Chapter 50

The Probabilistic Method IV

By Sariel Har-Peled, April 26, 2022[1]

> 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

50.1. The Method of Conditional Probabilities

In previous lectures, we encountered the following problem.

Problem 50.1.1 (Set Balancing/Discrepancy). Given a binary matrix M of size $n \times n$, find a vector $\mathbf{v} \in \{-1, +1\}^n$, such that $\|\mathsf{M}\mathbf{v}\|_\infty$ is minimized.

Using random assignment and the Chernoff inequality, we showed that there exists $\mathbf{v}$, such that $\|\mathsf{M}\mathbf{v}\|_\infty \leq 4\sqrt{n \ln n}$. 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 ith level we expose the ith coordinate of $\mathbf{v}$. 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 – here randomly assigning the remaining bits can be interpreted as a random walk down the tree to a leaf.

Formally, the algorithm is ***successful*** if ends up with a vector $\mathbf{v}$, such that $\|\mathsf{M}\mathbf{v}\|_\infty \leq 4\sqrt{n \ln n}$.

Let v_l and 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 compute $P(\cdot)$ quickly (and deterministically), then we could derandomize the algorithm.

Let C_m^+ be the bad event that $\mathbf{r}_m \cdot \mathbf{v} > 4\sqrt{n \log n}$, where $\mathbf{r}_m$ is the mth row of M. Similarly, C_m^- is the bad event that $\mathbf{r}_m \cdot \mathbf{v} < -4\sqrt{n \log n}$, and let $C_m = C_m^+ \cup C_m^-$. Consider the probability, $\mathbb{P}\big[C_m^+ \,\big|\, \mathbf{v}_1, \ldots, \mathbf{v}_k\big]$ (namely, the first k coordinates of $\mathbf{v}$ are specified). Let $\mathbf{r}_m = (r_1, \ldots, r_n)$. We have that

$$\mathbb{P}\Big[C_m^+ \,\Big|\, \mathbf{v}_1, \ldots, \mathbf{v}_k\Big] = \mathbb{P}\left[\sum_{i=k+1}^{n} \mathbf{v}_i r_i > 4\sqrt{n \log n} - \sum_{i=1}^{k} \mathbf{v}_i r_i\right] = \mathbb{P}\left[\sum_{i \geq k+1, r_i \neq 0} \mathbf{v}_i r_i > L\right] = \mathbb{P}\left[\sum_{i \geq k+1, r_i = 1} \mathbf{v}_i > L\right],$$

where $L = 4\sqrt{n \log n} - \sum_{i=1}^{k} \mathbf{v}_i r_i$ is a known quantity (since $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are known). Let $V = \sum_{i \geq k+1, r_i = 1} 1$. We have,

$$\mathbb{P}\Big[C_m^+ \,\Big|\, \mathbf{v}_1, \ldots, \mathbf{v}_k\Big] = \mathbb{P}\left[\sum_{\substack{i \geq k+1 \\ \alpha_i = 1}} (\mathbf{v}_i + 1) > L + V\right] = \mathbb{P}\left[\sum_{\substack{i \geq k+1 \\ \alpha_i = 1}} \frac{\mathbf{v}_i + 1}{2} > \frac{L + V}{2}\right],$$

The last quantity is the probability that in V flips of a fair 0/1 coin one gets more than $(L+V)/2$ heads. Thus,

$$P_m^+ = \mathbb{P}\left[C_m^+ \,\middle|\, \mathbf{v}_1, \ldots, \mathbf{v}_k\right] = \sum_{i=\lceil (L+V)/2 \rceil}^{\mathsf{V}} \binom{\mathsf{V}}{i} \frac{1}{2^n} = \frac{1}{2^n} \sum_{i=\lceil (L+V)/2 \rceil}^{\mathsf{V}} \binom{\mathsf{V}}{i}.$$

This implies, that we can compute P_m^+ 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_m^-, and let $P_m = P_m^+ + P_m^-$. Clearly, P_m can be computed in polynomial time, by applying a similar argument to the computation of $P_m^- = \mathbb{P}\left[C_m^- \,\middle|\, \mathbf{v}_1, \ldots, \mathbf{v}_k\right]$.

For a node $v \in T$, let $\mathbf{v}_v$ denote the portion of $\mathbf{v}$ that was fixed when traversing from the root of T to v. Let $P(v) = \sum_{m=1}^n \mathbb{P}\left[C_m \,\middle|\, \mathbf{v}_v\right]$. By the above discussion $P(v)$ can be computed in polynomial time. Furthermore, we know, by the previous result on discrepancy that $P(r) < 1$ (that 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 $v = root(T)$. 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 $\mathbf{v}'$ such that $\|A\mathbf{v}'\|_\infty \leq 4\sqrt{n \log n}$.

