Probabilistic Methods in Combinatorics

Po-Shen Loh

25 June 2008

1 Warm-up

Solve the following problems via counting-in-two-ways.

1. (Iran Team Selection Test 2008/6) Suppose 799 teams participate in a tournament in which every pair of teams plays against each other exactly once. Prove that there exist two disjoint groups A and B of 7 teams each such that every team from A defeated every team from B.

   Solution: Sample A as a random 7-set. Let X be the number of guys that are totally dominated by A. Letting d_v^- denote the in-degree of v, we have $\mathbb{E}[X] = \sum_v \binom{d_v^-}{7} / \binom{799}{7}$. But $\sum_v d_v^- = \binom{799}{2}$, which means that the average in-degree is exactly 399. By convexity, $\mathbb{E}[X] \geq 799 \cdot \binom{399}{7} / \binom{799}{7} \approx 800 \cdot (1/2)^7 \approx 6.25$, which is enough since X is an integer. Pick 7 teams B from the dominated group.

2. (Russia 1996/4) In the Duma there are 1600 delegates, who have formed 16000 committees of 80 persons each. Prove that one can find two committees having at least four common members.

   Solution: Simply select a random 2-set of committees. Calculate expected number of people in both committees, and use convexity of $\sum x_i^2 \geq n(\text{avg})^2$.

2 Linearity of expectation

Definition. *Let X be a random variable which takes values in some finite set S. Then the **expected value** of X, denoted $\mathbb{E}[X]$, is:*

$$\mathbb{E}[X] = \sum_{x \in S} x \cdot \mathbb{P}[X = x]$$

Use the following exercises to get used to the concept of expected value.

1. What is the expected number of heads in 101 tosses of a fair coin? Prove this formally from the definition.

   Solution:

$$\sum_{i=0}^{101} i \binom{101}{i} \frac{1}{2^{101}} = 2^{-101} \sum_{i=1}^{101} i \cdot \frac{101}{i} \binom{100}{i-1} = 2^{-101} \sum_{j=0}^{100} 101 \binom{100}{j} = 2^{-101} \cdot 101 \cdot 2^{100} = 101/2.$$

2. What is the expected number of heads in 1 toss of a coin which lands heads with probability 1/10?

3. Can you calculate the expected number of heads in 101 tosses of a coin which lands heads with probability 1/10?

One of the most useful facts about the expected value is that if X and Y are random variables, ***not necessarily*** independent, then $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$. This is known as *linearity of expectation.*

Now use that fact to solve the following problems.

1. Calculate the expected number of heads in 101 tosses of a coin which lands heads with probability 1/10.

   Solution: This is the sum of 101 random variables, each of which has expectation 1/10.

2.1 That's nice, but how does this solve *combinatorics* problems?

The key lemma is the following apparently trivial result.

Lemma. *Let X be a random variable. Then there is some point in the probability space where $X \geq \mathbb{E}[X]$, and also some point in the probability space where $X \leq \mathbb{E}[X]$.*

This takes the place of the "prove-by-contradiction" approach common in many counting-in-two-ways arguments. But where can probability arise in an apparently deterministic combinatorial problem? Let us illustrate this by reworking one of the Warm-up problems.

Solution. Sample a pair of committees uniformly at random (i.e., randomly pick one of the $\binom{16000}{2}$ possible pairs). Let X be the number of people who are in both chosen committees. Note that $X = X_1 + \cdots + X_{1600}$, where each X_i is the $\{0,1\}$-random variable telling whether the i-th person was in both chosen committees. By linearity of expectation,

