CSE 521: Design and Analysis of Algorithms I **Spring 2023**

Lecture 1: The Probabilistic Method

Lecturer: Shayan Oveis Gharan *03-28-2023* *Scribe:*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

1.1 Introduction to the Probabilistic Method

An old math puzzle goes: Suppose there are six people in a room; some of them shake hands. Prove that there are at least three people who all shook each others' hands or three people such that no pair of them shook hands. Generalized a bit, this is the classic Ramsey problem. The diagonal Ramsey numbers $R(k)$ are defined as follows. $R(k)$ is the smallest integer n such that in every two-coloring of the edges of the complete graph K_n by red and blue, there is a monochromatic copy of K_k, i.e. there are k nodes such that all of the $\binom{k}{2}$ edges between them are red or all of the edges are blue. A solution to the puzzle above asserts that $R(3) \leq 6$ (and it is easy to check that, in fact, $R(3) = 6$).

In 1929, Ramsey proved that $R(k)$ is finite for every k. We want to show that $R(k)$ must grow pretty fast; in fact, we'll prove that for $k \geq 3$, we have $R(k) > 2^{k/2}$. This requires finding a coloring of K_n that doesn't contain any monochromatic K_k. To do this, we'll use the probabilistic method: We'll give a random coloring of K_n and show that it satisfies our desired property with positive probability. This proof appeared in a paper of Erdös from 1947, and this is the example that starts Alon and Spencer's famous book devoted to the probabilistic method which will be one of the main resources for this coruse.

Lemma 1.1. *If* $\binom{n}{k} 2^{1-\binom{k}{2}} < 1$*, then* $R(k) > n$*. In particular,* $R(k) > 2^{k/2}$ *for* $k \geq 3$*.*

Proof. Consider a uniformly random 2-coloring of the edges of K_n. Every edge is colored red or blue independently with probability half each. For any fixed set of k vertices H, let E_H denote the event that the induced subgraph on H is monochromatic. An easy calculation yields

$$\mathbb{P}[E_H] = 2 \cdot 2^{-\binom{k}{2}}.$$

Since there are $\binom{n}{k}$ possible choices for H, we can use the union bound:

$$\mathbb{P}[\exists H \text{ s.t., } E_H] \leq 2 \cdot 2^{-\binom{k}{2}} \cdot \binom{n}{k}.$$

Thus if $2^{1-\binom{k}{2}}\binom{n}{k} < 1$, then with positive probability, no event E_H occurs. Thus there must exist at least one coloring with no monochromatic K_k. One can check that if $k \geq 3$ and $n = 2^{k/2}$, then this is satisfied. □

In the proof, we employed the following fundamental tool:

Fact 1.2 (Union Bound)**.** *If* $A_1, A_2, \ldots, A_m$ *are arbitrary events, then* $\mathbb{P}[A_1 \cup A_2 \cup \cdots \cup A_m] \leq \mathbb{P}[A_1] + \mathbb{P}[A_2] + \cdots + \mathbb{P}[A_m]$.

1.2 Linearity of Expectations

Fact 1.3 (Linearity of Expectation)**.** *If $X_1, X_2, \ldots, X_n$ are real-valued random variables, then*

$$\mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n].$$

The great fact about this inequality is that we don't need to know anything about the relationships between the random variables; linearity of expectation holds no matter what the dependence structure is.

Let's consider a 3-CNF formula over the variables $x_1, x_2, \ldots, x_n$. Such a formula has the form $C_1 \wedge C_2 \wedge \cdots \wedge C_m$ where each clause is an **or** of three literals involving distinct variables: $C_i = z_{i,1} \vee z_{i_2} \vee z_{i,3}$. A literal is a variable or its negation. For instance, $(x_2 \vee \bar{x}_3 \vee \bar{x}_4) \wedge (x_3 \vee \bar{x}_5 \vee \bar{x}_1) \wedge (x_1 \vee x_5 \vee x_4)$ is a 3-CNF formula.

Lemma 1.4. *If ϕ is a 3-CNF formula with m clauses, then there exists an assignment that makes at least $7m/8$ clauses evaluate to true.*

Proof. We will prove this using the probabilistic method. For every variable independently, we choose a uniformly random truth assignment: true or false each with probability 1/2. Let A_i equal 1 if clause C_i is satisfied by our random assignment, and equal 0 otherwise. Then $\mathbb{P}[Ai = 1] \geq 7/8$ because there are 7 ways to satisfy a clause out of the 8 possible truth values for its literals. Let $A = A_1 + \cdots + A_m$ denote the total number of satisfied clauses. By linearity of expectation,

$$\mathbb{E}[A] = \sum_i \mathbb{E}[A_i] = 7m/8.$$

So, there must be an assignment that satisfies this many clauses. □

1.3 Method of Conditional Expectations

The above lemma asserts that there exists an assignment satisfying at least $7m/8$ many clauses, but what if we wish to actually find one? One way is to randomly sample from the underlying distribution and then check the resulting assignment. Analyzing the probability of success will require our tail bounds which we will discuss in future lectures.