Theorem 50.1.2. *Using the method of conditional probabilities, one can compute in polynomial time in n, a vector $\mathbf{v} \in \{-1, 1\}^n$, such that $\|A\mathbf{v}\|_\infty \leq 4\sqrt{n \log n}$.*

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

50.2. Independent set in a graph

Theorem 50.2.1. *Consider a graph $\mathsf{G} = (\llbracket n \rrbracket, \mathsf{E})$, with n vertices an m edges. Then G contains an independent set of size*

$$\geq f(n,m) = n/(2m/n+1).$$

In particular, a randomized algorithm can compute an independent set of expected size $\Omega(f(n,m))$.

Proof: Consider a random permutation of the vertices, and in the ith iteration add the vertex π_i to the independent set if none of its neighbors are in the independent set. Let I be the resulting independent set. We have for a vertex $v \in \llbracket n \rrbracket$ that

$$\mathbb{P}[v \in I] \geq \frac{1}{d(i)+1}.$$

As such, the expected size of the computed independent set is

$$\Gamma = \sum_{i=1}^n \mathbb{P}[i \in I] \geq \sum_{i=1}^n \frac{1}{d(i)+1}.$$

Observe that for $x > 0$, and $\alpha \geq x$, we have that

$$1/(1+x) + 1/(1+\alpha-x) = \frac{\alpha}{(1+x)(1+\alpha-x)}.$$

achieves its minimum when $x = \alpha/2$.

As such, $\sum_{i=1}^{n} \frac{1}{d(i)+1}$ is minimized when all the $d(\cdot)$ are equal. Which means that

$$\Gamma \geq \sum_{i=1}^{n} \frac{1}{d(i)+1}. \geq \sum_{i=1}^{n} \frac{1}{(2m/n)+1}. = \frac{n}{(2m/n)+1},$$

as claimed. ■

50.3. A Short Excursion into Combinatorics via the Probabilistic Method

In this section, we provide some additional examples of the Probabilistic Method to prove some results in combinatorics and discrete geometry. While the results are not directly related to our main course, their beauty, hopefully, will speak for itself.

50.3.1. High Girth and High Chromatic Number

Definition 50.3.1. For a graph G, let $\alpha(\mathsf{G})$ be the cardinality of the largest independent set in G, $\chi(\mathsf{G})$ denote the chromatic number of G, and let girth(G) denote the length of the shortest circle in G.

Theorem 50.3.2. *For all K, L there exists a graph G with* girth$(\mathsf{G}) > L$ *and* $\chi(\mathsf{G}) > K$.

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

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

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, $\mathbb{P}[X \geq n/2] = o(1)$.

Let $x = \left\lceil \frac{3}{p} \ln n \right\rceil + 1$. We remind the reader that $\alpha(\mathsf{G})$ denotes the size of the largest independent set in G. We have that

$$\mathbb{P}\Big[\alpha(\mathsf{G}) \geq x\Big] \leq \binom{n}{x}(1-p)^{\binom{x}{2}} < \left(n \exp\left(-\frac{p(x-1)}{2}\right)\right)^x < \left(n \exp\left(-\frac{3}{2} \ln n\right)\right)^x < \Big(o(1)\Big)^x = o(1).$$

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

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 $\alpha(G^*) \leq \alpha(\mathsf{G})$ (any independent set in G^* is also an independent set in G). Thus

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

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

50.3.2. 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 $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ 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*** $\mathrm{cr}(\mathsf{G})$ of G is the minimum possible crossing number in an embedding of it in the plane.

Theorem 50.3.3. *The crossing number of any simple graph* $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ *with* $|\mathsf{E}| \geq 4|\mathsf{V}|$ *is* $\geq \dfrac{|\mathsf{E}|^3}{64|\mathsf{V}|^2}$.

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 $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be a graph with $|\mathsf{E}| \geq 4|\mathsf{V}|$ embedded in the plane with $t = \mathrm{cr}(\mathsf{G})$ 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 vertices of H is $p|\mathsf{V}|$, the expected number of its edges is $p^2|\mathsf{E}|$, 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 $p^4 t \geq p^2|\mathsf{E}| - 3p|\mathsf{V}|$, implying that

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

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

Theorem 50.3.4. *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 points of P and the lines of L (that is, the number of pairs (p, ℓ) with $p \in P$, $\ell \in L$, and $p \in \ell$) is at most $c\left(m^{2/3}n^{2/3} + m + n\right)$, for some absolute constant c.*

Proof: Let I denote the number of such incidences. Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ 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 $|\mathsf{V}| = n$, and $|\mathsf{E}| = I - m$. 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 50.3.3, implying that $I - m = |\mathsf{E}| < 4|\mathsf{V}| = 4n$. Namely, $I \leq m + 4n$. Or alliteratively,

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

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 50.3.5. *For any three sets A, B and C of s real numbers each, we have*