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_{1600}].$$

The magic is that each $\mathbb{E}[X_i]$ is easy to calculate! Let n_i be the number of committees that the i-th person belongs to. Then, each $\mathbb{E}[X_i] = \mathbb{P}[i\text{-th person is in both picked committees}] = \binom{n_i}{2}/\binom{16000}{2}$. The only piece of information we know about the $\{n_i\}$ is that their sum $\sum_i n_i = 16000 \cdot 80$, so this suggests that we use convexity to bound $\mathbb{E}[X]$ in terms of the average of $\{n_i\}$, which we denote by $\overline{n} = (16000 \cdot 80)/1600 = 800$:

$$\mathbb{E}[X] \geq 1600 \cdot \binom{\overline{n}}{2} \Big/ \binom{16000}{2} = 1600 \cdot \binom{800}{2} \Big/ \binom{16000}{2} = 1600 \cdot \frac{800 \cdot 799}{16000 \cdot 15999} = 3.995.$$

(One could see that since $799 \approx 800$ and $15999 \approx 16000$, the last fraction is roughly 1/400.) But by the Lemma, we know that **some outcome of the probabilistic sampling produces an $X \geq 3.995$.** Since X is always an integer, that outcome must in fact have $X \geq 4$. In particular, we conclude that some pair of committees has ≥ 4 common members. $\square$

Remarks. Observe that once we decided to select the pair of committees at random, we could essentially "follow our nose" to finish the rest of the argument. In particular, we were led to apply convexity because that was the only way to proceed, given our information. Even more conveniently, the sharpening from 3.995 to 4 was automatic!

3 Problems

1. (MOP Test 2007/7/1) In an $n \times n$ array, each of the numbers $1, 2, \ldots, n$ appears exactly n times. Show that there is a row or a column in the array with at least $\sqrt{n}$ distinct numbers.

   Solution: Choose a random row or column ($2n$ choices). Let X be the number of distinct entries in it. Now $X = \sum I_i$, where each I_i is the indicator variable of i appearing (possibly more than once) in our random row or column. Clearly, each $\mathbb{E}[I_i] = \mathbb{P}[I_i \geq 1]$. To lower-bound this, observe that the worst-case is if all n appearances of i are in some $\sqrt{n} \times \sqrt{n}$ submatrix, which gives $\mathbb{P}[I_i \geq 1] \geq 2\sqrt{n}/(2n) = 1/\sqrt{n}$. Hence by linearity, $\mathbb{E}[X] \geq \sqrt{n}$.

2. (IMO Shortlist 1999/C4) Let A be any set of n residues mod n^2. Show that there is a set B of n residues mod n^2 such that at least half of the residues mod n^2 can be written as $a + b$ with $a \in A$ and $b \in B$.

Solution: Make n independent uniformly random choices from the n^2 residues, and collect them into a set B. Note that since we use independence, this final set may have size $< n$. But if we still have $A+B$ occupying at least half of the residues, then this is okay (we could arbitrarily augment B to have the full size n).

Let X be the number of residues achievable as $a+b$. For each potential residue i, there are exactly n ways to choose some b for which $A+b \ni i$, since $|A| = n$. Therefore, the probability that a given residue i appears in $A+B$ is precisely $1 - \left(1 - \frac{n}{n^2}\right)^n$. Then $\mathbb{E}[X]$ is exactly n^2 times that, because there are n^2 total residues. Hence it suffices to show that $1 - \left(1 - \frac{n}{n^2}\right)^n \geq 1/2$. But this follows from the bound $1 - \frac{1}{n} \leq e^{-1/n}$, using $e \approx 2.718$.

3. (Alon-Spencer, Theorem 2.4.1) Let $v_1, \ldots, v_n$ be unit vectors in $\mathbb{R}^d$. Prove that it is possible to assign weights $\epsilon_i \in \{\pm 1\}$ such that the vector $\sum \epsilon_i v_i$ has Euclidean norm† less than or equal to $\sqrt{n}$.

   Solution: Choose weights independently and uniformly at random. Consider the square of the Euclidean norm. By linearity of expectation, this is $\sum_i \|v_i\|^2 + \sum_{i \neq j} 0 v_i \cdot v_j = n$.

4 Harder problems

1. (Austrian-Polish Math Competition 1997/8) Let n be a natural number and M a set with n elements. Find the biggest integer k such that there exists a k-element family of 3-element subsets of M, no two of which are disjoint.

