Probabilistic Method and Random Graphs

Lecture 12. Sample&Modify and Second Moment Method

Xingwu Liu

Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

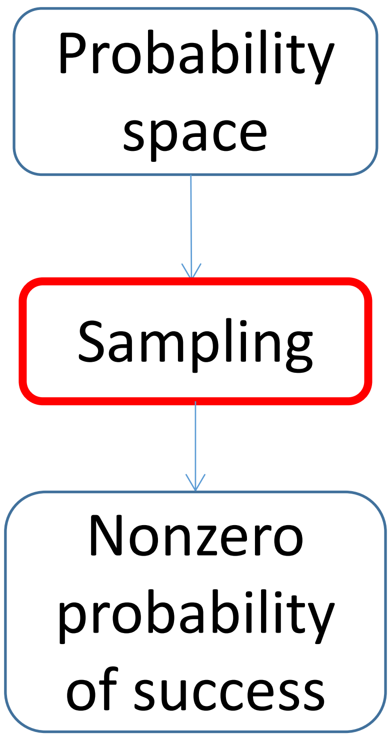
[1]The slides are mainly based on Chapter 6 of Probability and Computing.

Comments, questions, or suggestions?

Recap of Lecture 11

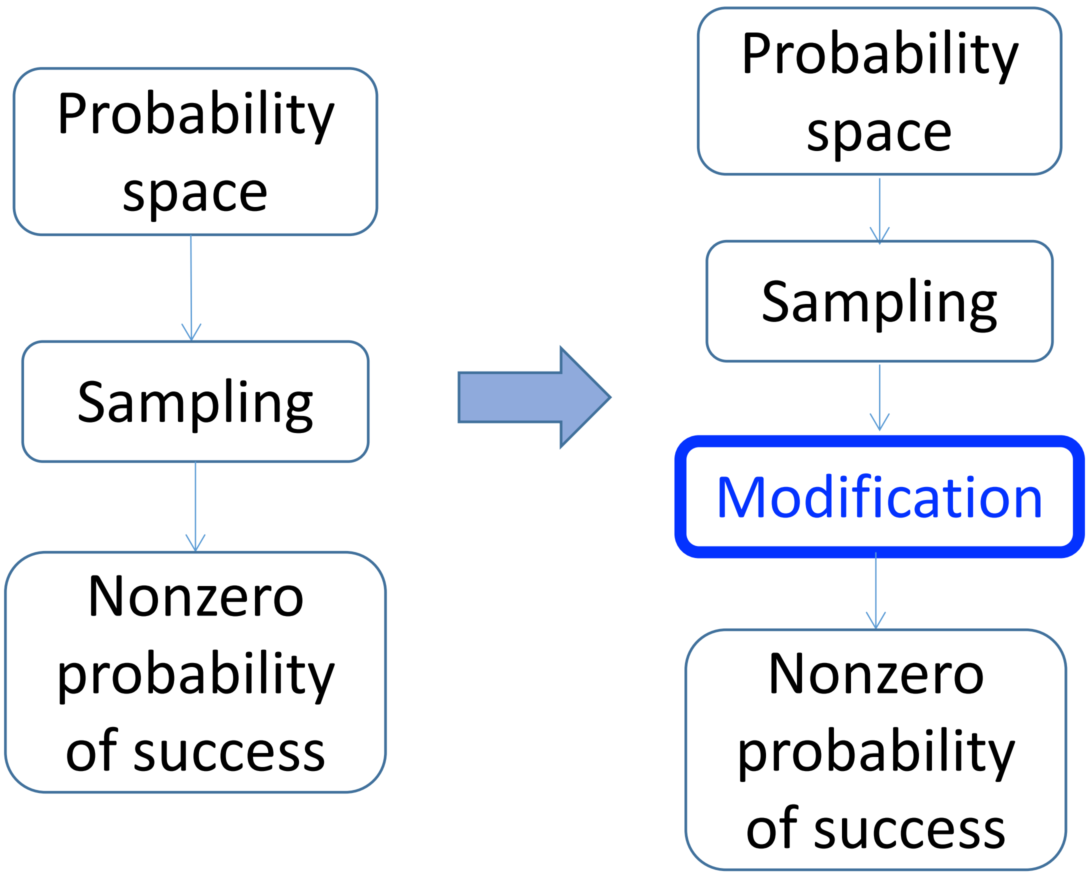
- De-randomization
  - Expectation argument leads to efficient randomized algo.
    - Sample and verify (succeed if lucky)
  - De-randomizing such algo. leads to deterministic algo
    - Sequentially make deterministic choices, maintaining conditional expectation
- Precondition
  - Only valid for expectation argument where randomness lies in a sequence of random variables
- Built on randomized algo.

