Variance and second moment methods

Definition 1. *The* variance *of a random variable X is defined as*

$$\mathrm{var}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2] = \mathbb{E}(X^2) - \mathbb{E}(X)^2$$

and can be interpreted as a measure of the spread of X. How far do we expect X to be from its expected value?

Exercise 1. *Show that* $\mathrm{var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$. 1

Definition 2. *The* covariance *of two random variables X and Y is defined as*

$$\mathrm{cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y))) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$$

and can be interpreted as a measure of the correlation between X and Y. The covariance is positive if X tends to be greater than its expectation when Y is greater than its expectation, and negative if the opposite is true. The covariance is 0 when there no such relationship.

Exercise 2. *Show that when two random variables X and Y are independent, we have* 1
$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ *(and hence* $\mathrm{cov}(X, Y) = 0$*).*

We used the first moment method to show that if a nonnegative random variable X has an expected value close to 0, this gives us a lower bound on the probability that $X = 0$. However, if X has expected value far away from 0, this does not give us an upper bound on the probability that $X = 0$. This may seem counterintuitive, but let for example X_n take value n^2 with probability $1/n$ and 0 otherwise. Then as $n \to \infty$, we have that $\mathbb{E}(X_n) \to \infty$, while $\mathbb{P}(X = 0) \to 1$. In order to find an upper bound on $\mathbb{P}(X = 0)$, we need X to stay closer to its expectation: we need to consider its variance.

Lemma 3 (Chebyshev's inequality)**.** *For a random variable X, and constant $t > 0$, we have*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathrm{var}(X)}{t^2}.$$

Exercise 3. *Prove Lemma 3.* 1

From here, we derive one version of the second moment method:

Lemma 4. *For a random variable X, we have*

$$\mathbb{P}(X = 0) \leq \frac{\mathrm{var}(X)}{\mathbb{E}(X)^2}.$$

Exercise 4. *Prove Lemma 4.* 1

We saw that $\mathbb{E}(X) \to \infty$ does not necessarily imply $\mathbb{P}(X = 0) \to 0$, as the variance could be large. Lemma 4 tells us that if $\mathrm{var}(X) = o(\mathbb{E}(X)^2)$ then $\mathbb{E}(X) \to \infty$ does imply $\mathbb{P}(X = 0) \to 0$.

Application: coin tosses

We start with a simple application: let X be the number of "tails" in a series of n (fair,independent) coin tosses. Then X has a binomial distribution with parameters n and $p = 1/2$, and, as we know, $\mathbb{E}(X) = n/2$. Consider the probability that X is "far away" from its expectation. For example, we can use Markov's inequality to obtain the following bound:

$$\mathbb{P}(X \geq 3n/4) \leq \frac{n/2}{3n/4} = \frac{2}{3}.$$

Now, let's find the variance of X. In this case, we can write

$$X = \sum_{i=1}^{n} Y_i,$$

where Y_i is the indicator random variable of the event that the ith trial is a success. Normally, we would use the letter I for indicator random variables, but the following result is a bit more general than that. We have

$$\begin{aligned}
\operatorname{var}(X) &= \mathbb{E}(X^2) - \mathbb{E}(X)^2 \\
&= \mathbb{E}\left(\left(\sum_{i=1}^{n} Y_i\right)^2\right) - \mathbb{E}\left(\sum_{i=1}^{n} Y_i\right)^2 \\
&= \mathbb{E}\left(\sum_{i,j\in\{1,\dots,n\}} Y_iY_j\right) - \left(\sum_{i=1}^{n} \mathbb{E}(Y_i)\right)^2 \\
&= \sum_{i,j\in\{1,\dots,n\}} \mathbb{E}(Y_iY_j) - \mathbb{E}(Y_i)\mathbb{E}(Y_j) \\
&= \sum_{i=1}^{n} (\mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2) + \sum_{i\neq j,\ i,j\in\{1,\dots,n\}} (\mathbb{E}(Y_iY_j) - \mathbb{E}(Y_i)\mathbb{E}(Y_j)) \\
&= \sum_{i=1}^{n} \operatorname{var}(Y_i) + \sum_{i\neq j,\ i,j\in\{1,\dots,n\}} \operatorname{cov}(Y_iY_j)
\end{aligned}$$

In this particular case, the random variables Y_i are all independent. We have

$$\operatorname{var}(Y_i) = \mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2 = \frac{1}{2} - \frac{1}{4} = \frac{1}{4}.$$

Therefore

$$\operatorname{var}(X) = \frac{n}{4}.$$

By Chebyshev, this gives us

$$\mathbb{P}(|X - \mathbb{E}(X)| \leq n/4) \leq \frac{4}{n}.$$

We can even say that

$$\mathbb{P}(|X - \mathbb{E}(X)| \leq \sqrt{n}) \leq \frac{1}{4},$$

which is a much stronger bound than the one obtained by Markov, since $\sqrt{n} = o(\mathbb{E}(X))$.

