Lecture 7: Probabilistic Methods

Rajat Mittal

IIT Kanpur

This lecture will focus on probabilistic methods. They are used to prove the existence of *good* structures using probability. Note that the problem description will not mention a randomized setting or have a probability connection.

The main idea is to *define* a probability distribution over the set of structures. Using the defined probability distribution, it is proven that the probability of good structure being present is non-zero, implying that a good structure exist. This might seem a bit vague and is best illustrated with the help of applications.

1 Ramsey numbers

You might have covered Ramsey numbers in a course on graph theory (CS201). We will only worry about undirected graphs in this section. There are two important and opposite concepts to remember,

- Clique: A set of vertices is called a *clique* if all vertices are connected to each other.
- Independent set : A set of vertices is called an *independent set* if no two vertices are connected to each other.

Let G denote a graph on n vertices. The elementary question in this field would be, for any graph on 6 vertices, there is either an independent set or a clique of size 3.

Exercise 1. Try to prove the above assertion.

We can generalize the above question. Given some k and l (two natural numbers), is there an n for which all graphs have either an independent set of size k or clique of size l? Notice that the previous question sets $k, l = 3$ and proves that $n \geq 6$ satisfies the constraint.

Exercise 2. Show a graph on 5 vertices which does not have a 3 clique (triangle) or an independent set of size 3.

Given k, l, it has been shown that there always exists a big enough n, s.t., there is an independent set of size k or clique of size l. The smallest such number n is called the *Ramsey number* $R(k, l)$.

It has been a big open question to find bounds, both upper and lower, on $R(k, l)$. Now, we will use probabilistic method to give a lower bound on the diagonal Ramsey number $R(k, k)$. This is the least n, for which, every graph on n vertices have either a clique or independent set of size k.
Call a graph on n vertices to be *good*, if there are no cliques/independent set of size k. Proving a lower bound of n on $R(k, k)$ amounts to showing, there always exists at least one good graph on n vertices. Notice the resemblance with the main idea of probabilistic methods discussed before this section. Existence of good graph does not have anything to do with probability and is purely a combinatorial question. The trick would be to define a probability distribution over graphs and show that there is a nonzero (positive) probability of obtaining a good graph, showing the existence.

Exercise 3. Can you think of a probability distribution over graphs?

The probability distribution over graphs is the simplest possible distribution, just randomly assign the edges of on every pair of n vertices. That means, we pick each graph uniformly with probability $2^{-\binom{n}{2}}$, where $\binom{n}{2}$ is the number of pairs on n vertices. If there is a positive probability (over the random graphs) that none of the k-size subgraphs are clique/independent sets, then there exist a graph which is good.

Other way to describe our probability distribution is, we keep every edge independently with probability $1/2$ in the graph. There are in total $\binom{n}{k}$ subgraphs of size k.

Exercise 4. Show that a particular subgraph of size k is a clique/independent set with probability $2 \cdot 2^{-\binom{k}{2}}$.

Hint: Subgraph would have all edges or all non-edges.

Index a subgraph of size k by i, and let C_i be the event that i-th subgraph is a clique/independent set. Union bounds shows that,

$$P(\cup_{i=1}^{\binom{n}{k}} C_i) \leq \sum_{i=1}^{\binom{n}{k}} P(C_i).$$

So, the total probability that any subgraph is a clique/independent set is at most $\binom{n}{k} 2^{1-\binom{k}{2}}$. If this probability is less than 1, then there is a positive probability that none of the subgraphs are clique/independent set. We showed,

$$\Pr_{\text{graph}}(\text{no subgraph is clique/independent set}) \geq 0.$$

Since the probability is over random graphs, there exist a good graph (such that no subgraphs are clique/independent set).

Theorem 1. *If* $2^{1-\binom{k}{2}}\binom{n}{k} < 1$*, then* n *is a lower bound on* $R(k,k)$*.*

To get an explicit lower bound, you can check that $n = \lceil 2^{k/2} \rceil$ will satisfy the above equation.