In this section, we will discuss a generic method that can turn many of the probabilistic method proofs into **even deterministic** algorithms. Let $S(x_1, x_2, \ldots, x_n)$ denote the expected number of satisfied clauses given a partial truth assignment to the input variables, where we choose the unassigned variables uniformly at random. We will use T to denote true, F to denote false, and * to denote that no assignment has been chosen for that variable. For instance, $S(\star, \star, \ldots, \star)$ denotes the expected number of satisfied clauses in a random assignment, and we have already seen that

$$S(\star, \star, \ldots, \star) = 7m/8.$$

Note that a simple linear-time algorithm can estimate $S(x_1, x_2, ..., x_n)$ for any partial assignment $x_1, ..., x_n \in \{T, F, \star\}$ by simply going through the clauses one by one.

As an example, consider the clause $x_1 \vee \bar{x} + 2 \vee \bar{x}_4$. The probability that a random assignment satisfies this is 7/8. If we assign $x_1 = F$, then the probability becomes 3/4, and if we set $x_1 = T$, then the probability becomes 1. Observe that

$$S(\star, \star, \ldots, \star) = \frac{1}{2} S(F, \star, \ldots, \star) + \frac{1}{2} S(T, \star, \ldots, \star).$$

Consequently, it must hold that

$$\max\{S(F,\star,\ldots,\star), S(T,\star,\ldots,\star)\} \geq S(\star,\star,\ldots,\star).$$

As we have just argued, it's possible to compute both these quantities and figure out which is larger. We can then set x_1 to the corresponding value and keep assigning truth values recursively. Since the value of S never goes down and it starts at $7m/8$, when the algorithm finishes we must satisfy **at least** 7m/8 fraction of clauses. Note that the algorithm may indeed satisfy more than $7m/8$ fraction of clauses.

1.4 Choosing the Right Distribution

Here is a more complicated example in which the choice of distribution requires a preliminary lemma. Let $V = V_1 \cup \cdots \cup V_k$, where the V_i's are disjoint sets, each of size n. Let $h : V_k \to \{\pm 1\}$ be a two-coloring of the k-sets. A k-set E is **crossing** if it contains precisely one point from each V_i. For $S \subset V$ set

$$h(S) := \sum_{E \in \binom{V}{k}} h(E). \tag{1.1}$$

Theorem 1.5. *Suppose $h(E) = +1$ for all crossing k-sets E. Then there is an $S \subset V$ for which*

$$h(S) \geq c_k n^k$$

Here c_k is a **positive** *constant, which is independent of n.*

Perhaps, the first attempt is to choose each element of V in S, independently, with probability $1/2$. It turns $\mathbb{E}[h(S)]$ for such a distribution can be even negative, e.g., assume $h(E) = -1$ for every non-crossing k-set. If you think about it deeply, you would wonder why $1/2$? As we will see, choosing elements of S independently is right, but we need to be careful on the marginals; we want to choose the marginals based on the function $h(.)$ given to us.

But how? Let $p_1, \ldots, p_k$ be the marginals of elements of $V_1, \ldots, V_k$ to be determined, i.e., we sample elements in S independently but elements from the same V_i are chosen with the same marignals. Given $p_1, \ldots, p_k$, we define a random set R where for every element $x \in V_i$, we add x to S with probability p_i, independent of every other element.

Define a random variable

$$X_R := h(R). \tag{1.2}$$

It turns out that we can write $\mathbb{E}[X_R]$ as a k-homogeneous polynomial in $p_1, \ldots, p_k$:

$$\begin{aligned}
\mathbb{E}[X_R] &\underset{(1.2)}{=} \sum_S \mathbb{P}[R = S]\, h(S) \\
&\underset{(1.1)}{=} \sum_{E \in \binom{V}{k}} \mathbb{P}[E \subseteq R] \cdot h(E) \\
&= \sum_{\substack{a_1,\ldots,a_k \in \mathbb{N}^k \\ \sum_i a_i = k}} \sum_{E : |E \cap V_i| = a_i} \prod_{i=1}^{k} p_i^{a_i} \cdot h(E) \\
&= \sum_{\substack{a_1,\ldots,a_k \in \mathbb{N}^k \\ \sum_i a_i = k}} \prod_{i=1}^{k} p_i^{a_i} \underbrace{\left(\sum_{E : |E \cap V_i| = a_i, \forall i} h(E) \right)}_{=: c_{a_1,\ldots,a_k}} = q(p_1, \ldots, p_k).
\end{aligned}$$

In the third equality, we classify all k-sets by their "type" namely the size of their intersections with $V_1, \ldots, V_k$.

