# 4 The Second Moment Method

Starting in this section, we shift the focus to that of **concentration**: essentially, can we say that the value of our random variable $X$ is realtively close to the mean?

## 4.1 Refresher on statistics and concentration

We've been discussing expectations of the form $\mathbb{E}[X]$ so far, and let's say that we find $\mathbb{E}[X]$ to be large. Can we generally conclude that $X$ is large or positive with high probability? No, because outliers can increase the mean dramatically.

So let's consider a sum of variables

$$X = X_1 + X_2 + \cdots + X_n, \quad X_i \sim \text{Bernoulli}(p).$$

If the $X_i$s are independent, we know a lot by the central limit theorem: a lot of random variables will converge to a Gaussian or other known distribution in the large limit. But most of the time, we only have that our variables are "mostly independent" or not independent at all. Is there any way for us to still understand the concentration of the sum?

**Definition 4.1**

The **variance** of a random variable $X$ is defined to be

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

We will often let $\mu$ denote the mean of a variable, $\sigma^2$ denote the variance, and define $\sigma$ to be the (positive) **standard deviation** of $X$.

**Proposition 4.2** (Chebyshev's inequality)

Given a random variable $X$ with mean $\mu$ and variance $\sigma^2$, then for all $\lambda$,

$$\Pr(|x - \mu| \geq \lambda\sigma) \leq \frac{1}{\lambda^2}.$$

*Proof.* The left hand side is equivalent to

$$\Pr((x-\mu)^2 \geq \lambda^2\sigma^2)$$

which, by Markov's inequality, is

$$\leq \frac{\mathbb{E}[|x-\mu|^2]}{\lambda^2\sigma^2} = \frac{\sigma^2}{\lambda^2\sigma^2} = \frac{1}{\lambda^2}.$$

□

Why do we care about these results? The central idea is that if our standard deviation $\sigma \ll \mu$, then we have "concentration" of polynomial decay by Chebyshev.

**Corollary 4.3** (of Chebyshev)

The probability that $X$ deviates from its mean by more than $\varepsilon$ times its mean is bounded as

$$\Pr(|X - \mathbb{E}[X]| \geq \varepsilon \mathbb{E}[X]) \leq \frac{\operatorname{var}(X)}{\varepsilon^2 \mathbb{E}[X]^2}.$$

In particular, if $\operatorname{var}(X) = o(\mathbb{E}[X]^2)$, then $X \sim \mathbb{E}[X]$ with high probability.

Usually, variance is easy to calculate. This is because

$$\operatorname{var}(X) = \operatorname{cov}[X, X],$$

where $\operatorname{cov}[X, Y]$ is the **covariance**

$$\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y]) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

Since this expression is bilinear, if $X = X_1 + \cdots + X_n$, we can expand this out as

$$\sum_{i,j} \operatorname{cov}[X_i, X_j] = \sum_i \operatorname{var}(X_i) + 2 \sum_{i<j} \operatorname{cov}[X_i, X_j]$$

Often the second term here is small, because each $X_i$ is independent with many other $X_j$s or there is low covariance between them.

**Example 4.4** (Binomial distribution)

If $X = X_1 + X_2 + \cdots + X_n$, where each $X_i$ is independently distributed via the Bernoulli distribution $\operatorname{Bernoulli}(p)$, the mean is $\mathbb{E}[X] = np$, and $\sigma^2 = np(1-p)$. As long as $np \gg 1$, we have $\sigma \ll \mu$, so $X \sim \mu$ with high probability.

Later on, we'll get much better bounds. (By the way, we want $np \gg 1$ so our distribution doesn't approach a Poisson distribution instead)

**Example 4.5**

Let $X$ be the number of triangles in a random graph $G(n, p)$, where each edge of $K_n$ is formed with probability $p$.

Is this variable concentrated around its mean? It's pretty easy to compute that mean: $X$ is the sum over all triangles

$$X = \sum_{\substack{i,j,k \in [n] \\ \text{distinct}}} X_{ijk}$$

where $X_{ijk}$ is 1 if they form a triangle and 0 otherwise. Each $X_{ijk}$ can be expanded out in terms of the indicator variables for edges:

$$X = \sum_{\substack{i,j,k \in [n] \\ \text{distinct}}} X_{ij} X_{jk} X_{ik}.$$

By linearity of expectation, each term is $p^3$, so $\mathbb{E}[X] = \binom{n}{3} p^3$. The variance is a bit harder, and we're mostly worried about the covariance term: when do those cross-terms come up?