   Solution: We need to be careful about case when n is small: for $n \leq 5$, we can simply take everything. Otherwise, it is Erdős-Ko-Rado.

2. (USAMO 1995/5) Suppose that in a certain society, each pair of persons can be classified as either amicable or hostile. We shall say that each member of an amicable pair is a friend of the other, and each member of a hostile pair is a foe of the other. Suppose that the society has n persons and q amicable pairs, and that for every set of three persons, at least one pair is hostile. Prove that there is at least one member of the society whose foes include $q(1 - 4q/n^2)$ or fewer amicable pairs.

   Solution: Sample the vertex randomly, and calculate the number of edges in its non-neighborhood. By linearity, this has expecation equal to:

$$\begin{aligned}
& \sum_{\text{edges } e} \frac{1}{n} \cdot \#\{v : \text{neither endpoint of } e \text{ sees } v\} \\
= & \sum_{\text{edges } e} \frac{1}{n} \big[(n-2) - \#\{v : \text{exactly one endpoint of } e \text{ sees } v\}\big] \\
= & \; q \cdot \frac{n-2}{n} - \frac{1}{n} \sum_{\text{edges } e} \#\{v : \text{exactly one endpoint of } e \text{ sees } v\},
\end{aligned}$$

   because there are no triangles. Let F be the number of "forks" $K_{1,2}$, where we are counting these up to automorphism, i.e., F is the number of unordered 3-sets of vertices that span exactly 2 edges. By convexity,

$$F = \sum_v \binom{d_v}{2} \geq n \binom{2q/n}{2}.$$

†The Euclidean norm of a vector is $\sqrt{\text{sum of squares of coordinates}}$.

But because of double-counting for each fork, the above bound for the expectation is precisely

$$\begin{aligned}
& q \cdot \frac{n-2}{n} - \frac{1}{n} \cdot 2F \\
\leq\ & q \cdot \frac{n-2}{n} - \frac{1}{n} \cdot 2 \cdot n \binom{2q/n}{2} \\
=\ & q \cdot \frac{n-2}{n} - \frac{1}{n} \cdot 2n \cdot \frac{1}{2} \cdot \frac{2q}{n} \left(\frac{2q}{n} - 1 \right) \\
=\ & q(1 - 4q/n^2).
\end{aligned}$$

3. (Taiwan 1997/9) For $n \geq k \geq 3$, let $X = \{1, 2, \dots, n\}$ and let F_k be a family of k-element subsets of X such that any two subsets in F_k have at most $k-2$ common elements. Show that there exists a subset M_k of X with at least $\lfloor \log_2 n \rfloor + 1$ elements containing no subset in F_k.

   Solution: First, observe that any $k-1$ elements are contained in at most 1 set in F_k, so $|F_k| \leq \binom{n}{k-1}$. Actually, this overcounts by a factor of k because each k-set has k many $(k-1)$-subsets, so in fact $|F_k| \leq \frac{1}{k}\binom{n}{k-1}$.

   Let $t = \lfloor \log_2 n \rfloor + 1$. If $t < k$ we are already done, so assume $t \geq k$. Sample a random t-set of $[n]$. Expected number of members of F_k that are contained in our random set is $|F_k| \cdot \binom{n-k}{t-k} / \binom{n}{t}$. Using the above bound for $|F_k|$, the final expression simplifies to exactly $\frac{1}{n-k+1}\binom{t}{k}$.

   Of course, $\binom{t}{k} \leq 2^t \leq 2n$, but this is not quite enough; it gives a bound of roughly 2. But actually the bound should be divided by a $\sqrt{n}$ factor, so it will get better as n grows. The remainder is left as an exercise to the reader.

4. (MOP Test 2008/7/2) Suppose that a, b, c are positive real numbers such that for every integer n,

   $$\lfloor an \rfloor + \lfloor bn \rfloor = \lfloor cn \rfloor.$$

   Prove that at least one of a, b, c is an integer. (This is not combinatorial, but the Probabilistic Method is still useful.)