Sample



The process is doomed!
Can we do anything?
De-randomization works, but conditionally.
Sample&modify speeds it up.

Sample and Modify



Big Chromatic Number and Big Girth

- Chromatic number vs local structure
  - Sparse local structure → small chro. number?
  - No! (Erdős 1959)
- One of the first applications of prob. Method
- Theorem: for any integers $g, k > 0$, there is a graph with girth$\geq g$ and chro. number$\geq k$
- We just prove the special case $g = 4$, i.e. triangle-free

Technical challenge

- It is hard to compute/estimate/check chro. number
  - $\chi(G)$: the chromatic number of G
- Often handled indirectly via easy-to-handle features
- Example:
  - $\mathbb{I}(G)$: the size of a maximum independent set of G
  - $\mathbb{I}(G)\chi(G) \geq n$
  - Small $\mathbb{I}(G)$ implies big $\chi(G)$

Basic Idea of the Proof

- Randomly pick a graph G from $G_{n,p}$
- With high probability $\mathbb{I}(G)$ is small
  - $\chi(G)$ is big w.h.p.
- With high probability G has few triangles
- Destroy the triangles while keeping $\mathbb{I}(G)$ small

Proof: $\mathbb{I}(G)$ is small w.h.p.

- S: a vertex set of size $\frac{n}{2k}$
- A_S: S is an independent set
- $\Pr\left(\mathbb{I}(G) \geq \frac{n}{2k}\right) = \Pr(\cup_S A_S)$

$$\leq \binom{n}{n/2k}(1-p)^{\binom{n/2k}{2}}$$

$$< 2^n e^{-\frac{pn(n-2k)}{8k^2}}$$

which is small if n is large and $p = \omega(n^{-1})$

Proof: triangles are few w.h.p.

- $\mathcal{T}(G)$: the number of triangles of G
- $\mathbb{E}[\mathcal{T}(G)] = \binom{n}{3}p^3 < \frac{(np)^3}{6} = \frac{n}{6}$ if $p = n^{-\frac{2}{3}}$
- By Markov ineq., $\Pr\left(\mathcal{T}(G) > \frac{n}{2}\right) \le \frac{1}{3}$
- Recall $\Pr\left(\mathbb{I}(G) \ge \frac{n}{2k}\right) < 2^n e^{-\frac{pn(n-2k)}{8k^2}}$

$$< e^n\, e^{-\frac{pn^2}{16k^2}} = e^{n-n^{\frac{4}{3}}/16k^2} \quad \text{if } n > 4k$$

$$< e^{-n} < \frac{1}{6} \quad \text{if } n^{1/3} \ge 32k^2$$

Proof: destroy triangles

- $\Pr\left(\mathbb{I}(G) < \frac{n}{2k}, \mathcal{T}(G) \leq \frac{n}{2}\right) > \frac{1}{2}$
  - Choose G s.t. $\mathbb{I}(G) < \frac{n}{2k}, \mathcal{T}(G) \leq \frac{n}{2}$
- Remove one vertex from each triangle of G, resulting in a graph G' with $n' \geq n - \mathcal{T}(G)$
- $\mathbb{I}(G') \leq \mathbb{I}(G) < \frac{n}{2k}$
- $\chi(G') \geq \frac{n'}{\mathbb{I}(G')} \geq \frac{n - \mathcal{T}(G)}{\frac{n}{2k}} \geq k$

Algorithm for finding such a graph

- Fix $n^{1/3} \geq 32k^2$ and $p = n^{-2/3}$
- Sample G from $G_{n,p}$
- Destroy the triangles

- Success probability > ½

- Do you have any idea of de-randomizing?

Main Probabilistic Methods

- Counting argument
- First-moment method
- Second-moment method
- Lovasz local lemma

Second moment argument

- Chebyshev Ineq.: $\Pr(|X - \mathbb{E}[X]| \geq a) \leq \frac{\mathrm{Var}[X]}{a^2}$
- A special case:

$$\Pr(X = 0) \leq \Pr(|X - \mathbb{E}[X]| \geq \mathbb{E}[X]) \leq \frac{\mathrm{Var}[X]}{(\mathbb{E}[X])^2}$$

  - Compare with $\Pr(X \neq 0) \leq \mathbb{E}[X]$ for integer r.v. X
- Typically works when nearly independent
  - Due to the difficulty in computing the variance

