2 Linearity of Expectations

Linearity of expectations refers to the following basic fact about the expectation: given random variables $X_1, \dots, X_n$ and constants $c_1, \dots, c_n$,

$$\mathbb{E}[c_1 X_1 + \cdots + c_n X_n] = c_1 \mathbb{E}[X_1] + \cdots + c_n \mathbb{E}[X_n].$$

This identity does not require any assumption of independence. On the other hand, generally $\mathbb{E}[XY] \neq \mathbb{E}[X]\mathbb{E}[Y]$ unless X and Y are uncorrelated (independent random variables are always uncorrelated).

Here is a simple application (there are also much more involved solutions via enumeration methods).

Question 2.0.1 (Expected number of fixed points)
What is the average number of fixed points of a uniform random permutation of an n element set?

Solution. Let X_i be the event that element $i \in [n]$ is fixed. Then $\mathbb{E}[X_i] = 1/n$. The expected number of fixed points is

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = 1. \qquad \square$$

2.1 Hamiltonian paths in tournaments

We frequently use the following fact:

with positive probability, $X \geq \mathbb{E}[X]$ (likewise for $X \leq \mathbb{E}[X]$).

A ***tournament*** is a directed complete graph. A ***Hamilton path*** in a directed graph is a directed path that contains every vertex exactly once.

Question 2.1.1 (Number of Hamilton paths in a tournament)
What is the maximum (and minimum) number of Hamilton paths in an n-vertex tournament?

The minimization problem is easier. The transitive tournament (i.e., respecting a fixed linear ordering of vertices) has exactly one Hamilton path. On the other hand, every tournament has at least one Hamilton path (Exercise: prove this! Hint: consider a longest directed path).

The maximization problem is more difficult and interesting. Here we have some asymptotic results.

Theorem 2.1.2 (Tournaments wit many Hamilton paths; Szele 1943)
There is a tournament on n vertices with at least $n!2^{-(n-1)}$ Hamilton paths

Proof. Consider a random tournament where every edge is given a random orientation chosen uniformly and independently. Each of the $n!$ permutations of vertices forms a directed path with probability 2^{-n+1}. So that expected number of Hamilton paths is $n!2^{-n+1}$. Thus, there exists a tournament with at least this many Hamilton paths. □

This was considered the first use of the probabilistic method. Szele conjectured that the maximum number of Hamilton paths in a tournament on n players is $n!/(2-o(1))^n$. This was proved by Alon (1990) using the Minc–Brégman theorem on permanents (we will see this later in Chapter 10 on the entropy method).

2.2 Sum-free subset

A subset A in an abelian group is ***sum-free*** if there do not exist $a, b, c \in A$ with $a + b = c$.

Does every n-element set contain a large sum-free set?

Theorem 2.2.1 (Large sum-free subsets; Erdős 1965)
Every set of n nonzero integers contains a sum-free subset of size $\geq n/3$.

Proof. Let $A \subseteq \mathbb{Z} \setminus \{0\}$ with $|A| = n$. For $\theta \in [0, 1]$, let

$$A_\theta := \{a \in A : \{a\theta\} \in (1/3, 2/3)\}$$

where $\{\cdot\}$ denotes fractional part. Then A_θ is sum-free since $(1/3, 2/3)$ is sum-free in $\mathbb{R}/\mathbb{Z}$.

For θ uniformly chosen at random, $\{a\theta\}$ is also uniformly random in $[0, 1]$, so $\mathbb{P}(a \in A_\theta) = 1/3$. By linearity of expectations, $\mathbb{E}|A_\theta| = n/3$. □

***Remark* 2.2.2** (Additional results). Alon and Kleitman (1990) noted that one can improve the bound to $\geq (n+1)/3$ by noting that $|A_\theta| = 0$ for θ close to zero (say, $|\theta| < (3 \max a \in A\, |a|)^{-1}$), so that $|A_\theta| < n/3$ with positive probability, and hence $|A_\theta| > n/3$ with positive probability. Note that since $|A_\theta|$ is an integer, being $> n/3$ is the same as being $\geq (n+1)/3$.

