The Probabilistic Method V

497 - Randomized Algorithms

Sariel Har-Peled

October 15, 2002

Once I sat on the steps by a gate of David's Tower, I placed my two heavy baskets at my side. A group of tourists was standing around their guide and I became their target marker. "You see that man with the baskets? Just right of his head there's an arch from the Roman period. Just right of his head." "But he's moving, he's moving!" I said to myself: redemption will come only if their guide tells them, "You see that arch from the Roman period? It's not important: but next to it, left and down a bit, there sits a man who's bought fruit and vegetables for his family." — Yehuda Amichai, Tourists

1 The Method of Conditional Probabilities

In previous lecture, we encountered the following problem:

Problem 1.1 (Set Balancing) Given a binary matrix A of size $n \times n$, find a vector $b \in \{-1,+1\}^n$, such that $\|Ab\|_\infty$ is minimized.

Using random assignment and the Chernoff inequality, we showed that there exists b, such that $\|Ab\|_\infty \leq 4\sqrt{n \ln n}$. Can we derandomize this algorithm? Namely, can we come up with an efficient *deterministic* algorithm that has low discrepancy?

To derandomize our algorithm, construct a computation tree of depth n, where in the i-th level we expose the i-th coordinate of b. This tree T has depth n. The root represents all possible random choices, while a node at depth i, represents all computations when the first i bits are fixed. For a node $v \in T$, let $P(v)$ be the probability that a random computation starting from v succeeds. Let v_l, v_r be the two children of v. Clearly, $P(v) = (P(v_l)+P(v_r))/2$. In particular, $\max(P(v_l), P(v_r)) \geq P(v)$. Thus, if we could could compute $P(\cdot)$ quickly (and deterministically), then we could derandomize the algorithm.

Let C_i^+ be the bad event that $v_i \cdot b > 4\sqrt{n \log n}$, where v_i is the i-th row of A. Similarly, C_i^- is the bad event that $v_i \cdot b < -4\sqrt{n \log n}$, and let $C_i = C_i^+ \cup C_i^-$. Consider the probability, $\mathbf{Pr}\left[C_i^+ \,\middle|\, b_1, \ldots, b_k\right]$ (namely, the first k coordinates of b are specified). Let $v_i = (\alpha_1, \ldots, \alpha_n)$. We have that

$$\begin{aligned}\mathbf{Pr}\left[C_i^+ \,\middle|\, b_1, \ldots, b_k\right] &= \mathbf{Pr}\left[\sum_{i=k+1}^{n} b_i \alpha_i > 4\sqrt{n \log n} - \sum_{i=1}^{k} b_i \alpha_i\right] \\ &= \mathbf{Pr}\left[\sum_{i \geq k+1, \alpha_i \neq 0} b_i \alpha_i > L\right] = \mathbf{Pr}\left[\sum_{i \geq k+1, \alpha_i = 1} b_i > L\right],\end{aligned}$$

where $L = 4\sqrt{n \log n} - \sum_{i=1}^{k} b_i \alpha_i$. Let $V = \sum_{i \geq k+1, \alpha_i = 1} 1$. We have,

$$\mathbf{Pr}\left[C_i^+ \,\middle|\, b_1, \ldots, b_k\right] = \mathbf{Pr}\left[\sum_{\substack{i \geq k+1 \\ \alpha_i = 1}} (b_i + 1) > L + V\right] = \mathbf{Pr}\left[\sum_{\substack{i \geq k+1 \\ \alpha_i = 1}} \frac{b_i + 1}{2} > \frac{L + V}{2}\right],$$

The last probability, is the probability that in V flips of a fair coin we will get more than $(L+V)/2$ heads. Thus,

$$P_i^+ = \mathbf{Pr}\left[C_i^+ \,\middle|\, b_1, \ldots, b_k\right] = \sum_{i=\lceil (L+V)/2 \rceil}^{V} \binom{V}{i} \frac{1}{2^n} = \frac{1}{2^n} \left(\sum_{i=\lceil (L+V)/2 \rceil}^{V} \binom{V}{i} \right).$$

This implies, that we can compute P_i^+ in polynomial time! Indeed, we are adding $V \leq n$ numbers, each one of them is a binomial coefficient that has polynomial size representation in n, and can be computed in polynomial time (why?). One can define in similar fashion P_i^-, and let $P_i = P_i^+ + P_i-$. Clearly, P_i can be computed in polynomial time, by applying a similar argument to the computation of $P_i^- = \mathbf{Pr}\left[C_i^- \,\middle|\, b_1, \ldots, b_k\right]$.

