20. TWENTIETH LECTURE : 8/12

An important class of results in probability theory are so-called *Central Limit Theorems*. A weaker set of results, called *Laws of Large Numbers*, capture the layman's notion that things tend to average out over time. The Central Limit Theorems are more precise : they tell you that random variables which are long-term averages tend to have *normal distributions*. Recall some definitions from statistics :

Definition. Let X be a real-valued random variable. The *standard deviation* of X, denoted $\sigma(X)$, is the quantity

$$\sigma(X) := \sqrt{\mathbb{E}[(X - \mathbb{E}(X))^2]}. \tag{20.1}$$

Example. Let X be the number of heads obtained when a fair coin is tossed n times. Thus

$$X = \sum_{i=1}^{n} X_i, \tag{20.2}$$

where the X_i are independent, identically distributed (i.i.d.) indicator variables, more precisely :

$$X_i = \begin{cases} 1, & \text{if the } i\text{:th toss yields a head,} \\ 0, & \text{otherwise.} \end{cases} \tag{20.3}$$

Since $\mathbb{E}(X_i) = 1/2$, linearity of expectation gives that $\mathbb{E}(X) = n/2$. Regarding the standard deviation of X, the following formula holds for any sum of indicator variables :

$$\sigma^2(X) = \sum_{i=1}^{n} \sigma^2(X_i) + \sum_{i \neq j} \text{cov}(X_i, X_j). \tag{20.4}$$

The terms in the last sum are called *covariances*. These are zero when the X_i are independent. Since each X_i is either 0 or 1, each with probability $1/2$, it is easy to compute that $\sigma(X_i) = 1/2$. Hence, (20.4) implies that $\sigma(X) = \sqrt{n}/2$.

The standard deviation of a random variable X measures, in some sense, the ‘average’ amount by which X deviates from its mean. Actually, a better way to think about it is that X is unlikely to deviate from its mean by significantly more than a standard deviation. A precise and very general statement of this type is

Theorem 20.1. (Chebyshev's inequality) *Let X be a real-valued random variable with mean μ and standard deviation σ. Then, for any $\lambda \geq 1$,*

$$\mathbb{P}(|X - \mu| > \lambda\sigma) \leq \frac{1}{\lambda^2}. \tag{20.5}$$

Definition. Let $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$. The *normal distribution* with mean μ and standard deviation σ is the real-valued random variable $X = N(\mu, \sigma^2)$ satisfying, for every $z \in \mathbb{R}_+$, that

$$\mathbb{P}(|X - \mu| \geq z) = 2 \cdot \int_z^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-t^2/2\sigma^2}\, dt. \tag{20.6}$$

A normal distribution is highly concentrated within a standard deviation of its mean. For example, (20.6) implies that, for any $\lambda > 0$,

$$\mathbb{P}(|X - \mu| > \lambda\sigma) \leq e^{-\lambda^2/2}. \tag{20.7}$$

This is a much stronger result than the totally general Chebyshev inequality above.

Classically, 'the' Central Limit Theorem is about sums of i.i.d random variables. It is an old result of somewhat obscure origins which says, basically, that if $X_1, X_2, ...$ is a sequence of i.i.d. variables, each with mean μ and standard deviation σ, and $Y_n := (X_1 + \cdots + X_n)/n$ is the average of the first n of them, then Y_n approaches $N(\mu, \sigma^2)$. What do we mean here by 'approaches' ? Well, there are different notions, but the simplest one (which is also the weakest and thus the easiset to get rigorous results about) concerns *convergence in proabability*, which means concretely that, for any $z \in \mathbb{R}_+$,

$$\lim_{n\to\infty} \mathbb{P}(|Y_n - \mu| \geq z) = 2 \cdot \int_z^\infty \frac{1}{\sqrt{2\pi\sigma}} e^{-t^2/2\sigma^2}\, dt. \tag{20.8}$$

While the CLT is a fundamental theoretical result, there are several problems with its application :

(I) it assumes identical distributions
(II) it assumes independence
(III) it is a qualitative, not a quantitative result. In other words, it doesn't say anything about the rate of convergence to the limit in (20.8).