Bourgain (1997) improved it to $\geq (n+2)/3$ via a difficult Fourier analytic argument. This is currently the best lower bound known.

It remains an open problem to prove $\geq (n + f(n))/3$ for some function $f(n) \to \infty$.

In the other direction, Eberhard, Green, and Manners (2014) showed that there exist n-element sets of integers whose largest sum-free subset has size $\leq (1/3 + o(1))n$.

2.3 Turán's theorem and independent sets

Question 2.3.1 (Turán problem)

What is the maximum number of edges in an n-vertex K_k-free graph?

Taking the complement of a graph changes its independent sets to cliques and vice versa. So the problem is equivalent to one about graphs without large independent sets.

The following result, due to Caro (1979) and Wei (1981), shows that a graph with small degrees much contain large independent sets. The probabilistic method proof shown here is due to Alon and Spencer.

Theorem 2.3.2 (Caro 1979, Wei 1981)

Every graph G contains an independent set of size at least

$$\sum_{v \in V(G)} \frac{1}{d_v + 1},$$

where d_v is the degree of vertex v.

Proof. Consider a random ordering (permutation) of the vertices. Let I be the set of vertices that appear before all of its neighbors. Then I is an independent set.

For each $v \in V$, $\mathbb{P}(v \in I) = \frac{1}{1+d_v}$ (this is the probability that v appears first among $\{v\} \cup N(v)$). Thus $\mathbb{E}|I| = \sum_{v \in V(G)} \frac{1}{d_v+1}$. Thus with positive probability, $|I|$ is at least this expectation. □

***Remark* 2.3.3.** Equality occurs if G is a disjoint union of cliques.

Remark 2.3.4 (Derandomization)**.** Here is an alternative "greedy algorithm" proof of the Caro–Wei inequality. At each step, take a vertex of smallest degree, and remove it and all its neighbors. If each vertex v is assigned weight $1/(d_v + 1)$, then the total weight removed at each step is at most 1. Thus there must be at least $\sum_v 1/(d_v + 1)$ steps.

Some probabilistic proofs, especially those involving linearity of expectations, can be derandomized this way into an efficient deterministic algorithm. However, for many other proofs (such as Ramsey lower bounds from Section 1.1), it is not known how to derandomize the proof.

By taking the complement of the graph, independent sets become cliques, and so we obtain the following corollary.

Corollary 2.3.5
Every n-vertex graph G contains a clique of size at least

$$\sum_{v \in V(G)} \frac{1}{n - d_v}.$$

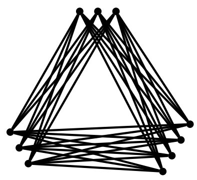
Note that equality is attained when G is multipartite.

Now let us answer the earlier question about maximizing the number of edges in a K_{r+1}-free graph.

The ***Turán graph*** $\boldsymbol{T_{n,r}}$ is the complete multipartite graph formed by partitioning n vertices into r parts with sizes as equal as possible (differing by at most 1).

Example:

$$T_{10,3} = K_{3,3,4} =$$



It is easy to see that $T_{n,r}$ is K_{r+1}-free.

Turán's theorem (1941) tells us that $T_{n,r}$ indeed maximizes the number of edges among n-vertex K_{r+1}-free graphs. We will prove a slightly weaker statement, below, which is tight when n is divisible by r.

Theorem 2.3.6 (Turán's theorem 1941)
The number of edges in an n-vertex K_{r+1}-free graph is at most

$$\left(1 - \frac{1}{r}\right)\frac{n^2}{2}.$$

Proof. Let m be the number of edges. Since G is K_{r+1}-free, by Corollary 2.3.5, the size $\omega(G)$ of the largest clique of G satisfies

$$r \geq \omega(G) \geq \sum_{v \in V} \frac{1}{n - d_v} \geq \frac{n}{n - \frac{1}{n}\sum_v d_v} = \frac{n}{n - \frac{2m}{n}}.$$

Rearranging gives $m \leq \left(1 - \frac{1}{r}\right)\frac{n^2}{2}$. □

Remark 2.3.7. By a careful refinement of the above argument, we can deduce Turán's theorem that $T_{n,r}$ maximizes the number of edges in a n-vertex K_{r+1}-free graph, by noting that $\sum_{v \in V} \frac{1}{n-d_v}$ is minimized over fixed $\sum_v d_v$ when the degrees are nearly equal.