To prove the theorem, we need to show that there is a choice of R such that $h(R) \geq c_k n^k$. By the probabilistic method it is enough to show that $\mathbb{E}[X_R] \geq c_k n^k$. By the above equation it is enough to show that there is a choice of $p_1, \ldots, p_k$ such that $q(p_1, \ldots, p_k) \geq c_k n^k$. That is what we show in the rest of the proof.

In the above equations, we get the multivariate polynomial q in terms of $p_1, \ldots, p_k$. The following properties of q are immediate:

- q is k-homogeneous; i.e., every monomial of q has degree k.
- Since $h(E) = +1$ for all crossing sets, we have $c_{1,\ldots,1} = |V_1 \times \cdots \times V_k| = n^k$.
- For every $a_1, \ldots, a_k \in [k]^k$ with $\sum_i a_i = k$, we have

$$c_{a_1,\ldots,a_k} \underset{h(E)=\pm 1, \forall E}{\leq} \sum_{E : |E \cap V_i| = a_i, \forall i} +1 \leq n^k$$

Finally, by the following fact, there exists a choice of $p_1, \ldots, p_k$ such that $q(p_1, \ldots, p_k)/n^k \geq c_k$ as desired.

Fact 1.6. *Let P_k denote the set of all k-homogeneous polynomials $f(p_1, \ldots, p_k)$ of with all coefficients having absolute value at most one and $p_1, \ldots, p_k$ having coefficient exactly one. Then, for all $f \in P_k$ there exist $p_1, \ldots, p_k \in [0, 1]$ with*

$$|f(p_1, \ldots, p_k)| \geq c_k$$

where $c_k > 0$ is an absolute constant only as a function of k.

Proof. Set

$$M(f) := \max_{p_1,\ldots,p_k \in [0,1]} |f(p_1, \ldots, p_k)|.$$

The main observation is that for any $f \in P_k$, $M(f) > 0$. This is simply because f is not the identically zero polynomial (it has one non-zero monomial. So over a field of size at least the degree of f, it cannot evaluate to zero. Lastly, we observe that P_k is compact and $M : P_k \to \mathbb{R}$ is a continuous map. So M must have a minimum value that is non-zero. □

1.5 The Alteration Method

Sometimes in our probabilistic method proof, we may not directly obtain the object of interest. Instead, we may try to sample a "good enough" object and then show that by a small number tweaks we can turn the object into a feasible object.

Recall that $R(k)$ is the smallest integer n such that in every two coloring of the edges of the complete graph K_n by red and blue there is a monochromatic copy of K_k. The following is a stronger variant of Lemma 1.1

Theorem 1.7. *For any integer n and k, if $R(k) > n - \binom{n}{k} 2^{1-\binom{k}{2}}$.*

Proof. As in Lemma 1.1, consider a uniformly random 2-coloring of the edges of K_n. Let E_H be the even that the subgraph on H is monochromatic. Let

$$X = \sum_{H \in \binom{n}{k}} E_H,$$

be the number of monochromatic copies of K_k in our two-colored graph. By linearity of expectations,

$$\mathbb{E}[X] = \sum_{H \in \binom{n}{k}} \mathbb{E}[E_H] = 2^{1-\binom{k}{2}} \cdot \binom{n}{k}$$

Now, it follows that there must exists a two-coloring such that the number of monochromatic copies of K_k is **at most** $\mathbb{E}[X]$. Consider such a coloring.

Now, we discuss the alteration part: We know that we have (at most) $\lfloor \mathbb{E}[X] \rfloor$ copies of K_k. We are going to delete one (arbitrary) vertex from each of these copies. Note that in principal these copies may share vertices so we may be able to delete all of them by removing a few vertices, but in the worst case, these copies are disjoint. So, we can delete all of them by removing at most $\lfloor \mathbb{E}[X] \rfloor$ vertices of G. The resulting graph has at least $n - \binom{n}{k}2^{1-\binom{k}{2}}$ vertices and has no copies of K_k. □

Now, we are left with the "calculus" problem of for what values of n, can we optimize the inequality. It turns out with a bit of calucluations that

$$R(k) > \frac{k}{e}2^{k/2}.$$

This is slightly better than what we can show with Lemma 1.1, that $R(k) > \frac{k}{e\sqrt{2}}2^{k/2}$.

In future lectures, we will see how to use a more sophisticated technique, called the Lovasz Local lemma, to get a slightly better bound.