Application: Arithmetic progressions

We select a random subset S from $\{1, 2, \ldots, n\}$, by including each element with probability p, independently of other elements. We are interested in the appearance of k-term arithmetic progressions. Just as with the isolated vertices in $G(n, p)$ last week, we will find a threshold for the appearance of k-term arithmetic progressions.
Let $\Phi(n, k)$ be the number of k-term arithmetic progressions contained in $\{1, 2, \ldots, n\}$.

Exercise 5. *Show that* $\Phi(n, k) = \Theta(n^2)$ *for fixed* k. 1

Given some arithmetic progression, A, $\mathbb{P}(A \subseteq S) = p^k$ (because for some fixed arithmetic progression A, with $|A| = k$, each element in A has probability p of being included). Let Y_i be an indicator such that

$$Y_i = \begin{cases} 1, & \text{if } A_i \subseteq S \\ 0, & \text{otherwise.} \end{cases},$$

for $1 \leq i \leq \Phi(n, k)$. In this case, it does not matter what labeling/ordering we gave to the arithmetic progressions. We let

$$X = \sum_i Y_i$$

be the number of k-term arithmetic progressions in S, and we have

$$\mathbb{E}(X) = \Phi(n, k) \cdot p^k = \Theta(n^2 p^k).$$

If $p \ll n^{-2/k}$ then $\mathbb{E}(X) \to 0$, which by the first moment method implies that $\mathbb{P}(X = 0) \to 1$.
If $p \gg n^{-2/k}$, then $\mathbb{E}(X) \to \infty$. In order to show that $\mathbb{P}(X = 0) \to 0$ by using Lemma 4, we need that $\mathrm{var}(X) \ll \mathbb{E}(X)^2$.
We have

$$\mathrm{var}(X) = \sum_{i=1}^{\Phi(n,k)} \mathrm{var}(Y_i) + \sum_{i \neq j,\ i,j \in \{1, \ldots, \Phi(n,k)\}} \mathrm{cov}(Y_i Y_j)$$

We have

$$\mathrm{var}(Y_i) = p^k(1 - p^k) \leq p^k,$$

which gives

$$\sum_{i=1}^{\Phi(n,k)} \mathrm{var}(Y_i) = O(n^2 p^k).$$

For the covariances, note that when two arithmetic progressions A_i and A_j have no elements in common, we have $\mathrm{cov}(Y_i, Y_j) = 0$.

Exercise 6. *Show that the number of pairs of k-term arithmetc progressions in* $\{1, 2, \ldots, n\}$ *that overlap in exactly one element is* $O(n^3)$. 1

Exercise 7. *Show that If A_i and A_j overlap in exactly one element, we have* $\mathrm{cov}(Y_i, Y_j) = O(p^{2k-1})$. 1

Exercise 8. *Show that the number of pairs of k-term arithmetc progressions in* $\{1, 2, \ldots, n\}$ *that overlap in more than one element is* $O(n^2)$. *Then, show that for such pairs, we have* $\mathrm{cov}(Y_i, Y_j) = O(p^k)$. 1

Exercise 9. *Put the above together to show that* $\mathrm{var}(X) \ll \mathbb{E}(X)^2$. 1

Threshold for appearance of K_3 in $G(n,p)$

Consider the random graph model $G(n,p)$, and let X be the number of triangles in a graph G sampled from $G(n,p)$. We will often write this as $G \sim G(n,p)$. We are interested in finding a threshold p^* such that when $p \ll p^*$, we have $\mathbb{P}(X = 0) \to 1$, and when $p \gg p^*$, we have $\mathbb{P}(X = 0) \to 0$.

Exercise 10. *Find a value for p^* that has this property, and carefully justify each step.* 5

Solutions to Selected Exercises

Exercise 5. *Show that* $\Phi(n,k) = \Theta(n^2)$ *for fixed* k.