   Solution: Suppose none of a, b, c are integers. Divide both sides by n and take the limit. This gives $a + b = c$, so we also know that the sum of fractional parts:

   $$\{an\} + \{bn\} = \{cn\}. \tag{1}$$

   If x is irrational, then $\{xn\}$ is equidistributed over the interval $[0, 1]$. In particular, if we choose n uniformly among $\{1, \dots, N\}$, then $\mathbb{E}[\{xn\}] \to 1/2$ as $N \to \infty$. On the other hand, if x is rational with reduced form p/q, then $\{xn\}$ has expectation tending to $\frac{q-1}{2q} = \frac{1}{2} - \frac{1}{2q}$. So, it is in the interval $[1/4, 1/2)$. **Conclusion:** for any noninteger x, $\mathbb{E}[\{xn\}] \to t$, with $t \in [1/4, 1/2]$.

   Taking the expectation of equation (1), and taking limits, we immediately see that the only way to have the equality is if $\mathbb{E}[\{an\}]$ and $\mathbb{E}[\{bn\}]$ both tend to $1/4$, and $\mathbb{E}[\{cn\}] \to 1/2$. But the only way to get expecation $1/4$ is when a, b are rational, and the only way to get the full expectation $1/2$ is when c is irrational. Yet our first deduction was that $a + b = c$, so we cannot have two rationals summing to one irrational. Contradiction.

5 Real problems

These are interesting results from research mathematics (as opposed to Olympiad mathematics) that have very elegant probabilistic proofs. They come from the excellent book titled *The Probabilistic Method*, by Noga Alon and Joel Spencer.

1. (Erdős-Ko-Rado Theorem) Let $n \geq 2k$ be positive integers, and let $\mathcal{C}$ be a collection of pairwise-intersecting k-element subsets of $\{1, \ldots, n\}$, i.e., every $A, B \in \mathcal{C}$ has $A \cap B \neq \emptyset$. Prove that $|\mathcal{C}| \leq \binom{n-1}{k-1}$.
Remark: this corresponds to the construction which takes all subsets that contain the element 1, say.

   Solution: Pick a random k-set A from $2^{[n]}$ by first selecting a random permutation $\sigma \in S_n$, and then picking a random index $i \in [n]$. Then define $A = \{\sigma(i), \ldots, \sigma(i+k-1)\}$, with indices after n wrapping around, of course. It suffices to show that $\mathbb{P}[A \in \mathcal{C}] \leq k/n$.

   Let us show that conditioned on any fixed σ, $\mathbb{P}[A \in \mathcal{C}|\sigma] \leq k/n$, which will finish our problem. But this is equivalent to the statement that $\mathcal{C}$ can only contain $\leq k$ intervals (wrapping after n) of the form $\{i, \ldots, i+k-1\}$, which is easy to show.

2. (Sperner's Lemma) Prove that the maximum antichain in $2^{[n]}$ has size $\binom{n}{\lfloor n/2 \rfloor}$. That is, show that if $\mathcal{F}$ is a collection of subsets of $\{1, \ldots, n\}$ such that no two distinct sets $A, B \in \mathcal{F}$ satisfy $A \subset B$, then $|\mathcal{F}| \leq \binom{n}{\lfloor n/2 \rfloor}$.

   Solution: The idea is given as a hint in Alon-Spencer. Take a random permutation σ and let the random variable $X = \#\{i : \sigma(1), \ldots, \sigma(i) \in \mathcal{F}\}$. Consider $\mathbb{E}[X]$.