Well, given a pair of triples $T_1, T_2$ of vertices, we can find the covariance for those triangles. If there is at most one vertex of overlap, no edges overlap, so there is no covariance. The others are a bit harder, but we use

$\operatorname{cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$:

$$\operatorname{cov}[X_{T_1}, X_{T_2}] = \begin{cases} 0 & |T_1 \cap T_2| \le 1 \\ p^5 - p^6 & |T_1 \cap T_2| = 2 \\ p^3 - p^6 & T_1 = T_2 \end{cases}$$

So we can now finish the computation:

$$\operatorname{var}(X) = \binom{n}{3}(p^3 - p^6) + \binom{n}{2}(n-2)(n-3)(p^5 - p^6) \lesssim n^3p^3 + n^4p^5,$$

and we have $\sigma \ll \mu$ if and only if $p \gg \frac{1}{n}$. So this means that the number of triangles is concentrated around its mean with high probability if $p$ is large enough! Later in the course, we will use other methods to prove better concentration.

**Fact 4.6**

It turns out that $X$ satisfies an asymptotic central limit theorem:

$$\frac{X - \mu}{\sigma} \to N(0, 1).$$

This fact was initially proved by taking moments of the form $\mathbb{E}[X^n]$, and the idea is that if the moments agree with the Gaussian moments, we have a Gaussian distribution. But there's a newer method that can be used called the method of projections.

## 4.2 Threshold functions for subgraphs

We're going to try to look for small subgraphs in a large random graph $G(n, p)$. Here's an example:

**Problem 4.7**

For which $p = p_n$ (a sequence in terms of $n$) does $G(n, p)$ have a $K_4$ subgraph with high probability $1 - o(1)$?

**Lemma 4.8**

For any random variable $X$ that takes on nonnegative values,

$$\Pr(X = 0) \le \frac{\operatorname{var}(X)}{\mathbb{E}[X]^2}.$$

*Proof.* The probability that $X = 0$ is at most the probability $|x - \mu| \ge \mu$, which is at most $\frac{\operatorname{var}(x)}{\mu^2}$ by Chebyshev's inequality. □

**Corollary 4.9**

Let $X$ take on only nonnegative values. If the variance of $X$ is much smaller than $\mu^2$, then $X > 0$ with high probability.

**Definition 4.10**

$r(n)$ is a **threshold function** for a property $P$ if $p = p_n \ll r(n)$ means that $G(n, p)$ satisfies $P$ with low probability, while $p = p_n \gg r(n)$ means that $G(n, p)$ satisfies $P$ with high probability.

**Proposition 4.11**

The threshold for a random graph to contain $K_3$ (triangles) is $\frac{1}{n}$, so the probability a graph contains a $K_3$ is 0 if $pn \to 0$ and 1 if $pn \to \infty$.

*Proof.* Let $X$ be the number of triangles in $G(n,p)$. Recall that

$$\mu = \binom{n}{3} p^3 \sim \frac{n^3 p^3}{6}, \sigma^2 = \text{var}(X).$$

If $p \ll \frac{1}{n}$, the mean $\mu = o(1)$, so by Markov's inequality, the probability $X$ has at least one triangle vanishes:

$$\Pr(X \geq 1) \leq \mathbb{E}[X] = o(1).$$

On the other hand, if $p \gg \frac{1}{n}$, $\mu \to \infty$, while $\sigma \ll \mu$. So $X$ is concentrated around its mean with high probability, making it positive with high probability. □

**Problem 4.12**

Given a subgraph $H$, what's the threshold for containing $H$?

