Notes 6 : First and second moment methods

Math 733-734: Theory of Probability *Lecturer: Sebastien Roch*

References: [Roc, Sections 2.1-2.3].

Recall:

THM 6.1 (Markov's inequality) *Let X be a non-negative random variable. Then, for all $b > 0$,*

$$\mathbb{P}[X \geq b] \leq \frac{\mathbb{E}X}{b}. \tag{1}$$

THM 6.2 (Chebyshev's inequality) *Let X be a random variable with $\mathbb{E}X^2 < +\infty$. Then, for all $\beta > 0$,*

$$\mathbb{P}[|X - \mathbb{E}X| > \beta] \leq \frac{\mathrm{Var}[X]}{\beta^2}. \tag{2}$$

1 First moment method

Recall that the expectation of a random variable has an elementary, yet handy property: *linearity*. If random variables $X_1, \ldots, X_n$ defined on a joint probability space have finite first moments

$$\mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n], \tag{3}$$

without any further assumption. In particular linearity holds whether or not the X_is are independent.

1.1 The probabilistic method

A key technique of probabilistic combinatorics is the so-called *the probabilistic method*. The idea is that one can establish the existence of an object satisfying a certain property—*without having to construct one explicitly*. Instead one argues that a randomly chosen object exhibits the given property with positive probability. The following "obvious" observation, sometimes referred to as the *first moment principle*, plays a key role in this context.

THM 6.3 (First moment principle) *Let X be a random variable with finite expectation. Then, for any $\mu \in \mathbb{R}$,*

$$\mathbb{E}X \leq \mu \implies \mathbb{P}[X \leq \mu] > 0.$$

Proof: We argue by contradiction, assume $\mathbb{E}X \leq \mu$ and $\mathbb{P}[X \leq \mu] = 0$. Write $\{X \leq \mu\} = \bigcap_{n \geq 1} \{X < \mu + 1/n\}$. That implies by monotonicity that, for any $\varepsilon \in (0,1)$, $\mathbb{P}[X < \mu + 1/n] < \varepsilon$ for n large enough. Hence

$$\begin{aligned}
\mu &\geq \mathbb{E}X \\
&= \mathbb{E}[X; X < \mu + 1/n] + \mathbb{E}[X; X \geq \mu + 1/n] \\
&\geq \mu\, \mathbb{P}[X < \mu + 1/n] + (\mu + 1/n)(1 - \mathbb{P}[X < \mu + 1/n]) \\
&> \mu,
\end{aligned}$$

a contradiction. ∎

The power of this principle is easier to appreciate on an example.

EX 6.4 (Balancing vectors) *Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be arbitrary unit vectors in $\mathbb{R}^n$. How small can we make the norm of the combination*

$$x_1 \mathbf{v}_1 + \cdots + x_n \mathbf{v}_n$$

by appropriately choosing $x_1, \ldots, x_n \in \{-1, +1\}$? We claim that it can be as small as $\sqrt{n}$, for any collection of $\mathbf{v}_i$s. At first sight, this may appear to be a complicated geometry problem. But the proof is trivial once one thinks of choosing the x_is at random. *Let $X_1, \ldots, X_n$ be independent random variables uniformly distributed in $\{-1, +1\}$. Then*

$$\begin{aligned}
\mathbb{E}\|X_1 \mathbf{v}_1 + \cdots + X_n \mathbf{v}_n\|^2 &= \mathbb{E}\left[\sum_{i,j} X_i X_j \mathbf{v}_i \cdot \mathbf{v}_j\right] \\
&= \sum_{i,j} \mathbb{E}[X_i X_j \mathbf{v}_i \cdot \mathbf{v}_j] \qquad (4) \\
&= \sum_{i,j} \mathbf{v}_i \cdot \mathbf{v}_j\, \mathbb{E}[X_i X_j] \\
&= \sum_{i} \|\mathbf{v}_i\|^2 \\
&= n,
\end{aligned}$$

where we used the linearity of expectation in (4)*. But note that a discrete random variable $Z = \|X_1 \mathbf{v}_1 + \cdots + X_n \mathbf{v}_n\|^2$ with expectation $\mathbb{E}Z = n$ must take a value $\leq n$ with positive probability by the first moment principle (Theorem 6.3). In other words, there must be a choice of X_is such that $Z \leq n$. That proves the claim.*

1.2 Union bound

Markov's inequality (THM 6.1) can be interpreted as a quantitative version of the first moment principle (THM 6.3). In this context, it is often stated in the following special form.