The essential argument in the above proof is, the fraction of graphs destroyed by a subgraph (by being clique/independent set) is much smaller than the total number of graphs.

A counting argument for the above theorem can also be constructed (assignment problem). Actually, in all our applications, a counting argument can always be given. But the probabilistic argument in general is much simpler and easier to construct.

Probabilistic algorithm One of the important thing to notice in a probabilistic method of proofs is that the proofs are non-constructive. For the previous example, it means that we were only able to show existence of a graph. This proof does not construct the required graph and hence is called non-constructive.

But suppose we choose n to be $\frac{1}{2}\lceil 2^{k/2} \rceil$. Then the probability of having a clique/independent set is very small. This shows that most of the random graphs will be good graphs.

This suggests a randomized algorithm. We just pick a random graph. Because of the argument above, with high probability we will get a good graph.

2 Using linearity of expectation

We have already discussed linearity of expectation. It is a simple result to prove, but has profound implications. Again, the importance of linearity lies in the fact that we can even take dependent random variables and still decompose the expectation into components. That means

$$E[X+Y] = E[X] + E[Y]$$

for any two random variables X and Y.

Notice that we used linearity of expectation for the proof in the previous section. We will take some more examples now.

First, let us look at the example of Ramsey number in the light of expectation. Suppose we pick each edge of G (on n vertices) uniformly at random. Define T to be the random variable which counts the number of clique/independent sets in the graph. We are interested in the expectation of T.

Define T_i (for i from 1 to $\binom{n}{k}$) to be the random variable which assigns 1 if a particular subgraph is clique/independent set otherwise 0. Convince yourself that $T = \sum_i T_i$.

Note 1. The random variables T_i are dependent on each other.

Then,

$$E[T] = \sum_i E[T_i] = \sum_i 2^{1-\binom{k}{2}} = \binom{n}{k} 2^{1-\binom{k}{2}}.$$

If $E[T] < 1$ then there exist a graph which has less than or equal to $E[T]$ number of clique/independent sets. Since number of clique/independent sets is an integer, there exist a graph for which number of clique/independent sets is zero.

Let's take another example of probabilistic method which utilizes linearity of expectation.

2.1 Sum-free subsets

Let's take another example. Given a set of integers S, $S+S$ is defined as the subset of integers which contain all possible sums of pair of elements in S.

$$S + S = \{t : t = s_1 + s_2,\ s_1, s_2 \in S\}$$

A set S is called *sum-free*, if S does not contain any element of $S + S$.

Exercise 5. Construct a set of 10 elements which is sum-free. Construct a set of n elements which is sum-free.

Using probabilistic method, we will show that every large subset of integers contain a big enough subset which is sum-free.

Theorem 2. *For any subset S of n non-zero integers, There exist a subset of S which is sum-free and has size more than $n/3$.*

Exercise 6. Again, notice that the statement of theorem does not have a randomized setting. Can you think of a proof?

Proof. Suppose $S = \{s_1, s_2, \cdots, s_n\}$. The idea would be to map S to $rS = \{rs_1, rs_2, \cdots, rs_n\}$ for a random r. If some subset of rS is sum-free then the corresponding set in S will also be sum-free.

Though, taking r to be uniformly at random from $\mathbb{Z}$ is not feasible. First pick a prime p of the form $3k + 2$, such that, p is at least 3 times bigger than the absolute value of any element of S. We will pick a random r from the set $\{1, 2, \cdots, p\}$.

You will show in the assignment that there are infinite primes of the form $3k + 2$. We will do the calculations modulo p.

Notice that the set $T = \{k + 1, k + 2, \cdots, 2k + 1\}$ is a sum-free subset when we do addition modulo $p = 3k + 2$.