For a node $v \in T$, let b_v denote the portion of b that was fixed when traversing from the root of T to v. Let $P(v) = \sum_{i=1}^{n} \mathbf{Pr}\left[C_i \,\middle|\, b_v\right]$. By the above discussion $P(v)$ can be computed in polynomial time. Furthermore, we know, by the previous result on set balancing that $P(r) < 1$ (thats was the bound used to show that there exist a good assignment).

As before, for any $v \in T$, we have $P(v) \geq \min(P(v_l), P(v_r))$. Thus, we have a polynomial *deterministic* algorithm for computing a set balancing with discrepancy smaller than $4\sqrt{n \log n}$. Indeed, set $v = \mathit{root}(T)$. And start traversing down the tree. At each stage, compute $P(v_l)$ and $P(v_r)$ (in polynomial time), and set v to the child with lower value of $P(\cdot)$. Clearly, after n steps, we reach a leaf, that corresponds to a vector b' such that $\|Ab'\|_\infty \leq 4\sqrt{n \log n}$.

Theorem 1.2 *Using the method of conditional probabilities, one can compute in polynomial time in n, a vertex b, such that $\|Ab\|_\infty \leq 4\sqrt{n \log n}$.*

Note, that this method might fail to find the best assignment.

2 Further Examples of the Probabilistic Method

2.1 High Girth and High Chromatic Number

Definition 2.1 For a graph G, let $\alpha(G)$ be the cardinality of the largest independent set in G, $\chi(G)$ denote the chromatic number of G, and let $\mathrm{girth}(G)$ denote the length of the shortest circle in G.

Theorem 2.2 *For all K, L there exists a graph G with* $\mathrm{girth}(G) > L$ *and* $\chi(G) > K$.

Proof: Fix $\mu < 1/L$, and let $G \approx G(n,p)$ with $p = n^{\mu-1}$; namely, G is a random graph on n chosen by picking each pair of vertices as an edge randomly and independently with probability p. Let X be the number of cycles of size at most L. Then

$$\mathbf{E}[X] = \sum_{i=3}^{L} \frac{n!}{(n-i)!} \cdot \frac{1}{2i} \cdot p^i \le \sum_{i=3}^{L} \frac{n^i}{2i} \cdot \left(n^{\mu-1}\right)^i \le \sum_{i=3}^{L} \frac{n^{\mu i}}{2i} = o(n),$$

as $\mu L < 1$, and since the number of different sequence of i vertices is $\frac{n!}{(n-i)!}$, and every cycle is being counted in this sequence $2i$ times.

In particular, $\mathbf{Pr}[X \ge n/2] = o(1)$.

Let $x = \left\lceil \frac{3}{p} \ln n \right\rceil + 1$. We have

$$\begin{aligned}\mathbf{Pr}[\alpha(G) \ge x] &\le \binom{n}{x}(1-p)^{\binom{x}{2}} < \left(n \exp\left(-\frac{p(x-1)}{2}\right)\right)^x < \left(n \exp\left(-\frac{3}{2}\ln n\right)\right)^x \\ &< (o(1))^x = o(1).\end{aligned}$$

Let n be sufficiently large so that both these events have probabilist less than $1/2$. Then there is a specific G with less than $n/2$ cycles of length at most L and with $\alpha(G) < 3n^{1-\mu}\ln n + 1$.

Remove from G a vertex from each cycle of length at most L. This gives a graph G^* with at least $n/2$ vertices. G^* has girth greater than L and $\alpha(G^*) \le \alpha(G)$ (any independent set in G^* is also an independent set in G). Thus

$$\chi(G^*) \ge \frac{|V(G^*)|}{\alpha(G^*)} \ge \frac{n/2}{3n^{1-\mu}\ln n} \ge \frac{n^\mu}{12 \ln n}.$$

To complete the proof, let n be sufficiently large so that this is greater than K. ■

2.2 Crossing Numbers and Incidences

The following problem has a long and very painful history. It is truly amazing that it can be solved by such a short and elegant proof.

And *embedding* of a graph $G = (V,E)$ in the plane is a planar representation of it, where each vertex is represented by a point in the plane, and each edge uv is represented by a curve connecting the points corresponding to the vertices u and v. The *crossing number* of such an embedding is the number of pairs of intersecting curves that correspond to pairs of edges with no common endpoints. The *crossing number* $\mathrm{cr}(G)$ of G is the minimum possible crossing number in an embedding of it in the plane.