An improved version by Shepp

- $\Pr(X = 0) \le \frac{\mathrm{Var}[x]}{\mathbb{E}[X^2]} \le \frac{\mathrm{Var}[X]}{(\mathbb{E}[X])^2}$
- Proof: $(\mathbb{E}[X])^2 = (\mathbb{E}[1_{X \neq 0} \cdot X])^2$
$$\le \mathbb{E}[1_{X\neq 0}^2]\mathbb{E}[X^2]$$
$$= \Pr(X \neq 0)\mathbb{E}[X^2]$$
$$= \mathbb{E}[X^2] - \Pr(X = 0)\mathbb{E}[X^2]$$
  - The inequality is due to $\left(\int fg\right)^2 \le \int f^2 \int g^2$
- When $X \ge 0$, $\Pr(X > 0) > \frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]}$

Generalizing Shepp's Theorem

- $\Pr(X > \theta\mathbb{E}[X]) \geq (1-\theta)^2 \frac{(\mathbb{E}[X])^2}{\mathbb{E}[X^2]}$, $\theta \in (0,1)$
- Paley&Zygmund, 1932
- Proof:

$$\mathbb{E}[X] = \mathbb{E}\left[X1_{X\leq\theta\mathbb{E}[X]}\right] + \mathbb{E}\left[X1_{X>\theta\mathbb{E}[X]}\right]$$
$$\leq \theta\mathbb{E}[X] + \left(\mathbb{E}[X^2]\Pr(X > \theta\mathbb{E}[X])\right)^{\frac{1}{2}}$$

- Further improvement, tight when X is constant

$$\Pr(X > \theta\mathbb{E}[X]) \geq \frac{(1-\theta)^2(\mathbb{E}[X])^2}{\mathrm{Var}[X]+(1-\theta)^2(\mathbb{E}[X])^2}$$

due to $\mathbb{E}\left[X - \theta\mathbb{E}[X]\right] \leq \mathbb{E}\left[(X - \theta\mathbb{E}[X])1_{X>\theta\mathbb{E}[X]}\right]$

App.: Erdős distinct sum problem

- $S \subset \mathbb{R}^+$ has distinct subset sums
  - Different subsets have different sums
  - Example: $S = \{2^0, 2^1, \dots 2^k\}$
- Fix $n \in \mathbb{Z}^+$. Let $f(n)$ be the max size of $S \subset [n]$ which has distinct subset sums.
- Easy lower bound: $f(n) \geq \lfloor \ln_2 n \rfloor + 1$
- Erdős promised 500$: $f(n) \leq \lfloor \ln_2 n \rfloor + c$
  - Now offered by Ron Graham?

An easy bound: $k \leq \ln_2 n + \ln_2 \ln_2 n + 1$

- Assume k-set $S \subseteq [n]$ has distinct subset sums
- There are 2^k subset sums
- Each subset sum $\in [nk]$
- So, $2^k \leq nk$
- $k \leq \ln_2 n + \ln_2 k \leq \ln_2 n + \ln_2(\ln_2 n + \ln_2 k)$

$$\leq \ln_2 n + \ln_2(2\ln_2 n)$$

$$= \ln_2 n + \ln_2 \ln_2 n + 1$$

- Can it be tighter? Yes!

A tighter upper bound

- Intuition underlying the proof:
  - A small interval ($[nk]$) has many (2^k) distinct sums
- If the sums are not distributed uniformly
  - Most of the sums lie in a much smaller interval
  - k must be smaller
  - It is the case by Chebyshev's Inequality

Proof: $f(n) = \ln_2 n + \frac{1}{2} \ln_2 \ln_2 n + O(1)$

- Fix a k-set $S \subset [n]$ with distinct subset sums
- X: the sum of a random subset of S
  - $\mu = \mathbb{E}[X], \sigma^2 = Var[X]$
- $\Pr(|X - \mu| \geq \alpha\sigma) \leq \frac{1}{\alpha^2} \Rightarrow$

$$1 - \frac{1}{\alpha^2} \leq \Pr(|X - \mu| < \alpha\sigma) \Rightarrow$$

$$1 - \frac{1}{\alpha^2} \leq \sum_{|i-\mu|<\alpha\sigma} \Pr(X = i)$$

Proof: $f(n) = \ln_2 n + \frac{1}{2}\ln_2\ln_2 n + O(1)$

- Fix a k-set $S \subset [n]$ with distinct subset sums
- X: the sum of a random subset of S
  - $\mu = \mathbb{E}[X], \sigma^2 = \mathrm{Var}[X]$
- $\Pr(|X-\mu| \geq \alpha\sigma) \leq \frac{1}{\alpha^2} \Rightarrow$

$$1 - \frac{1}{\alpha^2} \leq \Pr(|X-\mu| < \alpha\sigma) \Rightarrow$$

$$1 - \frac{1}{\alpha^2} \leq \sum_{|i-\mu|<\alpha\sigma} \Pr(X=i) \leq \frac{2\alpha\sigma}{2^k}$$