   By definition of $\mathcal{F}$, X is bounded by 1, and the events $\{\sigma(1), \ldots, \sigma(K)\} \in \mathcal{F}$ are disjoint for distinct K. Let N_k be the number of subsets of size k in $\mathcal{F}$.

   $$\mathbb{E}[X] = \sum_{k=1}^{n} \mathbb{P}[\{\sigma(1), \ldots, \sigma(k)\} \in \mathcal{F}] = \sum_{k=1}^{n} \frac{N_k}{\binom{n}{k}} \geq \sum_{k=1}^{n} \frac{N_k}{\binom{n}{\lfloor n/2 \rfloor}}.$$

   Since $\mathbb{E}[X] \leq 1$, we conclude that $\sum N_k$ is bounded by $\binom{n}{\lfloor n/2 \rfloor}$, and we are done.

3. (Crossing Lemma) No matter how you draw a graph with V vertices and E edges in the plane, there will be $\geq \frac{E^3}{64V^2}$ pairs of crossing edges, as long as $E \geq 4V$.

   Solution: Since planar graphs have $E \leq 3V - 6$, we automatically find that the crossing number is always $\geq E - (3V - 6) > E - 3V$. Now take a drawing with, say, t crossings, and sample vertices randomly with probability p. V goes down to pV, E goes down to p^2E, and cr goes down to p^4t. But this new drawing needs to satisfy the above, so

   $$p^4 t > p^2 E - 3pV.$$

   Substituting $p = 4V/E$, we get the desired result.

4. (Bollobás, 1965) Let $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$ be distinct subsets of $\mathbb{N}$ such that:

   - every $|A_i| = r$ and every $|B_i| = s$,
   - for every i, $A_i \cap B_i = \emptyset$, and
   - for every $i \neq j$, $A_i \cap B_j \neq \emptyset$.

   Prove that $n \leq \binom{r+s}{r}$.

   Solution: (Proof from Alon-Spencer.) Let the universe X be the union of all A_i and B_j. Take random permutation of X. Define event X_i to be that all elements of A_i precede all elements of B_i in permutation. Easy to check that all n of the events X_i are pairwise disjoint, and each $\mathbb{P}[X_i] = \binom{r+s}{r}^{-1}$. But sum of probabilities is ≤ 1, so done.

5. (Lovász, 1970) Let $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$ be distinct subsets of $\mathbb{N}$ such that:

   - every $|A_i| = r$ and every $|B_i| = s$,
   - for every i, $A_i \cap B_i = \emptyset$, and
   - for every $i < j$, $A_i \cap B_j \neq \emptyset$.

Prove that $n \leq \binom{r+s}{r}$. **Remark:** this is much more difficult, and the proof uses linear algebra.

6. (Alon-Spencer, Exercise 1.7) Let $A_1, \ldots, A_n$ and $B_1, \ldots, B_n$ be distinct subsets of $\mathbb{N}$ such that:

   - every $|A_i| = r$ and every $|B_i| = s$,
   - for every i, $A_i \cap B_i = \emptyset$, and
   - for every $i \neq j$, $(A_i \cap B_j) \cup (A_j \cap B_i) \neq \emptyset$.

   Prove that $n \leq \frac{(r+s)^{r+s}}{r^r s^s}$.

   Solution: Let $X := \bigcup_{i=1}^n A_i \cup B_i$ be the base set. Define $p := r/(r+s)$, and consider a "coin" which has one side saying "A" and one side saying "B", with the A-side appearing with probability p. For each element in X, independently flip the coin. This defines a mapping $f : X \to \{A, B\}$. Define the family of events $\{E_i\}_1^n$ by having E_i occur when all elements $x \in A_i$ have $f(x) = A$, and all elements $y \in B_i$ have $f(y) = B$.

   By definition of the family of sets, it is impossible for E_i and E_j to occur simultaneously if $i \neq j$, because in particular, there would exist some element in either $A_i \cap B_j$ or $A_j \cap B_i$, and it could neither be A nor B. So, just as in the previous problem, consider the probability of any E_i occurring. Trivially, $P(E_i) = p^r(1-p)^s$ for any i. Since the events are disjoint, the total probability is $np^r(1-p)^s$. Yet all probabilities are bounded by 1, so just as in the previous problem, $n \leq p^{-r}(1-p)^{-s}$, which turns out to be the desired bound. (In fact, this choice of p is optimal.)