Problem (I) is not serious : the CLT can be extended to sums of variables with different distributions. (II) and (III) are much more serious, though. There are CL Theorems that concern dependent variables, but results are limited. In a seminal paper, Chernoff (1952) dealt significantly with problem (III). His results concern sums of indicator variables. Chernoff was interested in statistics, and his results are of great importance in that field. Sums of indicator variables are also ubiquitous in combinatorial applications, so Chernoff's results deserve the combinatorialist's attention. On the other hand, the fact that they don't address the issue of independence limits their applicability[1].

In the literature, you will find several different results all referred to as *Chernoff's inequality*. The following is one of the more common formulations, and the one which we shall use in our proof of Theorem 18.2 for $h = 2$:

Theorem 20.2. **(Chernoff 1952)** *Let X be a random variable which is a sum of independent indicator variables. Let $\mathbb{E}(x) := \mu$. Then, for any $\epsilon > 0$, there exists a constant $c_\epsilon > 0$, depending only on ϵ, such that*

$$\mathbb{P}(|X - \mu| > \epsilon\mu) < 2e^{-c_\epsilon \mu}. \tag{20.9}$$

[1] Note that, in the example of Ramsey numbers, the variables X_A in (19.19) are not independent. If A and B are two K_k subgraphs which have at least one common edge (i.e.: at least two common vertices), then X_A and X_B are dependent. We are thus really lucky that in that application, we only needed to know $\mathbb{E}(X)$, and nothing at all about the concentration of X.

In fact one can take

$$c_\epsilon = \min\left\{\frac{\epsilon^2}{2}, (1+\epsilon)\log(1+\epsilon) - \epsilon\right\}. \tag{20.10}$$

The crucial point here is that c_ϵ does not depend on X, i.e.: it doesn't depend on how many indicator variables X is the sum of, nor on the distributions of these. It is in this sense that Chernoff's inequality is a quantitative version of CLT. By the way, note that while there is neither any explicit reference to the standard deviation $\sigma(X)$, the fact that c_ϵ depends quadratically on ϵ can be shown to imply that the bound in (20.9) does implicitly depend on $\sigma(X)$.

We will deduce Theorem 20.2 from a *normalised* version of it. Let X_i be an indicator variable, say

$$X_i = \begin{cases} 1, & \text{with probability } p_i, \\ 0, & \text{with probability } 1 - p_i. \end{cases}$$

The *normalisation* of X_i, which we denote $\hat{X}_i$, is the variable $X_i - p_i$, i.e.:

$$\hat{X}_i = \begin{cases} 1 - p_i, & \text{with probability } p_i, \\ -p_i, & \text{with probability } 1 - p_i. \end{cases} \tag{20.11}$$

Thus $\hat{X}_i$ has mean zero. It has the same variance as X_i, namely $p_i(1-p_i)$.

Now let $\hat{X}$ be a random variable which is a sum of n normalised indicator variables, for some fixed n. Write $\hat{X} = \hat{X}_1 + \cdots + \hat{X}_n$, with the $\hat{X}_i$ as above, and define the number p by $np = p_1 + \cdots + p_n$. Finally, let a be any positive real number. We will prove the following two inequalities:

$$\mathbb{P}(\hat{X} > a) < \exp\left[a - pn\ln\left(1 + \frac{a}{pn}\right) - a\ln\left(1 + \frac{a}{pn}\right)\right], \tag{20.12}$$

$$\mathbb{P}(\hat{X} < -a) < \exp\left[-\frac{a^2}{2pn}\right]. \tag{20.13}$$

Note that the theorem follows from (20.12) and (20.13) upon setting $a = \epsilon pn$. We will prove (20.12) in detail. The proof of (20.13) is very similar and is left as an exercise on Homework 3. First, though, a couple of remarks are in order:

(i) there is an obvious asymmetry in the estimates (20.12) and (20.13), depending on whether $\hat{X}$ is positive or negative. Unfortunately, this is a feature of Chernoff's method.
(ii) the connection to the normal distribution is clear in (20.13), as the variance of $\hat{X}$ is about np if the individual p_i are small, as is usually the case in applications. With (20.12), the connection is not so obvious. However, if a is small compared to pn and we use the fact that $\ln(1+u) \geq u - u^2/2$ when $0 < u < 1$, then we can deduce from (20.12) that

$$\mathbb{P}(\hat{X} > a) < \exp\left[-\frac{a^2}{2pn} + \frac{a^3}{2(pn)^2}\right]. \tag{20.14}$$

Note that (20.14) gives no information when a is large compared to np as then the cubic term dominates. Again, this is a feature of Chernoff's method, but is not important, since we're only interested in having concentration close to the mean anyway.