We can identify a particular k-AP by two parameters: the first term and the difference between consecutive terms. For every difference b, we have $n-(k-1)b$ possible starting terms a_1. Furthermore, b can take any value between 1 and $\left\lfloor \frac{n-1}{k-1} \right\rfloor$. This gives

$$\Phi(n,k) = \sum_{b=1}^{\left\lfloor \frac{n-1}{k-1} \right\rfloor} n-(k-1)b \leq \sum_{b=1}^{\left\lfloor \frac{n-1}{k-1} \right\rfloor} n \leq \sum_{b=1}^{n} n = n^2.$$

This gives $\Phi(n,k) = O(n^2)$. We also have

$$\Phi(n,k) = \sum_{b=1}^{\left\lfloor \frac{n-1}{k-1} \right\rfloor} n-(k-1)b \geq \sum_{b=1}^{\left\lfloor \frac{1}{2}\cdot\frac{n-1}{k-1} \right\rfloor} n - \left\lfloor \frac{1}{2}\cdot(n-1) \right\rfloor \geq \sum_{b=1}^{\left\lfloor \frac{1}{2}\cdot\frac{n-1}{k-1} \right\rfloor} \frac{n}{2} \geq c(k)n^2,$$

for some constant $c(k)$ (that depends on k). This gives $n^2 = O(\Phi(n,k))$. Therefore,

$$\Phi(n,k) = \Theta(n^2).$$

Exercise 6. *Show that the number of pairs of* k*-term arithmetc progressions in* $\{1,2,\ldots,n\}$ *that overlap in exactly one element is* $O(n^3)$.

Fix one k-AP and call this set S_1. Now we will count the number of possible second k-APs S_2 that overlap with S_1 in exactly one element. We have seen that there are $O(n)$ possible values for the difference between consecutive elements in S_2. For a fixed difference b, we have k choices of element in S_1 and k choices of element in S_2 that form the overlap. Note that this set of choices fixes S_2. Some of these choices may not be valid, since they either cause elements of S_2 to be outside of the set $[1,\ldots,n]$ and/or cause more than one element to be in $S_1 \cap S_2$, but this is fine since we just want an upper bound. This gives us

$$O(n^2) * n * k^2 = O(n^3)$$

possible pairs S_1 and S_2 that overlap in one element.

Exercise 7. *Show that If* A_i *and* A_j *overlap in exactly one element, we have* $\mathrm{cov}(Y_i,Y_j) = O(p^{2k-1})$.

We have

$$\mathrm{cov}(Y_i,Y_j) = \mathbb{E}(Y_iY_j) - \mathbb{E}(Y_i)\mathbb{E}(Y_j) = \mathbb{P}(Y_i{=}1, Y_j{=}1) - \mathbb{P}(Y_i{=}1)\mathbb{P}(Y_j{=}1) = p^{2k-1} - p^{4k} = O(p^{2k-1}),$$

since a pair of k-term arithmetic progressions overlapping in exactly one element have a total of $2k-1$ elements that need to appear (for both progressions to appear).

Exercise 9. *Put the above together to show that* $\mathrm{var}(X) \ll \mathbb{E}(X)^2$.

Let $p = \omega(n)n^{-2/k}$ for some function $\omega(n) \to \infty$. We have

$$\text{var}(Y_i) = \mathbb{E}(Y_i^2) - \mathbb{E}(Y_i)^2 = p^k - p^{2k} = O(p^k)$$

for all i. This gives us

$$\sum_{i=1}^{\Phi(n,k)} \text{var}(Y_i) = O(n^2 p^k) = O(\omega(n)^k).$$

Putting the above exercises together, we obtain

$$\begin{aligned}\sum_{i \neq j,\ i,j \in \{1,\ldots,\Phi(n,k)\}} \text{cov}(Y_i Y_j) &= O(n^3 p^{2k-1}) + O(n^2 p^k) \\ &= O(n^{-1+2/k}\omega(n)^{2k-1}) + O(\omega(n)^k) \\ &= O(\omega(n)^{2k-1}).\end{aligned}$$

$$\text{var}(X) = \sum_{i=1}^{\Phi(n,k)} \text{var}(Y_i) + \sum_{i \neq j,\ i,j \in \{1,\ldots,\Phi(n,k)\}} \text{cov}(Y_i Y_j) = O(\omega(n)^{2k-1}).$$

Since $\mathbb{E}(X)^{2k} = O(\omega(n)^{2k})$, we have $\text{var}(X) \ll \mathbb{E}(X)^2$ as desired.

Exercise 10. *Find a value for p^* that has this property, and carefully justify each step.*

The threshold function $p^* = \frac{1}{n}$ works here, and we will solve the more general problem of the appearance of any constant size subgraph in $G(n,p)$ next week.