0$

(2) $\Pr[X \leq E[X]] > 0$

What's more?
We may impose additional restrictions on X.

Inequalities on Expectation

If $X \geq 0$ and $E[X] > 0$, then we have:

$\Pr[X \geq \lambda E[X]] \leq 1/\lambda$ for any $\lambda \geq 1$, (Markov inequality)

Inequalities on Expectation

If $X \geq 0$ and $E[X] > 0$, then we have:

$\Pr[X \geq \lambda E[X]] \leq 1/\lambda$ for any $\lambda \geq 1$, (Markov inequality)

together with $X \leq cE[X]$ for some $c \geq 1$, we get:

$$E[X] \leq cE[X]\Pr[X > (1-\varepsilon)E[X]] + (1-\varepsilon)E[X]\Pr[X \leq (1-\varepsilon)E[X]]$$

$$\Rightarrow \Pr[X > (1-\varepsilon)E[X]] \geq \varepsilon/(c-1+\varepsilon) \quad \text{for any } \varepsilon \text{ in } (0, 1)$$

Sum-free Subsets

Theorem 1. [Erdős 1965] Every set $B = \{b_1, b_2, ..., b_n\}$ of n non-zero integers has a sum-free subset A of size $|A| > n/3$.

Definition. A set S is sum-free if $\forall$ a, b, c $\in$ S (possibly repeat), $a+b \neq c$.

Sum-free Subsets

Proof Strategy.

If S is sum-free in $\mathbf{Z}_p$, then S is sum-free in $\mathbf{Z}$. It suffices to show that B has a sum-free subset A of size $|A| > n/3$ in $\mathbf{Z}_p$.

Let p be a prime in the form of 3k+2 so that $p > 2 \max |b_i|$.
(By Dirichlet Theorem, such p exists.)

Then $C = \{k+1, ..., 2k+1\}$ is sum-free in $\mathbf{Z}_p$
$\Rightarrow xC = \{xc \pmod p): c \in C\}$ is sum-free in $\mathbf{Z}_p$ for any non-zero x's.

Find such an x (exists?) that $|B \cap xC| > n/3$. Then output $B \cap xC$ as the sum-free subset.

Sum-free Subsets

x	xc_1	xc_2	...
1	$\pi_1(1)$	$\pi_2(1)$	
2	$\pi_1(2)$	$\pi_2(2)$	
...	...	...	
p-1	$\pi_1(p-1)$	$\pi_2(p-1)$	

Every b in $\{1, ..., p-1\}$ appears exactly once in each column, and appears at most once in each row.

Sum-free Subsets

x	xc_1	xc_2	...
1	$\pi_1(1)$	$\pi_2(1)$	
2	$\pi_1(2)$	$\pi_2(2)$	
...	...	...	
p-1	$\pi_1(p-1)$	$\pi_2(p-1)$	

For an x picked uniformly at random from $\{1, ..., p-1\}$, $\Pr[b \in xC] = |C|/(p-1)$.

Sum-free Subsets

x	xc_1	xc_2	...
1	$\pi_1(1)$	$\pi_2(1)$	
2	$\pi_1(2)$	$\pi_2(2)$	
...	...	...	
p-1	$\pi_1(p-1)$	$\pi_2(p-1)$	

Let $X = \sum_b X_b$. where X_b be the indicator random varaible denoting whether $b \in xC$ for a random x.

Sum-free Subsets

x	xc_1	xc_2	...
1	$\pi_1(1)$	$\pi_2(1)$	
2	$\pi_1(2)$	$\pi_2(2)$	
...	...	...	
p-1	$\pi_1(p-1)$	$\pi_2(p-1)$	

$E[X] = \sum E[X_b] = n|C|/(p-1) = n(k+1)/(3k+1) > n/3.$

Construct a Large Sum-free Subset

Observe that $X \geq 0$ and $E[X] > 0$.

Because $X < n$ and $E[X] > n/3$, $X \leq 3E[X]$.

By the second form of Markov inequality, we have:

$$\Pr[X > (1-\varepsilon)E[X]] \geq \varepsilon/(2+\varepsilon) \quad \text{for any } \varepsilon \text{ in } (0, 1).$$

w.h.p. means with probability $1-1/n^{\Omega(1)}$.

By trying O(log n) random x's, we can have a sum-free subset of size $(1-\varepsilon)n/3$ for any constant ε in (0, 1) w.h.p.