THM 6.5 (First moment method) *If X is a non-negative, integer-valued random variable, then*

$$\mathbb{P}[X > 0] \leq \mathbb{E}X. \tag{5}$$

Proof: Take $b = 1$ in Markov's inequality (Theorem 6.1). ■

In words THM 6.3 implies that, if a non-negative integer-valued random variable X has expectation smaller than 1, then its value is 0 with positive probability. THM 6.5 adds: if X has "small" expectation, then its value is 0 with "large" probability. This simple fact is typically used in the following manner: one wants to show that a certain "bad event" does not occur with probability approaching 1; the random variable X then counts the number of such "bad events." See the examples below. In that case, X is a sum of indicators and THM 6.5 reduces to the standard *union bound*, also known as *Boole's inequality*.

COR 6.6 *Let $B_m = A_1 \cup \cdots \cup A_m$, where $A_1, \ldots, A_m$ is a collection of events. Then, letting*

$$\mu_m := \sum_i \mathbb{P}[A_i],$$

we have

$$\mathbb{P}[B_m] \leq \mu_m.$$

In particular, if $\mu_m \to 0$ then $\mathbb{P}[B_m] \to 0$.

Proof: This is of course a fundamental property of probability measures. (Or take $X = \sum_i \mathbb{1}_{A_i}$ in THM 6.5.) ■

Applications of THM 6.3 and 6.5 in the probabilistic method are referred to as the *first moment method*. We give another example in the next section.

1.3 Random permutations: longest increasing subsequence

In this section, we bound the expected length of a longest increasing subsequence in a random permutation. Let σ_n be a uniformly random permutation of $[n] := \{1, \ldots, n\}$ and let L_n be the length of a longest increasing subsequence of σ_n.

CLAIM 6.7

$$\mathbb{E}L_n = \Theta(\sqrt{n}).$$

Proof: We first prove that

$$\limsup_{n\to\infty} \frac{\mathbb{E}L_n}{\sqrt{n}} \le e,$$

which implies half of the claim. Bounding the expectation of L_n is not straightforward as it is the expectation of a *maximum*. A natural way to proceed is to find a value ℓ for which $\mathbb{P}[L_n \ge \ell]$ is "small." More formally, we bound the expectation as follows

$$\mathbb{E}L_n \le \ell\,\mathbb{P}[L_n < \ell] + n\,\mathbb{P}[L_n \ge \ell] \le \ell + n\,\mathbb{P}[L_n \ge \ell], \tag{6}$$

for an ℓ chosen below. To bound the probability on the r.h.s., we appeal to the first moment method by letting X_n be the number of increasing subsequences of length ℓ. We also use the indicator trick, i.e., we think of X_n as a sum of indicators over subsequences (not necessarily increasing) of length ℓ. There are $\binom{n}{\ell}$ such subsequences, each of which is increasing with probability $1/\ell!$. Note that these subsequences are not independent. Nevertheless, by the linearity of expectation and the first moment method,

$$\mathbb{P}[L_n \ge \ell] = \mathbb{P}[X_n > 0] \le \mathbb{E}X_n = \frac{1}{\ell!}\binom{n}{\ell} \le \frac{n^\ell}{(\ell!)^2} \le \frac{n^\ell}{e^2[\ell/e]^{2\ell}} \le \left(\frac{e\sqrt{n}}{\ell}\right)^{2\ell},$$

where we used a standard bound on factorials. Note that, in order for this bound to go to 0, we need $\ell > e\sqrt{n}$. The first claim follows by taking $\ell = (1+\delta)e\sqrt{n}$ in (6), for an arbitrarily small $\delta > 0$.