Let $X = X_1 + \cdots + X_m$, where each $X_i$ is an indicator variable for $A_i$. We let $i \sim j$ for $i \neq j$ to mean that $A_i$ and $A_j$ are not independent. So if $i \not\sim j$, then $\text{cov}[X_i, X_j] = 0$, but if $i \sim j$,

$$\text{cov}[X_i, X_j] = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] \leq \mathbb{E}[X_i X_j] = \Pr(A_i \cap A_j).$$

So expanding out the expression for variance,

$$\text{var}(X) = \sum_{i,j} \text{cov}[X_i, X_j] \leq \mathbb{E}[X] + \Delta,$$

where $\Delta$ is defined as (the bounded covariance term)

$$\sum_{i<j, i\sim j} \Pr(A_i \cap A_j).$$

So we approximate covariances by probabilities, but if there are very few dependent pairs, we really just care about the number of them. It's possible that all the $X_i$s are all correlated, or that $X_i$s are all nearly independent, but that's not the case here.

**Corollary 4.13**

If $\mathbb{E}[X] \to \infty$ and $\Delta = o(\mathbb{E}[X]^2)$, then with high probability, $X$ is positive and concentrated around its mean.

Simplifying $\Delta$,

$$\Delta = \sum_{i<j, i\sim j} \Pr(A_i \cap A_j) = \sum_i \Pr(A_i) \sum_{j: j\sim i} \Pr(A_j | A_i)$$

and usually the inner sum doesn't depend on $i$ by symmetry. In such cases, we can define

$$\Delta^* = \sum_{j: j\sim i} \Pr(A_j | A_i).$$

We then have

$$\Delta = \sum_i \Pr(A_i)\Delta^* = \Delta^* \cdot \mathbb{E}[X],$$

and this means that if $\mathbb{E}[X] \to \infty$ and $\Delta^* \ll \mu$, $X$ is positive and concentrated around its mean with high probability.

**Proposition 4.14**
The threshold for having $K_4$ as a subgraph is $n^{-2/3}$.

*Proof.* Let $X$ be the random variable which is the number of $K_4$ graphs in $G(n,p)$. The expected value of $X$ is

$$\mathbb{E}[X] = \binom{n}{4} p^6 \sim \frac{n^4 p^6}{24},$$

and if $p \ll n^{-2/3}$, then $\mu = o(1)$, so again by Markov, $X$ is 0 with high probability.

On the other hand, if $p \gg n^{-2/3}$, the mean goes to infinity, and we'll look at the second moment by letting $A_S$ be the event that we induce a $K_4$ on any set $S$ of four vertices. then

$$\Delta^* \lesssim n^2 p^5 + np^3,$$

where $n^2p^5$ comes from sets sharing two vertices (which means we need to find two more and have 5 edges chosen with probability $p$), and $np^3$ comes from sets sharing three vertices (meaning we find one more and have 3 more edges chosen). Provided that $p \gg n^{-2/3}$, both terms here are small: $\Delta^* = o(\mathbb{E}[X])$, and we are done by Corollary 4.13. □

So it seems we should be able to do this with any graph $H$. But the idea with $K_3$ and $K_4$ was that any $p$ with $\mu \to \infty$ gave $X > 0$ with high probability. In general, the answer isn't quite so simple.

**Question 4.15.** *Consider a $K_4$ with an extra edge attached to a vertex as the subgraph that we're looking for. What is its threshold density?*

The expected number of copies of this is $\mathbb{E}[X_H] \asymp n^5 p^7$, so we might predict that the threshold is $p = n^{-5/7}$. Indeed, if $p \ll n^{-5/7}$, $\mathbb{E}[X]$ is very small, and we have zero copies with small probability. But now let's say $p \gg n^{-5/7}$ but $p \ll n^{-2/3}$. There are no $K_4$s, so there's no way we can have this graph at all. Finally, when $p \gg n^{-2/3}$, we have a bunch of $K_4$s: it can be shown that we can easily find another edge to connect to our $K_4$. Therefore, the threshold density is $n^{-2/3}$, and that threshold is not just dependent on the number of edges and vertices of our subgraph $H$!

In a way, this is saying that $K_4$s are the "hard part" of the graph to hit, and the next definition helps us quantify that.