Also, Theorem 2.3.6 is asymptotically tight in the sense that the Turán graph $T_{n,r}$, for fixed r and $n \to \infty$, as $(1 - 1/r - o(1))n^2/2$ edges.

For more on this topic, see Chapter 1 of my textbook *Graph Theory and Additive Combinatorics* and the class with the same title.

2.4 Sampling

By Turán's theorem (actually Mantel's theorem, in this case for triangles, the maximum number of edges in an n-vertex triangle-free graph is $\lfloor n^2/4 \rfloor$.

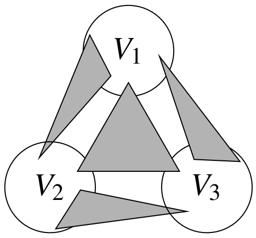
How about the problem for hypergraphs? A ***tetrahedron***, denoted $K_4^{(3)}$, is a complete 3-uniform hypergraph (3-graph) on 4 vertices (think of the faces of a usual 3-dimensional tetrahedron).

Question 2.4.1 (Hypergraph Turán problem for tetrahra)
What is the maximum number of edges in an n-vertex 3-uniform hypergraph not containing any tetrahedra?

This turns out to be a notorious open problem. Turán conjectured that the answer is

$$\left(\frac{5}{9} + o(1)\right)\binom{n}{3},$$

which can be achieved using the 3-graph illustrated below:



Above, the vertices are partitioned into three nearly equal sets V_1, V_2, V_3, and all the edges come in two types: (i) with one vertex in each of the three parts, and (ii) two vertices in V_i and one vertex in V_{i+1}, with the indices considered mod 3.

Let us give some easy upper bounds, in order to illustrate a simple yet important technique of **bounding by sampling**.

Proposition 2.4.2 (A cheap sampling bound)
Every tetrahedron-free 3-graph on $n \geq 4$ vertices has at most $\frac{3}{4}\binom{n}{3}$ edges.

Proof. Let S be a 4-vertex subset chosen uniformly at random. If the graph has $p\binom{n}{3}$ edges, then the expected number of edges induced by S is $4p$ by linearity of expectations (why?).

Since the 3-graph is tetrahedron-free, S induces at most 3 edges. Therefore, $4p \leq 3$. Thus the total number of edges is $p\binom{n}{3} \leq \frac{3}{4}\binom{n}{3}$. □

Why stop at sampling four vertices? Can we do better by sampling five vertices? To run the above argument, we will know how many edges can there be in a 5-vertex tetrahedron-free graph.

Lemma 2.4.3
A 5-vertex tetrahedron-free 3-graph has at most 7 edges.

Proof. We can convert a 5-vertex 3-graph H to a 5-vertex graph G, by replacing each triple by its complement. Then H being tetrahedron-free is equivalent to G not having a vertex of degree 4. The maximum number of edges in a 5-vertex graph with maximum degree at most 3 is $\lfloor 3 \cdot 5/2 \rfloor = 7$ (check this can be achieved). □

We can improve Proposition 2.4.2 by sampling 5 vertices instead of 4 in its proof. This yields (check):

Proposition 2.4.4

Every tetrahedron-free 3-graph on $n \geq 4$ vertices has at most $\frac{7}{10}\binom{n}{3}$ edges.

By sampling s vertices and using brute-force search to solve the s-vertex problem, we can improve the upper bound by taking larger values of s. In fact in principle, if we had unlimited computational power, we can arbitrarily close to optimum by taking sufficiently large s (why?). However, this is not a practical method due to the cost of the brute-force search. There are more clever ways to get better bounds (also with the help of a computer). The best known upper bound notably via a method known as *flag algebras* (using sums of squares) due to Razborov, which can give $\leq (0.561\cdots)\binom{n}{3}$).

For more on the Hypergraph Turán problem, see the survey by Keevash (2011).