For the other half of the claim, we show that

$$\frac{\mathbb{E}L_n}{\sqrt{n}} \ge 1.$$

This part does not rely on the first moment method (and may be skipped). We seek a lower bound on the expected length of a longest increasing subsequence. The proof uses the following two ideas. First observe that there is a natural symmetry between the lengths of the longest *increasing* and *decreasing* subsequences—they are identically distributed. Moreover if a permutation has a "short" longest increasing subsequence, then intuitively it must have a "long" decreasing subsequence, and vice versa. Combining these two observations gives a lower bound on the expectation of L_n. Formally, let D_n be the length of a longest decreasing subsequence. By symmetry and the arithmetic mean-geometric mean inequality, note that

$$\mathbb{E}L_n = \mathbb{E}\left[\frac{L_n + D_n}{2}\right] \ge \mathbb{E}\sqrt{L_n D_n}.$$

We show that $L_n D_n \geq n$, which proves the claim. We use a clever combinatorial argument. Let $L_n^{(k)}$ be the length of a longest increasing subsequence ending at position k, and similarly for $D_n^{(k)}$. It suffices to show that the pairs $(L_n^{(k)}, D_n^{(k)})$, $1 \leq k \leq n$ are *distinct*. Indeed, noting that $L_n^{(k)} \leq L_n$ and $D_n^{(k)} \leq D_n$, the number of pairs in $[L_n] \times [D_n]$ is at most $L_n D_n$ which must then be at least n. Let $1 \leq j < k \leq n$. If $\sigma_n(k) > \sigma_n(j)$ then we see that $L_n^{(k)} > L_n^{(j)}$ by appending $\sigma_n(k)$ to the subsequence ending at position j achieving $L_n^{(j)}$. The opposite holds for the decreasing case, which implies that $(L_n^{(j)}, D_n^{(j)})$ and $(L_n^{(k)}, D_n^{(k)})$ must be distinct. This combinatorial argument is known as the *Erdös-Szekeres theorem*. That concludes the proof of the second claim. ■

2 Second moment method

The first moment method gives an upper bound on the probability that a non-negative, integer-valued random variable is positive—provided its expectation is small enough. In this section we seek a *lower bound* on that probability. We first note that a large expectation does not suffice in general. Say X_n is n^2 with probability $1/n$, and 0 otherwise. Then $\mathbb{E}X_n = n \to +\infty$, yet $\mathbb{P}[X_n > 0] \to 0$. That is, although the expectation diverges, the probability that X_n is positive can be arbitrarily small.

So we turn to the second moment. Intuitively the basis for the so-called second moment method is that, if the expectation of X_n is large *and* its variance is relatively small, then we can bound the probability that X_n is close to 0. As we will see in applications, the first and second moment methods often work hand-in-hand.

2.1 Paley-Zygmund inequality

As an immediate corollary of Chebyshev's inequality (THM 6.2), we get a first version of the so-called *second moment method*: if the standard deviation of X is less than its expectation, then the probability that X is 0 is bounded away from 1. Formally, let X be a non-negative, integer-valued random variable (not identically zero). Then

$$\mathbb{P}[X > 0] \geq 1 - \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2}. \tag{7}$$

Indeed, by (2),

$$\mathbb{P}[X = 0] \leq \mathbb{P}[|X - \mathbb{E}X| \geq \mathbb{E}X] \leq \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2}.$$

The following tail inequality, a simple application of Cauchy-Schwarz, leads to an improved version of the second moment method.

THM 6.8 (Paley-Zygmund inequality) *Let X be a non-negative random variable. For all $0 < \theta < 1$,*

$$\mathbb{P}[X \geq \theta\, \mathbb{E}X] \geq (1-\theta)^2 \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}. \tag{8}$$

Proof: We have

$$\begin{aligned} \mathbb{E}X &= \mathbb{E}[X \mathbb{1}_{\{X < \theta \mathbb{E}X\}}] + \mathbb{E}[X \mathbb{1}_{\{X \geq \theta \mathbb{E}X\}}] \\ &\leq \theta \mathbb{E}X + \sqrt{\mathbb{E}[X^2]\mathbb{P}[X \geq \theta \mathbb{E}X]}, \end{aligned}$$

where we used Cauchy-Schwarz. Rearranging gives the result. ∎
As an immediate application:

THM 6.9 (Second moment method) *Let X be a non-negative random variable (not identically zero). Then*

$$\mathbb{P}[X > 0] \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}. \tag{9}$$

Proof: Take $\theta \downarrow 0$ in (8). ∎
Since

$$\frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} = 1 - \frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2 + \mathrm{Var}[X]},$$

we see that (9) is stronger than (7). We typically apply the second moment method to a sequence of random variables (X_n). The previous theorem gives a uniform lower bound on the probability that $\{X_n > 0\}$ when $\mathbb{E}[X_n^2] \leq C(\mathbb{E}[X_n])^2$ for some $C > 0$.

Just like the first moment method, the second moment method is often applied to a sum of indicators.

COR 6.10 *Let $B_m = A_1 \cup \cdots \cup A_m$, where $A_1, \ldots, A_m$ is a collection of events. Write $i \sim j$ if $i \neq j$ and A_i and A_j are* not *independent. Then, letting*

$$\mu_m := \sum_i \mathbb{P}[A_i], \qquad \gamma_m := \sum_{i \sim j} \mathbb{P}[A_i \cap A_j],$$

where the second sum is over ordered pairs, we have $\lim_m \mathbb{P}[B_m] > 0$ whenever $\mu_m \to +\infty$ and $\gamma_m \leq C\mu_m^2$ for some $C > 0$. If moreover $\gamma_m = o(\mu_m^2)$ then $\lim_m \mathbb{P}[B_m] = 1$.

Proof: Take $X := \sum_i \mathbb{1}_{A_i}$ in the second moment method (THM 6.9). Note that

$$\mathrm{Var}[X] = \sum_i \mathrm{Var}[\mathbb{1}_{A_i}] + \sum_{i \neq j} \mathrm{Cov}[\mathbb{1}_{A_i}, \mathbb{1}_{A_j}],$$

where

$$\mathrm{Var}[\mathbb{1}_{A_i}] = \mathbb{E}[(\mathbb{1}_{A_i})^2] - (\mathbb{E}[\mathbb{1}_{A_i}])^2 \leq \mathbb{P}[A_i],$$

and, if A_i and A_j are independent,

$$\mathrm{Cov}[\mathbb{1}_{A_i}, \mathbb{1}_{A_j}] = 0,$$

whereas, if $i \sim j$,

$$\mathrm{Cov}[\mathbb{1}_{A_i}, \mathbb{1}_{A_j}] = \mathbb{E}[\mathbb{1}_{A_i}\mathbb{1}_{A_j}] - \mathbb{E}[\mathbb{1}_{A_i}]\mathbb{E}[\mathbb{1}_{A_j}] \leq \mathbb{P}[A_i \cap A_j].$$

Hence

$$\frac{\mathrm{Var}[X]}{(\mathbb{E}X)^2} \leq \frac{\mu_m + \gamma_m}{\mu_m^2} = \frac{1}{\mu_m} + \frac{\gamma_m}{\mu_m^2}.$$

Noting

$$\frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} = \frac{(\mathbb{E}X)^2}{(\mathbb{E}X)^2 + \mathrm{Var}[X]} = \frac{1}{1 + \mathrm{Var}[X]/(\mathbb{E}X)^2},$$

and applying THM 6.9 gives the result. ■

We give an application of the second moment method in the section.

2.2 Erdös-Rényi random graph: small subgraphs

We start with some definitions.

Definitions An *undirected graph* (or *graph* for short) is a pair $G = (V, E)$ where V is the set of *vertices* (or *nodes* or *sites*) and

$$E \subseteq \{\{u, v\} \,:\, u, v \in V\},$$

is the set of *edges* (or *bonds*). A *subgraph* of $G = (V, E)$ is a graph $G' = (V', E')$ with $V' \subseteq V$ and $E' \subseteq E$. A subgraph containing all possible non-loop edges between its vertices is called a *complete subgraph* or *clique*.

We consider here random graphs. The Erdös-Rényi random graph is defined as follows.