The proof of (20.12) uses the *exponential generating function* of $\hat{X}$, namely: Let $\lambda > 0$. Then we will consider the random variable

$$e^{\lambda \hat{X}} := \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \hat{X}^k.$$

Now $\hat{X} > a$ if and only if $e^{\lambda \hat{X}} > e^{\lambda a}$. The simple Markov inequality gives a bound

$$P(e^{\lambda \hat{X}} > e^{\lambda a}) < \frac{\mathbb{E}[e^{\lambda \hat{X}}]}{e^{\lambda a}}. \tag{20.15}$$

We will estimate the expectation and then the clever part of the proof is that λ, which at this point is still some arbitrary positive real number, will be chosen so as to minimise the right-hand side of (20.15). The estimate of the expectation will use the concavity of the logarithm. Let us begin by formally defining what this means:

Definition. A function f on the positive reals is said to be *concave* if, for any n, any positive reals $x_1 \leq x_2 \leq \cdots \leq x_n$ and any positive reals $a_1, ..., a_n$ satisfying $\sum a_i = 1$, it holds that

$$f\left(\sum_{i=1}^{n} a_i x_i\right) \geq \sum_{i=1}^{n} a_i f(x_i).$$

Concavity has a simple geometric interpretation, namely that the graph of f lies on or above the straight line drawn between any two points on it.

CLAIM: *Let $C > 0$. Then the function $f(x) = \ln(Cx + 1)$ is concave.*

The claim follows immediately from the observation that $f'(x) > 0$ and $f''(x) < 0$ for all $x > 0$.

Now let us return to the proof of (20.12). Since $\hat{X} = \sum \hat{X}_i$, one easily sees that

$$e^{\lambda \hat{X}} = \prod_{i=1}^{n} e^{\lambda \hat{X}_i}.$$

Now we use the independence of the $\hat{X}_i$. Recall that if A, B are independent random variables, then $\mathbb{E}[AB] = \mathbb{E}[A]\mathbb{E}[B]$. Thus, by induction,

$$\mathbb{E}[e^{\lambda \hat{X}}] = \prod_{i=1}^{n} \mathbb{E}[e^{\lambda \hat{X}_i}]. \tag{20.16}$$

But from (20.11), the definition of e.g.f. and linearity of expectation (convergence is not a problem), one easily computes that

$$\mathbb{E}[e^{\lambda \hat{X}_i}] = p_i e^{\lambda(1-p_i)} + (1-p_i)e^{-\lambda p_i} = e^{-\lambda p_i}\left[p_i(e^{\lambda} - 1) + 1\right].$$

Substituting into (20.16) and recalling the definition of p, we thus have

$$\mathbb{E}[e^{\lambda\hat{X}}] = e^{-\lambda pn}\prod_{i=1}^{n}[p_i(e^{\lambda}-1)+1]. \tag{20.17}$$

But

$$\prod_{i=1}^{n}[p_i(e^{\lambda}-1)+1] \leq [p(e^{\lambda}-1)+1]^n. \tag{20.18}$$

Indeed this follows from taking logarithms and using the Claim. So substituting (20.18) back into (20.17) and in turn back into (20.15), we have the estimate

$$\mathbb{P}(\hat{X} > a) < e^{-\lambda pn}[pe^{\lambda}+(1-p)]^n e^{-\lambda a}. \tag{20.19}$$

It is now a horrid calculus exercise to compute the precise value of λ which minimises the right hand side of (20.19)[2]. However, a good approximation when $a \ll np$ is to take $\lambda = \ln(1+a/pn)$. Substituting this into (20.19) we get the desired relation (20.12) upon noticing that, with this choice of λ,

$$[pe^{\lambda}+(1-p)]^n = (1+a/n)^n \leq e^a.$$

This completes the proof of Theorem 20.2.

Before coming to our application to thin bases, we state and prove one final lemma which will be needed :

Lemma 20.3. (Borel-Cantelli lemma) *Let $(\mathscr{E}_n)_1^{\infty}$ be an infinite sequence of events in a probability space. If*

$$\sum_{n=1}^{\infty}\mathbb{P}(\mathscr{E}_n) < \infty \tag{20.20}$$

then, with probability 1, only finitely many of these events occur.