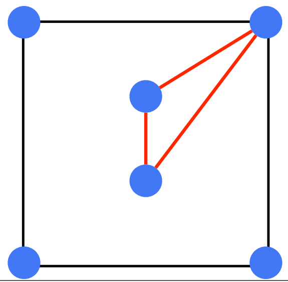
Alteration

Maximize the Minimum-area Triangle

Let S be a set of n points in a unit square, and let T(S) be the minimum area of traingles whose vertices are three distinct points in S. Let T(n) be the maximum possible T(S) for all S.

<u>Theorem 2</u>. (Thm 3.3.1) There is a set S so that $T(S) \geq 1/(100n^2)$ i.e. $T(n) = \Omega(1/n^2)$.

For n = 6,



Maximize the Minimum-area Triangle

Proof Strategy. Randomly sample n+C points, if there are at most C triangles of area < $1/(100n^2)$, then remove a vertex for each triangle. In this way, all triangles induced by some n points have area ≥ $1/(100n^2)$. This technique is called alteration.

Maximize the Minimum-area Triangle

Let P, Q, R be points sampled independently and uniformly from the unit square. Let **μ** be the area of triangle P, Q, R.

Claim. $\Pr[\boldsymbol{\mu} \le \varepsilon] \le 16\pi\varepsilon$.

Sample 2n points indepently and uniformly from the unit square, then let X be the number of triangles of area < $1/(100n^2)$.

$$E[X] \le \binom{2n}{3} 16\pi/(100n^2) < n$$

There exists an arrangement of 2n points that induce at most n triangles of area < $1/(100n^2)$.

Maximize the Minimum-area Triangle

Let P, Q, R be points sampled independently and uniformly from the unit square. Let **μ** be the area of triangle P, Q, R.

Claim. $\Pr[\boldsymbol{\mu} \le \varepsilon] \le 16\pi\varepsilon$.

Sample 2n points indepently and uniformly from the unit square, then let X be the number of triangles of area < $1/(100n^2)$.

$$E[X] \le \binom{2n}{3} 16\pi/(100n^2) < n$$

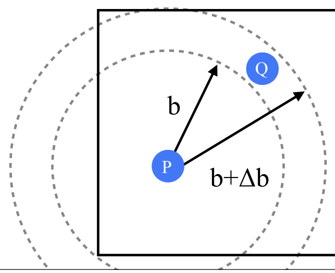
There exists an arrangement of n points that induce no triangle of area < $1/(100n^2)$.

Proof of $\Pr[\boldsymbol{\mu} \le \varepsilon] \le 16\pi\varepsilon$

Let ℓ be the distance between P and Q. Then we have

$$\Pr[b \le \ell \le b+\Delta b] = \pi(b+\Delta b)^2 - \pi b^2, \text{ and in the limit}$$

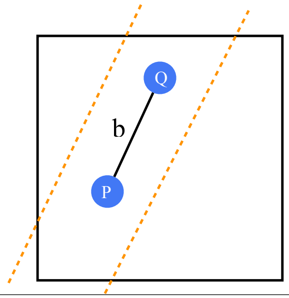
$$\Pr[b \le \ell \le b+db] = 2\pi b\, db.$$



Proof of Pr[$\boldsymbol{\mu} \leq \varepsilon$] $\leq 16\pi\varepsilon$

Given PQ has length b, to make PQR has area $\leq \varepsilon$, then R must fall within the following strip of width = $2*2\varepsilon/b$ and length $\leq \sqrt{2}$.

$$\text{Thus, } \Pr[\boldsymbol{\mu} \leq \varepsilon] \leq \int_0^{\sqrt{2}} 2\pi b \frac{4\sqrt{2}\varepsilon}{b} db = 16\pi\varepsilon$$



Exercise 1

What is the running time of finding the smallest-area triangle of n given points?

Suppose that given n points, testing whether there are 3 points colinear requires $\Omega(n^2)$ time. What can we infer?

The Second Moment

Chebyshev Inequality

Recall the definition of the varience,

$$\mathrm{Var}(X) = E[(X-\boldsymbol{\mu})^2] = \sigma^2.$$

By Markov inequality, we get

$$\Pr[(X-\boldsymbol{\mu})^2 \geq \lambda^2\sigma^2] \leq 1/\lambda^2,$$

or equivalently

$$\Pr[|X-\boldsymbol{\mu}| \geq \lambda\sigma] \leq 1/\lambda^2. \text{ (Chebyshev Inequality)}$$

What's more?
We may impose additional restrictions on X.