DEF 6.11 (Erdös-Rényi graphs) *Let $V = [n]$ and $p \in [0, 1]$. The* Erdös-Rényi graph $G = (V, E)$ *on n vertices with density p is defined as follows: for each pair $x \neq y$ in V, the edge $\{x, y\}$ is in E with probability p independently of all other edges. We write $G \sim \mathbb{G}_{n,p}$ and we denote the corresponding measure by $\mathbb{P}_{n,p}$.*

Threshold phenomena are common in random graphs. Formally, a *threshold function* for a graph property P is a function $r(n)$ such that

$$\lim_n \mathbb{P}_{n,p_n}[G_n \text{ has property } P] = \begin{cases} 0, & \text{if } p_n \ll r(n) \\ 1, & \text{if } p_n \gg r(n), \end{cases}$$

where, under $\mathbb{P}_{n,p_n}$, $G_n \sim \mathbb{G}_{n,p_n}$ is an Erdös-Rényi graph with n vertices and density p_n. In this section, we first illustrate this definition on the clique number.

Cliques Let $\omega(G)$ be the *clique number* of a graph G, i.e., the size of its largest clique.

CLAIM 6.12 *The property $\omega(G) \geq 4$ has threshold function $n^{-2/3}$.*

Proof: Let X_n be the number of 4-cliques in the Erdös-Rényi graph $G_n \sim \mathbb{G}_{n,p_n}$. Then, noting that there are $\binom{4}{2} = 6$ edges in a 4-clique,

$$\mathbb{E}_{n,p_n}[X_n] = \binom{n}{4} p_n^6 = \Theta(n^4 p_n^6),$$

which goes to 0 when $p_n \ll n^{-2/3}$. Hence the first moment method (THM 6.5) gives one direction.

For the other direction, we apply the second moment method for sums of indicators, COR 6.10. For an enumeration $S_1, \ldots, S_m$ of the 4-tuples of vertices in G_n, let $A_1, \ldots, A_m$ be the events that the corresponding 4-cliques are present. By the calculation above we have $\mu_m = \Theta(n^4 p_n^6)$ which goes to $+\infty$ when $p_n \gg n^{-2/3}$. Also $\mu_m^2 = \Theta(n^8 p_n^{12})$ so it suffices to show that $\gamma_m = o(n^8 p_n^{12})$. Note that two 4-cliques with disjoint edge sets (but possibly sharing one vertex) are independent. Suppose S_i and S_j share 3 vertices. Then

$$\mathbb{P}_{n,p_n}[A_i \mid A_j] = p_n^3,$$

as the event A_j implies that all edges between three of the vertices in S_i are present, and there are 3 edges between the remaining vertex and the rest of S_i. Similarly if

$|S_i \cap S_j| = 2$, $\mathbb{P}_{n,p_n}[A_i \mid A_j] = p_n^5$. Putting these together we get

$$\begin{aligned}
\gamma_m &= \sum_{i \sim j} \mathbb{P}_{n,p_n}[A_j]\, \mathbb{P}_{n,p_n}[A_i \mid A_j] \\
&= \binom{n}{4} p_n^6 \left[\binom{4}{3}(n-4)p_n^3 + \binom{4}{2}\binom{n-4}{2} p_n^5 \right] \\
&= O(n^5 p_n^9) + O(n^6 p_n^{11}) \\
&= O\left(\frac{n^8 p_n^{12}}{n^3 p_n^3}\right) + O\left(\frac{n^8 p_n^{12}}{n^2 p_n}\right) \\
&= o(n^8 p_n^{12}) \\
&= o(\mu_m^2),
\end{aligned}$$

where we used that $p_n \gg n^{-2/3}$ (so that for example $n^3 p_n^3 \gg 1$). COR 6.10 gives the result. ■

Roughly speaking, the first and second moments suffice to pinpoint the threshold in this case because the indicators in X_n are "mostly" pairwise independent and, as a result, the sum is concentrated around its mean.

References

[Roc] Sebastien Roch. Modern Discrete Probability: An Essential Toolkit. *Book in preparation*. http://www.math.wisc.edu/ roch/mdp/.