CS174 Fall 98: Lecture Note 5

Alistair Sinclair, September 1998; based on earlier notes by Manuel Blum/Douglas Young.

The Probabilistic Method

- Suppose you are allowed to color the edges of the complete graph on 1000 vertices with two colors. Can you do this in such a way that there is no set of 20 vertices all of whose internal edges have the same color? (Such a set is called a "monochromatic complete subgraph.")

This kind of question is typical of a part of Combinatorics known as Ramsey Theory.

The answer to the above question is "yes." A beautifully elegant proof is provided by the *probabilistic method.*

Idea: Rather than trying to construct an explicit coloring, we instead consider a random coloring and show that it has the desired property with non-zero probability.

Theorem 1: If $n \leq 2^{\frac{k}{2}}$, then it is possible to color the edges of the complete graph K_n on n vertices in such a way that there is no monochromatic complete subgraph of size k.

Proof: Consider a random coloring of the edges of K_n. Let S be (the subgraph defined by) any subset of k vertices. Then

$$\Pr[S \text{ is monochromatic}] = 2^{-\binom{k}{2}+1}. \qquad \text{(Why?)}$$

But there are only $\binom{n}{k}$ different such sets S, so

$$\begin{aligned}
\Pr[\textit{some } S \text{ is monochromatic}] &\leq \tbinom{n}{k} \cdot 2^{-\binom{k}{2}+1} \\
&\leq \tfrac{n^k}{k!} \cdot 2^{-\binom{k}{2}+1} \\
&\leq \tfrac{1}{k!} \cdot 2^{k^2/2-\binom{k}{2}+1} \qquad (\text{since } n \leq 2^{\frac{k}{2}}) \\
&= \tfrac{2^{\frac{k}{2}+1}}{k!} \\
&< 1 \quad (\text{for } k \geq 3).
\end{aligned}$$

Hence $\Pr[\text{no } S \text{ is monochromatic}] > 0$.

Since a random coloring has the desired property with probability > 0, there must exist a coloring with the property, which proves the theorem. □

Ex: Check that the theorem implies that the answer to our original question is "yes." □

So, the probabilistic method consists of proving the existence of an object having certain properties by showing that random objects have that property.

Does the method help us to construct such objects? Not immediately. All we have shown above is that a good coloring exists. However, if we look more closely at the proof we see that

$$\Pr[\text{coloring is "bad"}] \leq \frac{2^{\frac{k}{2}+1}}{k!},$$

which is tiny as long as k isn't too small. So, if we just take a random coloring we can be very confident that it has the desired property (though unfortunately we don't have an efficient way to *check* that the random coloring has the property).

Ex: For the example above ($n = 1000$, $k = 20$), give an upper bound on the probability that a random coloring is bad. □

Tournaments and Hamilton Paths

Definition: A tournament graph is a directed graph $G = (V, E)$ (without self-loops) such that

$$\forall i, j \in V \quad \text{exactly one of } (i,j) \text{ and } (j,i) \text{ is in } E.$$

Tournament graphs first appeared in an old biology paper studying the pecking order of chickens. The paper pointed out that the tournament graph of chickens always has a Hamilton path!

(Recall that a Hamilton path in a directed graph $G = (V, E)$ with vertex set $V = \{1, 2, \ldots, n\}$ is a permutation $(i_1, i_2, \ldots, i_n)$ of V such that $(i_j, i_{j+1}) \in E$ for all $j = 1, 2, \ldots, n-1$.)

Easy(ish) Theorem: every TG has a Hamilton path. $\square$

The proof of the theorem is left as an exercise (which doesn't make any use of probability). Hint: suppose you have constructed a Hamilton path on the first k vertices; show how to extend it to include vertex $k+1$.

Ex: Is there a TG with only one Hamilton path? $\square$

In the first ever published application of the probabilistic method, Szele proved in 1943 the following remarkable fact:

Theorem 2: For every n, there exists a TG with n vertices and at least $\frac{n!}{2^{n-1}}$ Hamilton paths.