**Definition 4.16**
Define $\rho(H) = \frac{e_H}{v_H}$, sometimes called the **density of $H$**, to be the ratio of edges to vertices in our graph $H$. $H$ is **balanced** if every subgraph $H'$ has $\rho(H') \le \rho(H)$. If $H$ is not balanced, define the **maximum subgraph density** $m(H)$ to be the maximum of $\rho(H')$ across all subgraphs $H'$.

**Example 4.17**
Cliques are balanced: the initial density is $\frac{k-1}{2}$, and we can't do better. On the other hand, the $K_4$ plus an edge is not balanced, since $\rho = \frac{7}{5}$ but the $\rho$ of $K_4$ is $\frac{3}{2}$.

In fact, $m(H)$ is actually what designates the threshold density:

**Theorem 4.18**

If we pick each edge of $K_n$ with probability $p$, the threshold for having $H$ as a subgraph is $p = n^{-\frac{1}{m(H)}}$.

The proof is very similar to what we've been doing.

*Proof.* Let $H'$ be the subgraph with maximum density $\rho(H') = m(H)$. If $p$ is below the threshold, the expected number of copies of $H'$

$$\mathbb{E}[X'_H] \asymp n^{V_{H'}} p^{e_{H'}} = o(1),$$

so with high probability $G(n,p)$ has no copies of $H'$ and therefore no $H$.

Now if $p \gg n^{-1/m(H)}$, we want to compute the number of copies of $H$. For sets $S$ of vertices with $|S| = v_H$,

$$\Delta^* = \sum_{T:|T|=v_H, |T\cap S|\geq 2} \Pr(A_T|A_S)$$

where $T$ is the event that $T$ contains a copy of $H$.

Doing cases based on the size of $T \cap S$ (like we did before), let's say $T$ intersects $S$ in $k$ spots. Here's the key step where we use the maximum subgraph density: overlaps in the covariance terms are subgraphs of $H$. If $H'$ is the overlap between $S$ and $T$, the contribution to $\Delta^*$ is

$$\lesssim n^{v'_H} p^{e'_H} \ll n^{v_H} p^{e_H}$$

for all $H'$, so if we keep track of all the overlaps, we find that $\Delta^* = o(1)$, meaning all overlaps don't contribute much. This finishes the proof by Corollary 4.13. □

## 4.3 Clique number

**Question 4.19.** *What can we say about $\omega(G)$, the number of vertices in the maximum size clique of $G$, if each edge in $K_n$ is included with probability $\frac{1}{2}$?*

We can't quote any of the results from last time, since we're not sticking to fixed-size subgraphs. But this is still not too hard to calculate from first principles.

Let $f(k)$ be the expected number of $k$-cliques: this is just $\binom{n}{k}2^{-\binom{k}{2}}$ by linearity of expectation. We can have a naive guess: perhaps we have a clique whenever this quantity goes to infinity and not when the quantity goes to 0.

**Theorem 4.20**

Let $k = k(n)$ be a function such that $f(k) = \binom{n}{k}2^{-\binom{k}{2}}$ goes to infinity. Then

$$\omega\left(G\left(n, \frac{1}{2}\right)\right) \geq k$$

with high probability.

*Proof.* For all subsets $S$ of the vertices of size $k$, let $A_S$ be the event that $S$ is a clique, and let $\chi_S$ be the indicator variable for $A_S$. Then the number of $k$-cliques

$$X = \sum_S \chi_S$$

has expectation $f(k)$, and we want to show that the variance is much smaller than the mean squared. This is very similar to the earlier proof: fixing $S$, we can find $\Delta^*$ by summing over all $T$ that intersect $S$ in at least two vertices

(those are the only ones that can be dependent on $S$):

$$\Delta^* = \sum_{T:|T\cap S|\geq 2} \Pr(A_T|A_S).$$

We can write this down explicitly, since the expression $\Pr(A_T|A_S)$ just depends on the size of the intersection:

$$= \sum_{i=2}^{k} \left( \binom{k}{i}\binom{n-k}{k-i} \right) 2^{\binom{i}{2}-\binom{k}{2}}$$

where the first term is the number of ways to choose $T$ with an overlap of $i$ vertices, and the power of 2 is the probability that $T$ is a clique given that the $i$ vertices in $S$ are all connected. This does indeed turn out to be small enough: omitting the detailed calculations,

$$\Delta^* \ll \binom{n}{k} 2^{-\binom{n}{k}} = \mathbb{E}[X],$$

so we're done. □

We also know by Markov's inequality that if the expected value goes to 0, the probability of having a $k$-clique is $o(1)$. The idea is that if there's some value $k$ such that $f(k+1) \gg 1$ and $f(k) \ll 1$, then we have a distinctive threshold. But it might be that one of the $f$s is constant order, and then the theorem doesn't actually let us know what happens for that specific value of $k$.