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

Proof: Let $R = A \cdot B + C$, $|R| = r$ and define $P = \left\{(a,t) \;\middle|\; a \in A, t \in R\right\}$, and $L = \left\{y = bx + c \;\middle|\; b \in B, c \in C\right\}$. Clearly $n = |P| = sr$, and $m = |L| = s^2$. Furthermore, 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 least s^3. By Theorem 50.3.4, we have

$$s^3 \le 4\Big(m^{2/3}n^{2/3} + m + n\Big) = 4\Big(\big(s^2\big)^{2/3}(sr)^{2/3} + s^2 + sr\Big) = 4\Big(s^2 r^{2/3} + s^2 + sr\Big).$$

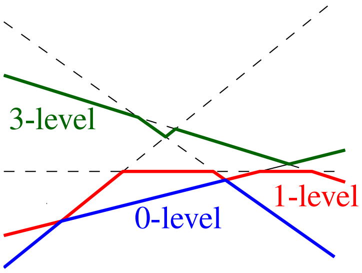
For $r < s^3$, we have that $sr \le s^2 r^{2/3}$. Thus, for $r < s^3$, we have $s^3 \le 12 s^2 r^{2/3}$, implying that $s^{3/2} \le 12r$. Namely, $|R| = \Omega(s^{3/2})$, as claimed. ■

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.

50.3.3. Bounding the at most k-level

Let L be a set of n lines in the plane. Assume, without loss of generality, that no three lines of L pass through a common point, and none of them is vertical. The complement of union of lines L break the plane into regions known as ***faces***. An intersection of two lines, is a ***vertex***, and the maximum interval on a line between two vertices is am ***edge***. The whole structure of vertices, edges and faces induced by L is known as ***arrangement*** of L, denoted by $\mathcal{A}(\mathsf{L})$.

Let L be a set of n lines in the plane. A point $\mathsf{p} \in \bigcup_{\ell \in \mathsf{L}} \ell$ is of ***level*** k if there are k lines of L strictly below it. The ***k-level*** is the closure of the set of points of level k. Namely, the k-level is an x-monotone curve along the lines of L.t



The 0-level is the boundary of the "bottom" face of the arrangement of L (i.e., the face containing the negative y-axis). It is easy to verify that the 0-level has at most $n-1$ vertices, as each line might contribute at most one segment to the 0-level (which is an unbounded convex polygon).

It is natural to ask what the number of vertices at the k-level is (i.e., what the combinatorial complexity of the polygonal chain forming the k-level is). This is a surprisingly hard question, but the same question on the complexity of the at most k-level is considerably easier.

Theorem 50.3.6. *The number of vertices of level at most k in an arrangement of n lines in the plane is $O(nk)$.*

Proof: Pick a random sample R of L, by picking each line to be in the sample with probability $1/k$. Observe that

$$\mathbb{E}[|\mathsf{R}|] = \frac{n}{k}.$$

Let $\mathbb{L}_{\le k} = \mathbb{L}_{\le k}(\mathsf{L})$ be the set of all vertices of $\mathcal{A}(\mathsf{L})$ of level at most k, for $k > 1$. For a vertex $\mathsf{p} \in \mathbb{L}_{\le k}$, let X_{p} be an indicator variable which is 1 if p is a vertex of the 0-level of $\mathcal{A}(\mathsf{R})$. The probability that p is in the 0-level of $\mathcal{A}(\mathsf{R})$ is the probability that none of the j lines below it are picked to be in the sample, and the two lines that define it do get selected to be in the sample. Namely,

$$\mathbb{P}\big[X_{\mathsf{p}} = 1\big] = \left(1 - \frac{1}{k}\right)^j \left(\frac{1}{k}\right)^2 \ge \left(1 - \frac{1}{k}\right)^k \frac{1}{k^2} \ge \exp\left(-2\frac{k}{k}\right)\frac{1}{k^2} = \frac{1}{e^2 k^2}$$

since $j \leq k$ and $1 - x \geq e^{-2x}$, for $0 < x \leq 1/2$.

On the other hand, the number of vertices on the 0-level of R is at most $|\mathsf{R}| - 1$. As such,

$$\sum_{\mathsf{p} \in \mathbb{L}_{\leq k}} X_{\mathsf{p}} \leq |\mathsf{R}| - 1.$$

Moreover this, of course, also holds in expectation, implying

$$\mathbb{E}\left[\sum_{\mathsf{p} \in \mathbb{L}_{\leq k}} X_{\mathsf{p}}\right] \leq \mathbb{E}\Big[|\mathsf{R}| - 1\Big] \leq \frac{n}{k}.$$

On the other hand, by linearity of expectation, we have

$$\mathbb{E}\left[\sum_{\mathsf{p} \in \mathbb{L}_{\leq k}} X_{\mathsf{p}}\right] = \sum_{\mathsf{p} \in \mathbb{L}_{\leq k}} \mathbb{E}\big[X_{\mathsf{p}}\big] \geq \frac{|\mathbb{L}_{\leq k}|}{e^2 k^2}.$$

Putting these two inequalities together, we get that $\dfrac{|\mathbb{L}_{\leq k}|}{e^2 k^2} \leq \dfrac{n}{k}$. Namely, $|\mathbb{L}_{\leq k}| \leq e^2 n k$. ■