2.5 Unbalancing lights

Consider an $n \times n$ array of light bulbs. Initially some arbitrary subset of the light bulbs are turned on. We are allowed up toggle the lights (on/off) for an entire row or column at a time. How many lights can be guarantee to turn on?

If we flip each row/column independently with probability $1/2$, then on expectation, we get exactly half of the lights to turn on. Can we do better?

In the probabilistic method, not every step has to be random. A better strategy is to first flip all the columns randomly, and then decide what to do with each row greedily based on what has happened so far. This is captured in the following theorem, where the left-hand side represents

$$\#\{\text{bulbs on}\} - \#\{\text{bulbs off}\}.$$

Theorem 2.5.1

Let $a_{ij} \in \{-1, 1\}$ for all $i, j \in [n]$. There exists $x_i, y_j \in \{-1, 1\}$ for all $i, j \in [n]$ such that

$$\sum_{i,j=1}^{n} a_{ij} x_i y_j \geq \left(\sqrt{\frac{2}{\pi}} + o(1)\right) n^{3/2}.$$

Proof. Choose $y_1, \ldots, y_n \in \{-1, 1\}$ independently and uniformly at random. For each i, let

$$R_i = \sum_{j=1}^{n} a_{ij} y_j$$

and set $x_i \in \{-1, 1\}$ to be the sign of R_i (arbitrarily choose x_i if $R_i = 0$. Then the LHS sum is

$$\sum_{i=1}^{n} R_i x_i = \sum_{i=1}^{n} |R_i|\,.$$

For each i, R_i has the same distribution as a sum of n i.i.d. uniform $\{-1, 1\}$: $S_n = \varepsilon_1 + \cdots + \varepsilon_n$ (note that R_i's are not independent for different i's). Thus, for each i

$$\mathbb{E}[|R_i|] = \mathbb{E}[|S_n|] = \left(\sqrt{\frac{2}{\pi}} + o(1)\right)\sqrt{n},$$

since by the central limit theorem

$$\lim_{n\to\infty} \mathbb{E}\left[\frac{|S_n|}{\sqrt{n}}\right] = \mathbb{E}[|X|] \qquad \text{where } X \sim \text{Normal}(0, 1)$$
$$= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} |x| e^{-x^2/2}\, dx = \sqrt{\frac{2}{\pi}}$$

(one can also use binomial sum identities to compute exactly: $\mathbb{E}[|S_n|] = n2^{1-n}\binom{n-1}{\lfloor (n-1)/2 \rfloor}$, though it is rather unnecessary to do so.) Thus

$$\mathbb{E}\sum_{i=1}^{n} |R_i| = \left(\sqrt{\frac{2}{\pi}} + o(1)\right) n^{3/2}.$$

Thus with positive probability, the sum is $\geq \left(\sqrt{\frac{2}{\pi}} + o(1)\right) n^{3/2}$. ◻

The next example is tricky. The proof will set up a probabilistic process where the parameters are not given explicitly. A compactness argument will show that a good choice of parameters exists.

Theorem 2.5.2

Let $k \geq 2$. Let $V = V_1 \cup \cdots \cup V_k$, where $V_1, \ldots, V_k$ are disjoint sets of size n. The edges of the complete k-uniform hypergraph on V are colored with red/blue. Suppose that every edge formed by taking one vertex from each $V_1, \ldots, V_k$ is colored blue. Then there exists $S \subseteq V$ such that the number of red edges and blue edges in S differ by more than $c_k n^k$, where $c_k > 0$ is a constant.

Proof. We will write this proof for $k = 3$ for notational simplicity. The same proof works for any k.

Let p_1, p_2, p_3 be real numbers to be decided. We are going to pick S randomly by including each vertex in V_i with probability p_i, independently. Let

$$a_{i,j,k} = \#\{\text{blue edges in } V_i \times V_j \times V_k\} - \#\{\text{red edges in } V_i \times V_j \times V_k\}.$$

Then

$$\mathbb{E}[\#\{\text{blue edges in } S\} - \#\{\text{red edges in } S\}]$$

equals to some polynomial

$$f(p_1, p_2, p_3) = \sum_{i \le j \le k} a_{i,j,k} p_i p_j p_k = n^3 p_1 p_2 p_3 + a_{1,1,1} p_1^3 + a_{1,1,2} p_1^2 p_2 + \cdots .$$

