The Probabilistic Method

497 - Randomized Algorithms

Sariel Har-Peled

September 30, 2002

"Shortly after the celebration of the four thousandth anniversary of the opening of space, Angary J. Gustible discovered Gustible's planet. The discovery turned out to be a tragic mistake.

Gustible's planet was inhabited by highly intelligent life forms. They had moderate telepathic powers. They immediately mind-read Angary J. Gustible's entire mind and life history, and embarrassed him very deeply by making up an opera concerning his recent divorce."

— From Gustible's Planet, Cordwainer Smith

1 Introduction

The probabilistic method is a combinatorial technique to use probabilistic algorithms to create objects having desirable properties, and furthermore, prove that such objects exist. The basic technique is based on two basic observations:

1. If $\mathbf{E}[X] = \mu$, then there exists a value x of X, such that $x \geq \mathbf{E}[X]$.
2. If the probability of event $\mathcal{E}$ is larger than zero, then $\mathcal{E}$ exists and it is not empty.

The surprising thing is that despite the elementary nature of those two observations, they lead to a powerful technique that leads to numerous nice and strong results. Including some elementary proofs of theorems that previously had very complicated and involved proofs.

The main proponent of the probabilistic method, was Paul Erdős. An excellent text on the topic is the book by Noga Alon and Joel Spencer [AS00].

This topic is worthy of its own course. The interested student is refereed to the course "Math 475 — The Probabilistic Method".

1.1 Examples

Theorem 1.1 *For any undirected graph $G(V, E)$ with n vertices and m edges, there is a partition of the vertex set V into two sets A and B such that*

$$\left| \left\{ uv \in E \;\middle|\; u \in A \text{ and } v \in B \right\} \right| \geq \frac{m}{2}.$$

Proof: Consider the following experiment: randomly assign each vertex to A or B, independently and equal probability.

For an edge $e = uv$, the probability that one endpoint is in A, and the other in B is $1/2$, and let X_e be the indicator variable with value 1 if this happens. Clearly,

$$\mathbf{E}\left[\left|\left\{uv \in E \;\middle|\; u \in A \text{ and } v \in B\right\}\right|\right] = \sum_{e \in E(G)} \mathbf{E}[X_e] = \sum_{e \in E(G)} \frac{1}{2} = \frac{m}{2}.$$

Thus, there must be a partition of V that satisfies the theorem. ■

Definition 1.2 For a vector $v = (v_1, \ldots, v_n) \in \mathbb{R}^n$, $\|v\|_\infty = \max_i |v_i|$.

Theorem 1.3 *Let A be an $n \times n$ binary matrix (i.e., each entry is either 0 or 1), then there always exists a vector $b \in -1, +1^n$ such that $\|Ab\|_\infty \leq 4\sqrt{n \log n}$.*

Proof: Let $v = (v_1, \ldots, v_n)$ be a row of A. Chose a random $b = (b_1, \ldots, b_n) \in -1, +1^n$. Let $i_1, \ldots, i_m$ be the indices such that $v_{i_j} = 1$. Clearly,

$$\mathbf{E}[v \cdot b] = \sum_i \mathbf{E}[v_i b_i] = \sum_j v_i \, \mathbf{E}\left[b_{i_j}\right] = 0.$$

Let $X_j = 1$ if $b_{i_j} = +1$, for $j = 1, \ldots, m$. We have $\mathbf{E}\left[\sum_j X_j\right] = n/2$, and