Proof. Let $\mathscr{E}$ be the event that infinitely many of the $\mathscr{E}_n$ occur. We must show that $\mathbb{P}(\mathscr{E}) = 0$. We will show that $\mathbb{P}(\mathscr{E}) < \epsilon$, for every $\epsilon > 0$. So let $\epsilon > 0$ be given. By (20.11), there must exist some $N_\epsilon > 0$ such that

$$\sum_{n=N_\epsilon}^{\infty}\mathbb{P}(\mathscr{E}_n) < \epsilon. \tag{20.21}$$

But if $\mathscr{E}$ occurs, then so must some $\mathscr{E}_n$, for some $n \geq N_\epsilon$. Hence

$$\mathbb{P}(\mathscr{E}) \leq \mathbb{P}\left(\bigcup_{n=N_\epsilon}^{\infty}\mathscr{E}_n\right) \leq \sum_{n=N_\epsilon}^{\infty}\mathbb{P}(\mathscr{E}_n) < \epsilon. \tag{20.22}$$

□

I started, but did not finish, the proof of Theorem 18.2 ($h = 2$), so instead will present it in full in next day's notes ...

[2] The right value turns out to be

$$\lambda = \ln\left[\left(\frac{1-p}{p}\right)\left(\frac{a+np}{n-(a+np)}\right)\right].$$

21. Twenty-First Lecture : 10/12

Proof. of Theorem 18.2 ($h = 2$). Let K be a large positive constant - how large will be decided later on. We choose a subset $A \subseteq \mathbb{N}$ randomly according to the following rule : each $x \in \mathbb{N}$ is considered independently, and chosen to lie in A with probability

$$p_x := \min \left\{ 1, K\sqrt{\frac{\log x}{x}} \right\}. \tag{21.1}$$

I claim that, for a suitably large choice of K, the set A will satisfy, with probability 1, that

$$0.9\left(\frac{K^2\pi}{2}\right)\log n < r_2(A,n) < 1.1\left(\frac{K^2\pi}{2}\right)\log n, \tag{21.2}$$

for all $n \gg 0$. In fact, for each $n \in \mathbb{N}$, let $\mathscr{E}_n$ denote the event that (21.2) is not satisfied. By the Borel-Cantelli lemma, it suffices to prove that

$$\sum_{n=1}^{\infty} \mathbb{P}(\mathscr{E}_n) < \infty. \tag{21.3}$$

Let X_n denote the random variable $r_2(A,n)$. Then

$$X_n = \sum_{t=1}^{\lfloor n/2 \rfloor} X_{n,t}, \tag{21.4}$$

where

$$X_{n,t} = \left\{ \begin{array}{ll} 1, & \text{if both } t \text{ and } n-t \text{ are in } A, \\ 0, & \text{otherwise.} \end{array} \right. \tag{21.5}$$

Each $X_{n,t}$ is an indicator variable and, CRUCIALLY, for each fixed n, the $X_{n,t}$ are independent. Thus each X_n is a sum of independent, indicator variables and we can apply Chernoff's inequality. Let $\mu_n := \mathbb{E}[X_n]$. I claim that

$$\mu_n \sim \frac{K^2\pi}{2}\log n. \tag{21.6}$$

For the moment, let us assume this is true and show how to finish the proof. Take $\epsilon = 0.09$ and apply (20.9). For $n \gg 0$, $\mathscr{E}_n$ will be contained in the event that $|X_n - \mu_n| > \epsilon\mu_n$. Thus, it follows easily that, for $n \gg 0$,

$$\mathbb{P}(\mathscr{E}_n) \leq 2 \cdot \exp\left(-c_\epsilon \frac{K^2\pi}{2}\log n\right), \tag{21.7}$$

in other words,

$$\mathbb{P}(\mathscr{E}_n) \leq 2\dot{n}^{-C_\epsilon}, \tag{21.8}$$

where

$$C_\epsilon = c_\epsilon \frac{K^2\pi}{2}. \tag{21.9}$$

Hence, (21.3) will be satisfied provided

$$c_\epsilon \frac{K^2\pi}{2} > 1, \tag{21.10}$$

which tells us how large we need to choose K. Thus, we have proven Theorem 18.2 subject to establishing (21.6). First of all, by (21.4) and linearity of expectation, we have