(note that $a_{1,2,3} = n^3$ by hypothesis). We would be done if we can find $p_1, p_2, p_3 \in [0, 1]$ such that $|f(p_1, p_2, p_3)| > c$ for some constant $c > 0$ (not depending on the $a_{i,j,k}$'s). Note that $|a_{i,j,k}| \le n^3$. We are done after the following lemma

Lemma 2.5.3

Let P_k denote the set of polynomials $g(p_1, \ldots, p_k)$ of degree k, whose coefficients have absolute value ≤ 1, and the coefficient of $p_1 p_2 \cdots p_k$ is 1. Then there is a constant $c_k > 0$ such that for all $g \in P_k$, there is some $p_1, \ldots, p_k \in [0, 1]$ with $|g(p_1, \ldots, p_k)| \ge c$.

Proof of Lemma. Set $M(g) = \sup_{p_1,\ldots,p_k \in [0,1]} |g(p_1, \ldots, p_k)|$ (note that sup is achieved as max due to compactness). For $g \in P_k$, since g is nonzero (its coefficient of $p_1 p_2 \cdots p_k$ is 1), we have $M(g) > 0$. As P_k is compact and $M : P_k \to \mathbb{R}$ is continuous, M attains a minimum value $c = M(g) > 0$ for some $g \in P_k$. ■ □

2.6 Crossing number inequality

Consider drawings of graphs on a plane using continuous curves as edges.

The ***crossing number*** **cr(*G*)** is the minimum number of crossings in a drawing of G.

A graph is ***planar*** if $\mathrm{cr}(G) = 0$.

The graphs $K_{3,3}$ and K_5 are non-planar. Furthermore, the following theorem characterizes these two graphs as the only obstructions to planarity:

Kuratowski's theorem (1930). Every non-planar graph contains a subgraph that is topologically homeomorphic to $K_{3,3}$ or K_5.

Wagner's theorem (1937). A graph is planar if and only if it does not have $K_{3,3}$ or K_5 as a minor.

(It is not too hard to show that Wagner's theorem and Kuratowski's theorem are equivalent)

If a graph has a lot of edges, is it guaranteed to have a lot of crossings no matter how it is drawn in the plane?

Question 2.6.1
What is the minimum possible number of crossings that a drawing of:

- K_n? (Hill's conjecture)
- $K_{n,n}$? (Zarankiewicz conjecture; Turán's brick factory problem)
- a graph on n vertices and $n^2/100$ edges?

The following result, due to Ajtai–Chvátal–Newborn–Szemerédi (1982) and Leighton (1984), lower bounds the number of crossings for graphs with many edges.

Theorem 2.6.2 (Crossing number inequality)
In a graph $G = (V, E)$, if $|E| \geq 4|V|$, then

$$\operatorname{cr}(G) \gtrsim \frac{|E|^3}{|V|^2}.$$

***Remark* 2.6.3.** The constant 4 in $|E| \geq 4|V|$ can be replaced by any constant greater than 3 (at the cost of changing the constant in the conclusion). On the other hand, by considering a large triangular grid, we get a planar graph with average degree arbitrarily close to 6.

Corollary 2.6.4
In a graph $G = (V, E)$, if $|E| \gtrsim |V|^2$, then $\operatorname{cr}(G) \gtrsim |V|^4$.

Proof. The proof has three steps, starting with some basic facts on planar graphs.

Step 1: From zero to one.

Recall ***Euler's formula:*** $v - e + f = 2$ for every connected planar drawing of graph. Here v is the number of vertices, e the number of edges, and f the number of faces (connected components of the complement of the drawing, including the outer infinite region).

For every connected planar graph with at least one cycle, $3|F| \leq 2|E|$ since every face is adjacent to ≥ 3 edges, whereas every edge is adjacent to exactly 2 faces. Plugging into Euler's formula, $|E| \leq 3|V| - 6$.

Thus $|E| \le 3|V|$ for all planar graphs. Hence $\operatorname{cr}(G) > 0$ whenever $|E| > 3|V|$.