For applying the probabilistic method, pick a random $r \in \{1, 2, \cdots, p - 1\}$ and consider the set $rS \mod p = \{rs_1 \mod p, rs_2 \mod p, \cdots, rs_n \mod p\}$.

Exercise 7. Show that if we pick an r at random from $0, 1, \cdots, p - 1$ then rs_1 is also random with uniform probability.

Define a random variable Y which is the intersection size of $rS \mod p$ and T.

Using linearity of expectation,

$$E[Y] = \sum_i E[rs_i \mod p \in T].$$

Exercise 8. Show that $E[Y] = \frac{|S|}{3}$.

This implies that there exist at least one r for which $rS \mod p \cap T$ is of size at least $|S|/3$. Call that particular r, r_0. Then $T' = r_0 S \mod p \cap T$ is sum-free when addition is considered modulo p (T is sum-free). This implies that the pre-image in S which maps to T' is sum-free.

Exercise 9. Show that $r_0^{-1} T'$ is sum-free with respect to addition over integers.

□

We will take one more example of probabilistic method which utilizes linearity of expectation.

2.2 Vectors with small length

Theorem 3. *Given n unit vectors $v_i \in \mathbb{R}^n,\ i \in [n]$, there always exists a bit string $b \in \{-1,1\}^n$, such that,*

$$\left\| \sum_i b_i v_i \right\| \leq \sqrt{n}.$$

Proof. Again, we will pick b_i's uniformly at random from $\{-1,1\}$ and calculate the expected value of $N = \|\sum_i b_i v_i\|^2$.

From the definition of the length of a vector.

$$N = (\sum_i b_i v_i)^T (\sum_i b_i v_i) = \sum_{i,j} b_i b_j v_i^T v_j.$$

Notice that $v_i^T v_j$, the dot product between v_i and v_j, is a fixed number and the random variable are b_i's. Hence,

$$E[N] = \sum_{i,j} E[b_i b_j] v_i^T v_j.$$

By definition, we picked b_i and b_j independently. So b_i and b_j are independent if $i \neq j$. This implies that $E[b_i b_j] = E[b_i]E[b_j]$.

Exercise 10. Show that $E[b_i b_j] = 1$, if $i = j$ otherwise it is zero.

$$E[N] = \sum_i v_i^T v_i = n.$$

This implies that there is a choice of b_i's for which length of $\sum_i b_i v_i$ is less than or equal to $\sqrt{n}$.

□

Exercise 11. Given n unit vectors $v_i \in \mathbb{R}^n,\ i \in [n]$, there always exists a bit string $b \in \{-1,1\}^n$, such that,

$$\left| \sum_i b_i v_i \right| \geq \sqrt{n}.$$

3 Assignment

Exercise 12. Give a counting argument for Thm. 1.

Exercise 13. Read about Stirling's bound on factorial and binomial coefficients.

Exercise 14. Show that there are infinite primes of the form $3k + 2$.

Exercise 15. Suppose a vertex v has $deg(v)$ neighbors. Prove that the probability in a random permutation, v comes before any of its neighbors is, $\frac{1}{deg(v)+1}$.

Exercise 16. Consider a graph $G = (V, E)$. Show that G contains some independent set of size at least $\sum_{v \in V} \frac{1}{deg(v)+1}$.

Hint: Consider all permutations of $v_1, v_2, \cdots, v_n$. Take the independent set by considering vertices in the order of the permutation and once taken in the independent set, delete all its neighbors from the permutation.

Exercise 17. Let X be the random variable which counts the number of fixed points (i maps to i) in a random permutation. What is the expected value of X.

References

1. N. Alon and J. H. Spencer. The Probabilistic Method. *Wiley*, 2008.
2. H. Tijms Understanding Probability. *Cambridge University Press*, 2012.
3. D. Stirzaker. Elementary Probability. *Cambridge University Press*, 2003.
4. U. Schöning. Gems of Theoretical Computer Science. *Springer-Verlag*, 1998.