Chebyshev Inequality

If X is a nonnegative integral-valued random variable and E[X]>0,

$$\Pr[X = 0] \le \Pr[|X-\boldsymbol{\mu}| \ge \boldsymbol{\mu}] = \Pr[|X-\boldsymbol{\mu}| \ge (\boldsymbol{\mu}/\sigma)\sigma] \le \mathrm{Var}(X)/E[X]^2.$$

Can we do something similar for X=k? What is the assumption?

We may like to bound Pr[X = 0] small because we may set X=0 to be an undesirable case.

Exercise 2

Prove that for nonnegative integral-valued random variable X,

$$\Pr[X=0] \le \mathrm{Var}(X)/E[X^2].$$

Distinct Sums

We say a set $x_1, x_2, \ldots, x_k$ of positive integers has distinct sums if all subset sums

$$\sum_{i\in S} x_i \quad \text{for all } S \subseteq \{1, 2, \ldots, k\}$$

are distinct. Let f(n) denote the maximum k for the case $x_1, x_2, \ldots, x_k \le n$.

$1, 2, \ldots, 2^{\lfloor \log_2 n \rfloor}$ yields that $f(n) \ge \lfloor \log_2 n \rfloor + 1$.

Distinct Sums

We say a set $x_1, x_2, \ldots, x_k$ of positive integers has distinct sums if all subset sums

$$\sum_{i\in S} x_i \quad \text{for all } S \subseteq \{1, 2, \ldots, k\}$$

are distinct. Let f(n) denote the maximum k for the case $x_1, x_2, \ldots, x_k \le n$.

Since each subset sum is unique and less than nk,
$2^{f(n)} < nf(n) \Rightarrow f(n) < \log_2 n + \log_2 \log_2 n + O(1)$

A Tighter Upper Bound

Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$ be independent random variables so that

$$\Pr[\varepsilon_i = 1] = 1/2 \text{ and } \Pr[\varepsilon_i = 0] = 1/2 \text{ for each i in } [1, k].$$

Let $X = \sum_i \varepsilon_i x_i$ i.e. a random subset sum. Thus, $E[X] = (\sum_i x_i)/2$ and $\mathrm{Var}(X) = (\sum_i (x_i)^2)/4$ (Why?).

Let $X = \sum X_i$. If all X_i's are independent, then $\mathrm{Var}(X) = \sum \mathrm{Var}(X_i)$.

A Tighter Upper Bound

Let $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_k$ be independent random variables so that

$$\Pr[\varepsilon_i = 1] = 1/2 \text{ and } \Pr[\varepsilon_i = 0] = 1/2 \text{ for each i in } [1, k].$$

Let $X = \sum_i \varepsilon_i x_i$ i.e. a random subset sum. Thus, $E[X] = (\sum_i x_i)/2$ and $\mathrm{Var}(X) = (\sum_i (x_i)^2)/4 \le kn^2/4$.

By Chebyshev inequality,

$$\Pr[|X-\boldsymbol{\mu}| \ge \lambda n k^{1/2}/2] \le 1/\lambda^2.$$

or equivalently

$$\Pr[|X-\boldsymbol{\mu}| < \lambda n k^{1/2}/2] \ge 1-1/\lambda^2. \quad \text{(a)}$$

A Tighter Upper Bound

By Chebyshev inequality,

$$\Pr[|X-\boldsymbol{\mu}| \ge \lambda n k^{1/2}/2] \le 1/\lambda^2.$$

or equivalently

$$\Pr[|X-\boldsymbol{\mu}| < \lambda n k^{1/2}/2] \ge 1-1/\lambda^2. \quad \text{(a)}$$

Because each distinct sum occurs with probability either 0 or 2^{-k},

$$\Pr[|X-\boldsymbol{\mu}| < \lambda n k^{1/2}/2] < (\lambda n k^{1/2}+1)2^{-k}. \quad \text{(b)}$$

Combine (a), (b) and set $\lambda = \sqrt{3}$ yields that

$$f(n) \le \log_2 n + \frac{1}{2}\log_2 \log_2 n + O(1)$$