Step 2: From one to many.

The above argument gives us one crossing. Next, we will use it to obtain many crossings.

By deleting one edge for each crossing, we get a planar graph, so $|E| - \operatorname{cr}(G) \le 3|V|$, that is

$$\operatorname{cr}(G) \ge |E| - 3|V|.$$

This is a "cheap bound." For graphs with $|E| = \Theta(n^2)$, this gives $\operatorname{cr}(G) \gtrsim n^2$. This is not a great bound. We next will use the probabilistic method to boost this bound.

Step 3: Bootstrapping.

Let $p \in [0, 1]$ to be decided. Let $G' = (V', E')$ be obtained from G by randomly keeping each vertex with probability p. Then

$$\operatorname{cr}(G') \ge |E'| - 3|V'|.$$

So

$$\mathbb{E} \operatorname{cr}(G') \ge \mathbb{E}|E'| - 3\mathbb{E}|V'|$$

We have $\mathbb{E} \operatorname{cr}(G') \le p^4 \operatorname{cr}(G)$, $\mathbb{E}|E'| = p^2|E|$ and $\mathbb{E}|V'| = p\mathbb{E}|V|$. So

$$p^4 \operatorname{cr}(G) \ge p^2|E| - 3p|V|.$$

Thus

$$\operatorname{cr}(G) \ge p^{-2}|E| - 3p^{-3}|V|.$$

Setting $p = 4|V|/|E| \in [0, 1]$ (here is where we use the hypothesis that $|E| \ge 4|V|$) so that $4p^{-3}|V| = p^{-2}|E|$, we obtain $\operatorname{cr}(G) \gtrsim |E|^3/|V|^2$. □

***Remark* 2.6.5.** The above idea of boosting a cheap bound to a better bound is an important one. We saw a version of this idea in Section 2.4 where we sampled a constant number of vertices to deduce upper bounds on the hypergraph Turán number. In the above crossing number inequality application, we are also applying some preliminary cheap bound to some sampled induced subgraph, though this time the sampled subgraph has super-constant size.

It is tempting to modify the proof by sampling edges instead of vertices, but this does not work.

Exercises

1. Let A be a measurable subset of the unit sphere in $\mathbb{R}^3$ (centered at the origin) containing no pair of orthogonal points.

   a) Prove that A occupies at most $1/3$ of the sphere in terms of surface area.

   b) $\star$ Prove an upper bound smaller than $1/3$ (give your best bound).

2. $\star$ Prove that every set of 10 points in the plane can be covered by a union of disjoint unit disks.

3. Let $\boldsymbol{r} = (r_1, \dots, r_k)$ be a vector of nonzero integers whose sum is nonzero. Prove that there exists a real $c > 0$ (depending on $\boldsymbol{r}$ only) such that the following holds: for every finite set A of nonzero *reals*, there exists a subset $B \subseteq A$ with $|B| \geq c|A|$ such that there do not exist $b_1, \dots, b_k \in B$ with $r_1 b_1 + \cdots + r_k b_k = 0$.

4. Prove that every set A of n nonzero integers contains two disjoint subsets B_1 and B_2, such that both B_1 and B_2 are sum-free, and $|B_1| + |B_2| > 2n/3$.

5. Let G be an n-vertex graph with pn^2 edges, with $n \geq 10$ and $p \geq 10/n$. Prove that G contains a pair of vertex-disjoint and isomorphic subgraphs (not necessarily induced) each with at least cp^2n^2 edges, where $c > 0$ is a constant.

6. $\star$ Prove that for every positive integer r, there exists an integer K such that the following holds. Let S be a set of rk points evenly spaced on a circle. If we partition $S = S_1 \cup \cdots \cup S_r$ so that $|S_i| = k$ for each i, then, provided $k \geq K$, there exist r congruent triangles where the vertices of the i-th triangle lie in S_i, for each $1 \leq i \leq r$.

7. $\star$ Prove that $[n]^d$ cannot be partitioned into fewer than 2^d sets each of the form $A_1 \times \cdots \times A_d$ where $A_i \subsetneq [n]$.

18.226 Probabilistic Methods in Combinatorics
Fall 2022