**Theorem 4.21**

There exists a $k_0 = k_0(n)$ such that with high probability,

$$\omega\left(G\left(n, \frac{1}{2}\right)\right) \in \{k_0, k_0+1\}$$

and $k_0 \sim 2\log_2 n$.

This is known as **two-point concentration**. Rephrasing this, if we create this graph at random, we expect one of two values for the clique number.

*Proof sketch.* We can check that for $k \sim 2\log_2 n$,

$$\frac{f(k+1)}{f(k)} = \frac{n-k}{k+1}2^{-k} = n^{-1+o(1)} = o(1).$$

(In particular, the gap between two adjacent $k$s is too large to allow a bunch of $k$s to give constant order $f(k)$s.) Then let $k_0 = k_0(n)$ be the value such that

$$f(k_0) \geq 1 > f(k_0+1);$$

then $f(k_0-1) \gg 1$ and $f(k_0+2) \ll 1$. □

It turns out for most but not all values of $n$, there is only one $k_0$ that $\omega$ takes on with high probability! Later in this class, we'll be able to say something more specific.

## 4.4 Chromatic number

**Question 4.22.** *What is the expected chromatic number (maximum number of colors needed for a proper coloring) in a random graph $G\left(n, \frac{1}{2}\right)$?*

Remember that we have the result $\chi(G)\alpha(G) \geq n$, because each color class is an independent set (and therefore one of them has size at least $\frac{n}{\chi(G)}$).

**Corollary 4.23**

The expected independence number of $G$ is also $\sim 2\log_2 n$, since

$$\alpha(G) = \omega(\overline{G}),$$

since including an edge in $G$ with probability $\frac{1}{2}$ is equivalent to including it in $\overline{G}$ with probability $\frac{1}{2}$.

So this means we can guarantee

$$\chi(G) \geq \frac{n}{\alpha(G)} \sim \frac{n}{2\log_2 n}.$$

Do we also have an upper bound? Can we show that we can color $G\left(n, \frac{1}{2}\right)$ with that many colors?

**Theorem 4.24** (Bollobás, 1987)

The chromatic number

$$\chi\left(G\left(n, \frac{1}{2}\right)\right) \sim \frac{n}{2\log_2 n}.$$

We'll see how to prove this later on using martingale convergence.

## 4.5 Number theory

This class was advertised as using probability to solve problems that don't involved probability. The next few examples have no randomness inherently, but we'll still use the second moment method to solve them.

Let $\nu(n)$ denote the number of prime divisors of $n$, not counting multiplicity. Can we figure out the typical size of $\nu(n)$ just given $n$?

**Theorem 4.25** (Hardy - Ramanujan 1920)

For all $\varepsilon$, there exist a constant $c$ such that all but $\varepsilon$ fraction of the numbers $[1, n]$ satisfy

$$|\nu(x) - \log\log n| \leq c\sqrt{\log\log n}.$$

**Remark.** log *refers to natural log in number theory contexts.*

*Proof by Turán, 1934.* We're going to use a basic intuition about a "random model of the primes." Statistically, they have many properties that make them seem random, even if the primes themselves are not.