Ex: To gain some appreciation of why this fact is remarkable, try constructing TGs with this many Hamilton paths. For example, for $n = 6$ the above expression evaluates to 22.5, so try constructing a TG with 6 vertices and at least 23 Hamilton paths. $\square$

Proof: Let G be a random TG on n vertices. (What is the size of the sample space? How would you construct such a G?)
Define the r.v. $X = \#$ Hamilton paths in G.
Then we can write $X = \sum_\pi X_\pi$, where the sum is over all $n!$ permutations π of the vertices, and X_π is the r.v.

$$X_\pi = \begin{cases} 1 & \text{if } \pi \text{ is a Hamilton path;} \\ 0 & \text{otherwise.} \end{cases}$$

What is $\mathrm{E}(X_\pi)$? Well,

$$\mathrm{E}(X_\pi) = \Pr[\pi \text{ is a Hamilton path}] = \tfrac{1}{2^{n-1}}.$$

(Why is this?)
Therefore, we have

$$\mathrm{E}(X) = \textstyle\sum_\pi \mathrm{E}(X_\pi) = \sum_\pi \frac{1}{2^{n-1}} = \frac{n!}{2^{n-1}}.$$

But now we are done, since there must surely be some point in the sample space (i.e., some TG) on which X has value at least $\mathrm{E}(X)$. (Why?) $\square$

Ex (harder): What about actually finding such a TG? All we know from Theorem 2 is that there is some TG with many Hamilton paths; since there are $2^{\binom{n}{2}}$ TGs, a random TG might have this property with probability as tiny as $2^{-\binom{n}{2}}$. But suppose we knew more. For example, suppose we could also show that no TG has more than $\frac{n!}{2^{n-1}}$ Hamilton paths. Then it must follow from our above expectation result that none has less, either, so every TG has exactly $\frac{n!}{2^{n-1}}$ Hamilton paths. This supposition is in fact false, but something close to it is true: Noga Alon has proved that no TG has more than $c \cdot n^{3/2} \cdot \frac{n!}{2^{n-1}}$ Hamilton paths, where c is a constant. Use this fact to show that a random TG has probability at least $\frac{1}{2cn^{3/2}}$ (which is non-trivial) of containing at least $\frac{n!}{2^n}$ Hamilton paths. (Hint: show that if a r.v. X has $\mathrm{E}(X) = \mu$ and always takes values in the range $[0, .., M]$, then $\Pr[X \geq B] \geq \frac{\mu - B}{M - B}$, for any $B \leq \mu$.) $\square$

MAX-SAT

The MAX-SAT problem is the following:

Input : A Boolean formula ϕ in conjunctive normal form.

Question : Find a truth assignment for the variables of ϕ that satisfies the largest possible number of clauses.

Solving MAX-SAT exactly is NP-hard, so we cannot expect a polynomial time algorithm for it (unless P = NP). The probabilistic method gives a simple and efficient solution to this problem that is not necessarily optimal, but never too bad.

Theorem 3: For any Boolean formula ϕ as above, there is a truth assignment that satisfies at least half of the clauses of ϕ.

Proof: As usual, consider a random truth assignment.

For each clause i, introduce a r.v.

$$X_i = \begin{cases} 1 & \text{if clause } i \text{ is satisfied;} \\ 0 & \text{otherwise.} \end{cases}$$

Then $X = \sum_i X_i$ is the number of satisfied clauses.

But $\mathrm{E}(X_i) = 1 - 2^{-k}$, where k is the number of literals in clause i. (Why?) So $\mathrm{E}(X_i) \geq \frac{1}{2}$.

Therefore, we have $\mathrm{E}(X) = \sum_i \mathrm{E}(X_i) \geq \frac{m}{2}$, where m is the number of clauses.

So, there must exist an assignment that satisfies at least half of the clauses. □

Ex: By applying Markov's inequality to the r.v. $Y = m - X$, where m is the number of clauses in ϕ, show that, for a random truth assignment,

$$\Pr[\text{more than } \tfrac{1}{2}(1-\epsilon)m \text{ clauses are satisfied}] \geq 1 - \tfrac{1}{1+\epsilon}$$

for $0 < \epsilon < 1$. Hence design a polynomial-time algorithm that, with probability at least $\frac{1}{5}$, satisfies at least three eighths of the clauses in ϕ. How could you boost this probability to $1 - 10^{-6}$? (Hint: Consider independent repeated trials.) □

Ex: For a positive integer k, the problem MAX-kSAT is defined exactly as for MAX-SAT except that the input formula ϕ must have precisely k literals in every clause. MAX-kSAT is also known to be NP-hard for every $k \geq 2$.[1] Show that, for a MAX-3SAT formula ϕ, there always exists an assignment that satisfies at least seven eighths of the clauses of ϕ. □

From proofs to algorithms

Theorem 3 immediately gives us an efficient randomized algorithm for MAX-SAT which produces an assignment whose expected number of satisfied clauses is at least $\frac{m}{2}$: just output a random assignment.