7. (Erdős, 1965) A set S is called *sum-free* if there is no triple of (not necessarily distinct) elements $x, y, z \in S$ satisfying $x + y = z$. Prove that every set A of nonzero integers contains a subset $S \subset A$ of size $|S| > |A|/3$ which is sum-free.

   Solution: Choose a prime p of the form $3k + 2$ such that p is greater than twice the maximum absolute value of any element in A. Let $C = \{k + 1, \ldots, 2k + 1\}$, which is sum-free modulo p. Then pick a uniformly random $x \in \{1, \ldots, p\}$ and let B be the set obtained by multiplying each element of A by x, modulo p. For each element, probability of mapping into C is $|C|/(p-1) > 1/3$. So expected number of elements mapping into C is $> |A|/3$, and we can take S to be those that do.

8. (Alon-Spencer, Theorem 3.2.1) Prove that every n-vertex graph with $\frac{nd}{2}$ edges ($d \geq 1$) has a subset U of pairwise-nonadjacent vertices of size $|U| \geq \frac{n}{2d}$.

   Solution: This uses the method of alterations. Pick a random subset U_0 by taking each element independently with probability p. Then, for every edge that survives to U_0, delete one of the endpoints, and call the result U. This has expected size $\mathbb{E}[U] = np - (nd/2)p^2$, and using the optimal $p = 1/d$, we get $\mathbb{E}[U] = \frac{n}{2d}$.

9. (Alon-Spencer, Exercise 2.9) Suppose that every vertex of a n-vertex bipartite graph is given a personalized list of $> \log_2 n$ possible colors. Prove that it is possible to give each vertex a color from its list such that no two adjacent vertices receive the same color. (This is a statement about the *list-chromatic number* of a bipartite graph.)

   Solution: Let the bipartition of the vertex set be $V_1 \cup V_2$. Let X be the total set of all colors in all lists. For each color, independently flip a fair coin. The idea is that we would be done if for every vertex in V_1, we can choose a color that corresponds to a Head, and for every vertex in V_2, we can choose a color that was a Tail. So, let N count the number of vertices for which this fails. Clearly, for each vertex v, the probability of failure is $2^{-|S(v)|} < 1/n$. Thus, by linearity of expectation, $\mathbb{E}[N] < 1$, and we are guaranteed a circumstance in which $N = 0$.

10. (ChipLiar game, Alon-Spencer Theorem 14.2.1) Paul and Carole play a game on a board with positions labeled $\{0, 1, \ldots, k\}$. Initially, n stones are at position k. Paul and Carole play for r rounds, where each round has the following structure: Paul names a subset S of the stones on the board, and then, Carole either moves all stones in S one position to the left, or moves all stones in S^c one position to

the left. Any stone that is moved leftwards from 0 is discarded. If the number of stones on the board becomes 1 or 0, Carole loses. Prove that if $n \cdot \sum_{i=0}^{k} \binom{r}{i} 2^{-r} > 1$, then Carole has a winning strategy.

Solution: Perfect information game, so one of them has a winning strategy. Suppose it is Paul, and let him play it. Carole will respond by playing randomly. Note that if Paul indeed has winning strategy, then it can even always beat a random strategy. Calculate expected number of stones left on the board after r rounds. By linearity, this is n times the probability that a chip survives for r rounds. But with random play, each chip moves left $\mathrm{Bin}(r, 1/2)$ many times. So, expected number of surviving chips is:

$$n\mathbb{P}\left[\mathrm{Bin}(r, 1/2) \leq k\right],$$

which we are given to be greater than 1. Hence some play sequence will leave Carole with ≥ 2 stones, contradicting existence of Paul's perfect strategy.

6 Really harder problems

1. Prove that the Riemann Hypothesis has probability > 0 of being true.