Pick a random $x \in [n]$. For each prime $p$, let $X_p$ be the indicator variable

$$X_p = \begin{cases} 1 & p|x \\ 0 & \text{otherwise.} \end{cases}$$

Then the number of prime divisors of $x$ less than or equal to $M$ is approximately

$$X = \sum_{p \leq M} X_p,$$

where we pick $M = n^{1/10}$, a constant power of $n$. Then there are at most 10 prime factors of $x$ larger than $M$, so

$$\nu(x) - 10 \leq X \leq \nu(x).$$

Since we're dealing with asymptotics, that constant is okay for our purposes here. We're treating $X$ as a random variable: we want to show that it is concentrated and that its mean is around $\log\log n$. Each $X_p$ is also a random variable, so this is a good use of the second moment method: we have

$$\mathbb{E}[X_p] = \frac{\lfloor n/p \rfloor}{n} = \frac{1}{p} + O\left(\frac{1}{n}\right)$$

for each prime $p$, so the mean of the random variable is

$$\mathbb{E}[X] = \sum_{p \leq M} \left(\frac{1}{p} + O\left(\frac{1}{n}\right)\right).$$

We'll now use a basic result from analytic number theory:

**Theorem 4.26** (Merten's theorem)
Adding over all primes up to $N$,

$$\sum_{p \leq N} \frac{1}{p} = \log\log N + O(1).$$

To find the expected value of $X^2$, we need to understand the covariance between different $X_p$s. For any primes $p \neq q$,

$$\operatorname{cov}[X_p, X_q] = \mathbb{E}[X_p X_q] - \mathbb{E}[X_p]\mathbb{E}[X_q] = \frac{\lfloor n/(pq) \rfloor}{n} - \frac{\lfloor n/p \rfloor}{n}\frac{\lfloor n/q \rfloor}{n} \leq \frac{1}{pq} - \left(\frac{1}{p} - \frac{1}{n}\right)\left(\frac{1}{q} - \frac{1}{n}\right) \leq \frac{1}{n}\left(\frac{1}{p} + \frac{1}{q}\right).$$

The idea is that these variables are basically independent by Chinese Remainder Theorem, except for the "edge cases" near $n$. So the total sum of the covariances is

$$\sum_{p \neq q, p, q \leq M} \operatorname{cov}[X_p, X_q] \leq \frac{1}{n} \sum_{p \neq q, p, q \leq M} \left(\frac{1}{p} + \frac{1}{q}\right) \leq \frac{2M}{n} \sum_{p \leq M} \frac{1}{p} \lesssim n^{-9/10} \log\log n = o(1),$$

since $M = n^{1/10}$. Now the variance of $X$ is

$$\operatorname{var}(X) = \sum_p \operatorname{var}(X_p) + o(1) = \log\log n + O(1)$$

(which is not very large), and therefore the standard deviation is on the order of $\sqrt{\log\log n}$. Now by Chebyshev's inequality,

$$\Pr\left(|x - \log\log n| \geq \lambda\sqrt{\log\log n}\right) \leq \frac{1}{\lambda^2} + o(1),$$

and since $X$ is within 10 of $\nu(x)$, we've shown concentration with high probability (just pick $\lambda$ to be whatever constant we need in terms of $\varepsilon$). □

What's the distribution, though? Is $\sqrt{\log\log n}$ the right order of magnitude? If we really believe the $X_p$s are independent, we should believe in the central limit theorem.

**Theorem 4.27** (Erdős-Kac theorem)

Picking a random $x \in [n]$, $\nu(x)$ is asymptotically normal:

$$\Pr_{x\in[n]}\left(\frac{\nu(n)-\log\log n}{\sqrt{\log\log n}} \geq \lambda\right) = \frac{1}{\sqrt{2\pi}}\int_{\lambda}^{\infty} e^{-t^2/2}dt$$

for all $\lambda \in \mathbb{R}$.

We briefly mentioned the method of moments earlier: instead of looking at second moments, look at higher moments as well. There's a theorem in probability that if all the moments of our function are the same as certain distributions (including the normal distribution), then convergence happens.

We can do this explicitly if we want, but it gets a bit tedious. Here's a trick that simplifies the calculation: let's compare $\mathbb{E}[X^k]$ with that of an "idealized" random variable $Y$.