Since $\Pr(X=i)$ is either 0 or 2^{-k}

Proof (continued)

- Estimating σ (assume $S = \{a_1, \dots, a_k\}$):

$$\sigma^2 = \frac{a_1^2 + \cdots + a_k^2}{4} \le \frac{n^2 k}{4} \Rightarrow \sigma \le \frac{n\sqrt{k}}{2}$$

$$\Rightarrow 1 - \frac{1}{\alpha^2} \le \frac{2\alpha\sigma}{2^k} \le \frac{\alpha n \sqrt{k}}{2^k}$$

$$\Rightarrow n \ge \frac{2^k\left(1 - \frac{1}{\alpha^2}\right)}{\alpha\sqrt{k}}$$

- This holds for any $\alpha > 1$. Let $\alpha = \sqrt{3}$
- $n \ge \frac{2}{3\sqrt{3}} \frac{2^k}{\sqrt{k}} \Rightarrow k \le \ln_2 n + \frac{1}{2} \ln_2 \ln_2 n + O(1)$

Application: threshold function

- Consider a property P of random graph $G_{n,p}$
- Threshold function $t(n)$ for P is such that

$$\lim_{n\to\infty} \Pr\left(G_{n,p} \text{ has } P\right) = \begin{cases} 0 \ \text{ if } p = o(t(n)) \\ 1 \text{ if } p = \omega(t(n)) \end{cases}$$

- Example (clique number $c(G)$: max clique size)
  - P: $c(G) \geq 4$
  - $t(n) = n^{-\frac{2}{3}}$ is the threshold function for P

Proof: when $p = o(n^{-\frac{2}{3}})$

- S: a 4-subset of the n vertices
- X_S: indicator of whether S spans a clique
- $X = \sum_S X_S$: the number of 4-cliques
- $\mathbb{E}[X] = \binom{n}{4} p^6 = \Theta(n^4 p^6) = o(1)$
- By Markov's inequality

  $\Pr(c(G) \geq 4) = \Pr(X > 0) \leq \mathbb{E}[X] = o(1)$

Proof: when $p = \omega(n^{-\frac{2}{3}})$

- To derive $\Pr(X > 0) \to 1$
  - By Chebychev's Ineq.: $\Pr(X = 0) \le \frac{\mathrm{Var}[X]}{(\mathbb{E}[X])^2}$
  - Try to show $\mathrm{Var}[X] = o(\mathbb{E}[X])^2$
- Recall $\mathrm{Var}[X] = \sum_S \mathrm{Var}[X_S] + \sum_{S \ne T} \mathrm{Cov}(X_S, X_T)$
- X_S is an indicator $\Rightarrow \mathrm{Var}[X_S] \le \mathbb{E}[X_S]$
- $\mathrm{Cov}(X_S, X_T) \le \mathbb{E}[X_S X_T]$

$$= \Pr(X_S = 1, X_T = 1)$$

$$= \mathbb{E}[X_S]\Pr(X_T = 1 | X_S = 1)$$

And $\mathrm{Cov}(X_S, X_T)$=0 if independent

Proof: estimating the variance

- $\mathrm{Var}[X] \le \sum_S \mathbb{E}[X_S] + \sum_S \mathbb{E}[X_S] \sum_{T\sim S} \Pr(X_T = 1 | X_S = 1)$
  $= \sum_S \mathbb{E}[X_S]\Delta_S$
- $\Delta_S = 1 + \sum_{|T\cap S|=2} \Pr(X_T = 1|X_S = 1)$
  $+ \sum_{|T\cap S|=3} \Pr(X_T = 1|X_S = 1)$
  $= 1 + \binom{n-4}{2}\binom{4}{2}p^5 + \binom{n-4}{1}\binom{4}{3}p^3$
  $= o(n^4p^6) = o(\mathbb{E}[X])$
- $\mathrm{Var}[X] = o(\mathbb{E}[X]^2) \Rightarrow \Pr(X = 0) \le \frac{\mathrm{Var}[X]}{\mathbb{E}[X]^2} = o(1)$
  $\Rightarrow \Pr(X > 0) \to 1$

References

- http://www.openproblemgarden.org/

Thank you