$$\begin{aligned}
\mathbf{Pr}\left[|v \cdot b| \geq 4\sqrt{n \ln n}\right] = 2\,\mathbf{Pr}\left[v \cdot b \leq -4\sqrt{n \ln n}\right] &= 2\,\mathbf{Pr}\left[\sum_j X_j - \frac{n}{2} \leq -2\sqrt{n \ln n}\right] \\
&= 2\,\mathbf{Pr}\left[\sum_j X_j < \left(1 - 4\sqrt{\frac{\ln n}{n}\frac{n}{m}}\right)\frac{m}{2}\right] \\
&\leq 2\exp\left(-\frac{m}{2}\left(4\sqrt{\frac{\ln n}{n}\frac{n}{m}}\right)^2\right) = 2\exp\left(-\frac{m}{2}\left(16\frac{n \ln n}{m^2}\right)\right) \\
&= 2\exp\left(-\frac{8n \ln n}{m}\right) \\
&\leq 2\exp(-8 \ln n) = \frac{2}{n^8}
\end{aligned}$$

by the Chernoff inequality and symmetry. Thus, the probability that any entry in Ab exceeds $\sqrt{4n \ln n}$ is smaller than $2/n^7$. Thus, with probability at least $1 - 2/n^7$, all the entries of Ab have value smaller than $4\sqrt{n \ln n}$.

In particular, there exists a vector $b \in \{-1, +1\}^n$ such that $\|Ab\|_\infty \leq 4\sqrt{n \ln n}$. ■

2 Maximum Satisfiability

Theorem 2.1 *For any set of m clauses, there is a truth assignment of variables that satisfies at least $m/2$ clauses.*

Proof: Assign every variable a random value. Clearly, a clause with k variables, has probability $1 - 2^{-k}$ to be satisfied. Using linearity of expectation, and the fact that even clause has at least one variable, it follows, that $\mathbf{E}[X] = m/2$, where X is the random variable counting the number of clauses being satisfied. In particular, there exists an assignment for which $X \geq m/2$. ■

For an instant I, let $m_{\text{opt}}(I)$, denote the maximum number of clauses that can be satisfied by the "best" assignment. For an algorithm A, let $m_A(I)$ denote the number of clauses satisfied computed by the algorithm A. The *approximation factor* of A, is $m_A(I)/m_{\text{opt}}(I)$. Clearly, the algorithm of Theorem 2.1 provides us with 1/2-approximation algorithm.

For every clause, C_j in the given instance, let $z_j \in \{0,1\}$ be a variable indicating whether C_j is satisfied or not. Similarly, let $x_i = 1$ if the i-th variable is being assigned the value TRUE. Let C_j^+ be indices of the variables that appear in C_j in the positive, and C_j^- the indices of the variables that appear in the negative. Clearly, to solve MAX-SAT, we need to solve:

$$\begin{array}{ll} \textit{maximize} & \displaystyle\sum_{j=1}^{m} z_j \\ \textit{subject to} & y_i, z_j \in \{0,1\} \text{ for all } i,j \\ & \displaystyle\sum_{i \in C_j^+} y_i + \sum_{i \in C_j^-} (1 - y_i) \geq z_j \text{ for all } j. \end{array}$$

We relax this into the following linear program:

$$\begin{array}{ll} \textit{maximize} & \displaystyle\sum_{j=1}^{m} z_j \\ \textit{subject to} & 0 \leq y_i, z_j \leq 1 \text{ for all } i,j \\ & \displaystyle\sum_{i \in C_j^+} y_i + \sum_{i \in C_j^-} (1 - y_i) \geq z_j \text{ for all } j. \end{array}$$

Which can be solved in polynomial time. Let $\widehat{\cdot}$ denote the values assigned to the variables by the linear-programming solution. Clearly, $\sum_{j=1}^{m} \widehat{z_j}$ is an upper bound on the number of clauses of I that can be satisfied.