*Proof.* This time, set $M = n^{1/s(n)}$ where $s(n) \to \infty$ slowly. Choosing $s(n) = \log\log\log n$ is fine, but $s(n)$ can't grow too quickly because we have that

$$\nu(x) - s(n) \leq X \leq \nu(x).$$

(Joke: What's the sound a drowning number theorist makes?...) So now let

$$Y = \sum_{p\leq M} Y_p,$$

where $Y_p$ is now idealized to Bernoulli$\left(\frac{1}{p}\right)$, independent of the other variables. This is supposed to model $X_p$. So now let

$$\mu = \mathbb{E}[Y] \sim \mathbb{E}[X],$$

and

$$\sigma^2 = \text{var}(Y) \sim \text{var}(X).$$

Set

$$\tilde{X} = \frac{X-\mu}{\sigma}, \tilde{Y} = \frac{Y-\mu}{\sigma}.$$

By the central limit theorem, we know that $\tilde{Y}$ converges to $N(0,1)$. Now let's compare $\tilde{Y}$ and $\tilde{X}$, showing that for all $k$,

$$\mathbb{E}[\tilde{X}^k] = \mathbb{E}[\tilde{Y}^k],$$

which are (by the central limit theorem) also equal to $\mathbb{E}[Z^k]$ for the standard normal distribution.

When we expand out the factors of $\mathbb{E}[X^k - Y^k]$ for distinct primes $p_1, \cdots, p_r \leq M$, they look like

$$\mathbb{E}[X_{p_1}X_{p_2}\cdots X_{p_r} - Y_{p_1}\cdots Y_{p_r}] = \frac{1}{n}\left\lfloor\frac{n}{p_1\cdots p_r}\right\rfloor - \frac{1}{p_1\cdots p_r} = O\left(\frac{1}{n}\right).$$

So if we compare the expansions of $\tilde{X}^k$ in terms of the $X_p$s, there's $M^k = n^{o(1)}$ terms. Since each term contributes $O\left(\frac{1}{n}\right)$, the moments are essentially the same:

$$\mathbb{E}[\tilde{X}^k - \tilde{Y}^k] = n^{-1+o(1)} = o(1).$$

Since all moments converge, $\tilde{X}$ converges to the normal distribution asymptotically. □

## 4.6 Distinct sums

**Question 4.28.** *What's the size of the largest subset $S \subseteq [n]$ such that all $2^{|S|}$ subset sums of $S$ are distinct?*

**Example 4.29**
We can take $S = \{1, 2, 4, \cdots, 2^k\}$, where $k = \lfloor \log_2 n \rfloor$. All sums are distinct by base-2 expansion.

This set has size $\log_2(n)$. Is there any way we can do much better?

**Problem 4.30** (Open; Erdős offered \$300 for this one)
Prove or disprove: $|S| \leq \log_2 n + O(1)$.

One thing we know is that all subset sums have size at most $n|S|$, since there are only $|S|$ things we can add. There are $2^{|S|}$ sums, so if they're all distinct, by Pigeonhole, we must have $2^{|S|} \leq n|S|$, which rearranges to

$$|S| \leq \log_2 n + \log_2 \log_2 n + O(1).$$

Can we formulate a better argument than this? Let's try the second moment method! The idea is that if we pick a random subset sum, we should expect some concentration around the mean.

**Theorem 4.31**
Every subset $S \subseteq [n]$ with distinct subset sums has

$$|S| \leq \log_2 n + \frac{1}{2} \log_2 \log_2 n + O(1).$$

*Proof.* Given our set $S = \{x_1, \cdots, x_k\}$, define a random variable $X = \varepsilon_1 x_1 + \cdots + \varepsilon_k x_k$ where $\varepsilon_i \in \{0, 1\}$ uniformly and independently. The mean is (by linearity of expectation) just $\frac{1}{2}(x_1 + \cdots + x_k)$, and the variance is

$$\sigma^2 = \frac{1}{4}(x_1^2 + \cdots + x_k^2) \leq \frac{n^2 k}{4},$$

since all $x_i \leq n$. By Chebyshev's inequality, for all $\lambda > 1$,

$$\Pr\left[|X - \mu| < \frac{\lambda n \sqrt{k}}{2}\right] \geq 1 - \frac{1}{\lambda^2}.$$