The Exercise immediately following its proof shows how to get (almost) as good an approximation with high probability by repeated trials of the algorithm.

Can we do even better? Yes: we can derandomize the algorithm.

[1] Note the contrast here with the corresponding decision problem kSAT, in which we just have to decide whether it is possible to satisfy *all* the clauses of ϕ. The decision problem kSAT is NP-hard for all $k \geq 3$, but 2SAT is in P (as you may recall from CS170 as an application of finding strongly connected components).

View the algorithm that outputs a random assignment as a binary tree, as follows. Suppose ϕ has n variables $z_1, z_2, \ldots, z_n$. The two branches from the root correspond to setting $z_1 = \mathrm{T}$ or F respectively. Similarly, branches at level i correspond to the two possible choices for z_i. The algorithm chooses a random path from the root to one of the 2^n leaves, each of which corresponds to a complete assignment.

Label the nodes of the tree with formulae as follows. The root r has label $\phi_r = \phi$. A node v at level i has formula ϕ_v equal to ϕ with the appropriate values substituted for $z_1, \ldots, z_{i-1}$ (as determined by the path from r to v). Note that the tree below node v corresponds to a similar random process starting from the formula ϕ_v (except that ϕ_v may contain the constants T and F).

Define the r.v. $Y_v = \#$ of satisfied clauses in process starting from ϕ_v (including clauses in ϕ_v with constant value T).
Then Y_r is the r.v. X in the proof of Theorem 3, and $\mathrm{E}(Y_r) \geq \frac{m}{2}$.
Call a node v good if $\mathrm{E}(Y_v) \geq \frac{m}{2}$. We show how to find a path from the root to a leaf using only good nodes. If we can do this, we will arrive at a good leaf v, for which with $\mathrm{E}(Y_v) \geq \frac{m}{2}$. But now ϕ_v contains no further variables, so we have constructed an assignment with at least $\frac{m}{2}$ satisfied clauses. (Do you understand this?)

The key observation is the following: if node v has children v_1, v_2, then $\max\{\mathrm{E}(Y_{v_1}), \mathrm{E}(Y_{v_2})\} \geq \mathrm{E}(Y_v)$. Thus, if v is good, then at least one of its children must be good also.
To see this claim, suppose v is on level i; then v_1, v_2 correspond to setting the variable $z_i = \mathrm{T}/\mathrm{F}$ respectively. We have

$$\begin{aligned}
\mathrm{E}(Y_v) &= \mathrm{E}(Y_v | z_i = \mathrm{T}) \Pr[z_i = \mathrm{T}] + \mathrm{E}(Y_v | z_i = \mathrm{F}) \Pr[z_i = \mathrm{F}] \\
&= \tfrac{1}{2}(\mathrm{E}(Y_{v_1}) + \mathrm{E}(Y_{v_2})) \\
&\leq \max\{\mathrm{E}(Y_{v_1}), \mathrm{E}(Y_{v_2})\}.
\end{aligned}$$

How do we identify a good child? Well, we just evaluate $\mathrm{E}(Y_{v_1})$ and $\mathrm{E}(Y_{v_2})$ and pick the child with the larger value (breaking ties arbitrarily). Evaluating each of these expectations is easy, using the technique in the proof of Theorem 3.

So, the probabilistic method has led us to an efficient deterministic algorithm that is guaranteed to satisfy at least half of the clauses.

Ex: Show how to evaluate $\mathrm{E}(Y_v)$ exactly for any node v in the above tree. (You should use indicator r.v.'s X_i as in the proof of Theorem 3.) □

Ex: Explain carefully why the following alternative method of choosing a path through the tree is also guaranteed to satisfy at least half of the clauses: at each level, pick the child for which the number of newly satisfied clauses is larger (again breaking ties arbitrarily). □

Ex: Use a similar method to construct a deterministic algorithm that, for every MAX-3SAT formula ϕ, produces an assignment that satisfies at least seven eighths of the clauses. □