We set the variable y_i to 1 with probability $\widehat{y_i}$. This is called *randomized rounding.*

Lemma 2.2 *Let C_j be a clause with k literals. The probability that it is satisfied by randomized rounding is at least $\beta_k \widehat{z_j} \geq (1 - 1/e)\widehat{z_j}$, where $\beta_k = 1 - (1 - 1/k)^k$.*

Proof: Assume $C_j = y_1 \vee v_2 \ldots \vee v_k$. By the LP, we have $\widehat{y_1} + \cdots + \widehat{y_k} \geq \widehat{z_j}$. Furthermore, the probability that C_j is not satisfied is $\prod_{i=1}^{k}(1 - \widehat{y_i})$. Note that $1 - \prod_{i=1}^{k}(1 - \widehat{y_i})$ is minimized when all the $\widehat{y_i}$'s are equal (by symmetry). Namely, when $\widehat{y_i} = \widehat{z_j}/k$. Consider the function $f(x) = 1 - (1 - x/k)^k$. This is a concave function, which is larger than $g(x) = \beta_k x$ for all $0 \leq x \leq 1$, as can be easily verified, by checking the inequality at $x = 0$ and $x = 1$.

Thus,

$$\mathbf{Pr}[C_j \text{ is satisfied}] = 1 - \prod_{i=1}^{k} (1 - \widehat{y_i}) \geq f(\widehat{z_j}) \geq \beta_k \widehat{z_j}.$$

The second part of the inequality,m follows from the fact that $\beta_k \geq 1 - 1/e$, for all $k \geq 0$. Indeed, for $k = 1, 2$ the claim trivially holds. Furthermore,

$$1 - \left(1 - \frac{1}{k}\right)^k \geq 1 - \frac{1}{e} \Leftrightarrow \left(1 - \frac{1}{k}\right)^k \leq \frac{1}{e} \Leftrightarrow 1/\left(1 - \frac{1}{k}\right)^k \geq e \Leftrightarrow \left(1 + \frac{1}{k-1}\right)^k \geq e$$

Note that for

$$u(x) = \left(1 + \frac{1}{x}\right)^x \text{ and}$$

$$u'(x) = \left(1 + \frac{1}{x}\right)^x \ln\left(1 + \frac{1}{x}\right) - \frac{1}{x}\left(1 + \frac{1}{x}\right)^{(x-1)} < 0,$$

for $x \geq 3$. Thus $u(x)$ is monotone, decreasing for $x \geq 3$ and $\lim_{x \to \infty} u(x) = e$. Thus, for $x \geq 3$, we have $u(x) \geq e$. We conclude that

$$(1 + 1/(k-1))^k \geq (1 + 1/(k-1))^{k-1} = u(k-1) \geq e,$$

as required. ■

Theorem 2.3 *Given an instance of MAX-SAT, the expected number of clauses satisfied by linear programming and randomized rounding is at least* $(1 - 1/e)$ *times the maximum number of clauses that can be satisfied on that instance.*

Theorem 2.4 *Let* n_1 *be the expected number of clauses satisfied by randomized assignment, and let* n_2 *be the expected number of clauses satisfied by linear programming followed by randomized rounding. Then,* $\max(n_1, n_2) \geq \frac{3}{4} \sum_j \widehat{z_j}$.

Proof: It is enough to show that $(n_1 + n_2)/2 \geq \frac{3}{4} \sum_j \widehat{z_j}$. Let S_k denote the set of clauses that contain k literals. We know that

$$n_1 = \sum_k \sum_{C_j \in S_k} \left(1 - 2^{-k}\right) \geq \sum_k \sum_{C_j \in S_k} \left(1 - 2^{-k}\right) \widehat{z_j}.$$

By Lemma 2.2 we have $n_2 \geq \sum_k \sum_{C_j \in S_k} \beta_k \widehat{z_j}$. Thus,

$$\frac{n_1 + n_2}{2} \geq \sum_k \sum_{C_j \in S_k} \frac{1 - 2^{-k} + \beta_k}{2} \widehat{z_j}.$$

One can verify that $\left(1 - 2^{-k}\right) + \beta_k \geq 3/2$, for all k, so that we have

$$\frac{n_1 + n_2}{2} \geq \frac{3}{4} \sum_k \sum_{C_j \in S_k} \widehat{z_j} = \frac{3}{4} \sum_j \widehat{z_j}.$$

■

References

[AS00] N. Alon and J. H. Spencer. *The probabilistic method.* Wiley Inter-Science, 2nd edition, 2000.