But $X$ must take distinct values for all different instantiations, so the probability that $X = x$ is at most $2^{-k}$ for each $x$. This means that in the probability expression above, $X$ must lie in the range $\left[\mu - \frac{\lambda n \sqrt{k}}{2}, \mu + \frac{\lambda n \sqrt{k}}{2}\right]$, which has a probability

$$\Pr\left[|X - \mu| < \frac{\lambda n \sqrt{k}}{2}\right] \leq 2^{-k} \cdot (\lambda n \sqrt{k} + 1).$$

Putting these inequalities together,

$$1 - \frac{1}{\lambda^2} \leq 2^{-k}(\lambda n \sqrt{k} + 1),$$

which rearranges to

$$n \geq \frac{2^k(1 - \lambda^{-2}) - 1}{\sqrt{k}\lambda}.$$

We can choose $\lambda$ to optimize this expression: in this case, $\lambda = \sqrt{3}$ yields the desired result. □

## 4.7 An application to analysis

We're going to prove the following result using the second moment method:

**Theorem 4.32** (Weierstrass approximation theorem)
Let $f : [0,1] \to \mathbb{R}$ be a continuous function on a bounded interval. Given $\varepsilon > 0$, it is possible to approximate $f$ by a polynomial $p(x)$ such that

$$|p(x) - f(x)| \leq \varepsilon \ \forall x \in [0,1].$$

*Proof.* First of all, since $[0,1]$ is compact, $f$ is uniformly continuous and is therefore bounded. In other words, there exists a $\delta$ such that

$$|f(x) - f(y)| \leq \frac{\varepsilon}{2} \text{ for all } x, y \text{ with } |x - y| \leq \delta.$$

Rescale $f$ so that it is bounded by 1, so now $|f(x)| \leq 1$ for all $x$. Let $X$ be a random variable $X \sim \text{Binomial}(n, x)$: then

$$\Pr(X = j) = \binom{n}{j} x^j (1-x)^{n-j} \text{ for all } 0 \leq j \leq n.$$

We know the statistics $\mathbb{E}[X] = nx, \text{var}(X) = nx(1-x) \leq n$. So by Chebyshev,

$$\Pr\left[|X - nx| > n^{2/3}\right] \leq n^{-1/3}.$$

In particular, if we take $n$ fixed but large enough – let $n > \max(64\varepsilon^{-3}, \delta^{-3})$ – we can bound this in terms of $\varepsilon$:

$$\Pr\left[|X - nx| > n^{2/3}\right] < \frac{\varepsilon}{4}.$$

We can now write down our approximating polynomial explicitly:

$$\boxed{P_n(x) = \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} f\left(\frac{i}{n}\right)}.$$

Basically, chop up $[0,1]$ into $n$ intervals and sample the value at each one. We claim that this works: we do have $|P_n(x) - f(x)| \leq \varepsilon$ for all $x \in [0,1]$. To show this, note that by the triangle inequality,

$$|P_n(x) - f(x)| \leq \sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} \left| f\left(\frac{i}{n}\right) - f(x) \right|$$

implicitly using that the sum of $\binom{n}{i} x^i (1-x)^{n-i} = 1$. The idea is that this absolute value is small if $x \approx \frac{i}{n}$; otherwise, Chebyshev bounds the contribution! We'll split this up into two terms - those close and far away from our given $x$:

$$= \sum_{i : \left|\frac{i}{n} - x\right| \leq n^{-1/3}} \binom{n}{i} x^i (1-x)^{n-i} \left| f\left(\frac{i}{n}\right) - f(x) \right| + 2 \left( \sum_{i : \left|\frac{i}{n} - x\right| > n^{-1/3}} \binom{n}{i} x^i (1-x)^{n-i} \right),$$

where the 2 comes from the fact that $|f(x)| \leq 1$. But now note that the absolute value in the first term deals with those $x$ within $\delta$ of $\frac{i}{n}$, and the second term was bounded earlier:

$$\leq \sum_{i : \left|\frac{i}{n} - x\right| \leq n^{-1/3}} \left( \binom{n}{i} x^i (1-x)^{n-i} \cdot \frac{\varepsilon}{2} \right) + 2 \cdot \frac{\varepsilon}{4},$$

and now both terms are at most $\frac{\varepsilon}{2}$, so this is at most $\varepsilon$, as desired. □

# 18.218 Probabilistic Method in Combinatorics

Spring 2019