Theorem 2.3 *The crossing number of any simple graph $G = (V,E)$ with $|E| \ge 4|V|$ is at least $\frac{|E|^3}{64|V|^2}$.*

Proof: By Euler's formula any simple planar graph with n vertices has at most $3n-6$ edges. (Indeed, $f - e + v = 2$ in the case with maximum number of edges, we have that every face, has 3 edges around it. Namely, $3f = 2e$. Thus, $(2/3)e - e + v = 2$ in this case. Namely, $e = 3v - 6$.) This implies that the crossing number of any simple graph with n

vertices and m edges is at least $m - 3n + 6 > m - 3n$. Let $G = (V, E)$ be a graph with $|E| \geq 4|V|$ embedded in the plane with $t = \mathrm{cr}(G)$ crossings. Let H be the random induced subgraph of G obtained by picking each vertex of G randomly and independently, to be a vertex of H with probabilistic p (where P will be specified shortly). The expected number of vertex of H is $p|V|$, the expected number of its edges is $p^2|E|$, and the expected number of crossings in the given embedding is $p^4 t$, implying that the expected value of its crossing number is at most $p^4 t$. Therefore, we have $p^4 t \geq p^2|E| - 3p|V|$, implying that

$$\mathrm{cr}(G) \geq \frac{|E|}{p^2} - \frac{3|V|}{p^3},$$

let $p = 4|V| / |E| < 1$, and we have $\mathrm{cr}(G) \geq (1/16 - 3/64)|E|^3 / |V|^2 = |E|^3/(64|V|^2)$. ∎

Theorem 2.4 *Let P be a set of n distinct points in the plane, and let L be a set of m distinct lines. Then, the number of incidences between the members of P and), those of L (that is, the number of pairs (p, ℓ) with $p \in P$, $\ell \in L$, and $p \in \ell$) is at most $c(m^{2/3}n^{2/3} + m + n)$, for some absolute constant c.*

Proof: Let I denote the incidences number. Let $G = (V, E)$ be the graph whose vertices are all the points of P, where two are adjacent if and only if they are consecutive points of P on some line in L. Clearly $|V| = n$, and $|E| = I - m$. Note that G is already given embedded in the plane, where the edges are presented by segments of the corresponding lines of L.

Either, we can not apply Theorem 2.3, implying that $I - m = |E| < 4|V| = 4n$. Namely, $I \leq m + 4n$. Or alliteratively,

$$\frac{(I - m)^3}{(64n^2)} \leq \mathrm{cr}(G) \leq \binom{m}{2} \leq \frac{m^2}{2}.$$

Implying that $I \leq (32)^{1/3} m^{2/3} n^{2/3} + m$. In both cases, $I \leq 4(m^{2/3}n^{2/3} + m + n)$. ∎

This technique has interesting and surprising results, as the following theorem shows.

Theorem 2.5 *For any three sets A, B and C of s real numbers each,*

$$|A \cdot B + C| = \left|\left\{ ab + c \;\middle|\; a \in A, b \in B, mc \in C \right\}\right| \geq \Omega\left(s^{3/2}\right).$$

Proof: Let $R = A \cdot B + C$, $|R| = r$ and define $P = \left\{ (a, t) \;\middle|\; a \in t \in R \right\}$, and $L = \left\{ y = bx + c \;\middle|\; b \in B, c \in C \right\}$.

Clearly $n = |P| = sr$, and $m = |L| = s^2$. And, a line $y = bx + c$ of L is incident with s points of R, namely with $\left\{ (a, t) \;\middle|\; a \in A, t = ab + c \right\}$. Thus, the overall number of incidences is at at least s^3. By Theorem 2.4, we have

$$s^3 \leq 4(m^{2/3}n^{2/3} + m + n) = 4\left(\left(s^2\right)^{2/3}(sr)^{2/3} + s^2 + sr\right) = 4\left(s^2 r^{2/3} + s^2 + sr\right).$$

For $r < s^3$, we have that $sr \leq s^2 r^{2/3}$. Thus, for $r < s^3$, we have $s^3 \leq 12 s^2 r^{2/3}$, implying that $s^{3/2} \leq 12r$. Namely, $|R| = \Omega(s^{3/2})$, as claimed. ∎

Among other things, the crossing number technique implies a better bounds for k-sets in the plane than what was previously known. The k-set problem had attracted a lot of research, and remains till this day one of the major open problems in discrete geometry.