$$\mu_n = \sum_{t=1}^{\lfloor n/2 \rfloor} \mathbb{E}[X_{n,t}]. \tag{21.11}$$

By the definition of $X_{n,t}$ in (21.5), its expectation is just the probability that both t and $n-t$ lie in A. For $t = n/2$ (n even) this is just $p_{n/2}$, otherwise it is $p_t p_{n-t}$. Since we are only interested in asymptotic estimates, it is thus clear that

$$\mu_n \sim \sum_{t=1}^{\lfloor n/2 \rfloor} p_t p_{n-t}. \tag{21.12}$$

By (21.1) there will be a bounded number of terms in this sum which are equal to 1. In all other terms, the minimum in (21.1) will be the function of t. Hence,

$$\mu_n \sim K^2 \sum_{t=1}^{\lfloor n/2 \rfloor} \sqrt{\frac{\log t \log(n-t)}{t(n-t)}}. \tag{21.13}$$

Note that the summand above is symmetric about $n/2$. Hence, in order to establish (21.6), it remains to prove that

$$\sum_{t=1}^{n-1} \sqrt{\frac{\log t \log(n-t)}{t(n-t)}} \sim \pi \log n. \tag{21.14}$$

Applying the standard trick of replacing the sum by an integral, we will show instead that[3]

$$\int_1^n \sqrt{\frac{\log t \log(n-t)}{t(n-t)}}\, dt \sim \pi \log n. \tag{21.15}$$

We change variables $t := \xi n$, and are left with having to show that

$$\int_0^1 \sqrt{\frac{\log(\xi n) \log[(1-\xi)n]}{\xi(1-\xi)}}\, d\xi \sim \pi \log n. \tag{21.16}$$

At this point, we need a bit of calculus :

Lemma 21.1.

$$\int_0^1 \frac{d\xi}{\sqrt{\xi(1-\xi)}} = \pi. \tag{21.17}$$

In particular, the integral converges, hence

$$\lim_{\delta \to 0} \int_0^{\delta} \frac{d\xi}{\sqrt{\xi(1-\xi)}} = \lim_{\delta \to 0} \int_{1-\delta}^1 \frac{d\xi}{\sqrt{\xi(1-\xi)}} = 0. \tag{21.18}$$

[3] It needs to be justified that replacing the sum by the integral does not lead to a significant error in our estimates. It is easy to see that the error will be $o(\log n)$. A rigorous proof is technical, and hence I omit it, though if you read through the rest of the calculations presented here, you should be able to see how to do it.

The second assertion of the lemma follows from the first (note that the integrand is symmetric about $\xi = 1/2$), and the first is proven by making the trigonometric substitution $\xi := \sin^2 \theta$.

So back to the theorem. Let δ be a small positive number. At the end we will let $\delta \to 0$. Divide up the integral in (21.16) into three parts, (i) from 0 to δ, (ii) from δ to $1-\delta$, (iii) from $1-\delta$ to 1. Call these three sub-integrals I_1, I_2 and I_3 respectively. Now, for any fixed $\xi \in (0,1)$, we have

$$\begin{aligned}\log(\xi n)\log[(1-\xi)n] &= (\log n + \log \xi)(\log n + \log(1-\xi)) \qquad (21.19)\\ &= (\log n + O(1))(\log n + O(1)) \sim (\log n)^2,\end{aligned}$$

so that the numerator of the integrand in (21.16) is $\sim \log n$. From this and Lemma 21.1, it follows easily that, as $\delta \to 0$,

$$I_1 \lesssim (\log n) \cdot \int_0^{\delta} \frac{d\xi}{\sqrt{\xi(1-\xi)}} = o(\log n), \qquad (21.20)$$

$$I_2 \sim (\log n) \cdot \int_{\delta}^{1-\delta} \frac{d\xi}{\sqrt{\xi(1-\xi)}} \sim \pi \log n, \qquad (21.21)$$

$$I_3 \lesssim (\log n) \cdot \int_{1-\delta}^{1} \frac{d\xi}{\sqrt{\xi(1-\xi)}} = o(\log n). \qquad (21.22)$$

Eq. (21.16) follows immediately from these estimates, and hence the proof of Theorem 18.2, for $h = 2$, is complete. □

Remark 21.2. If you examine the proof above, you will see that by a suitable choice of the parameters K and ϵ, we can choose positive contants $c_1 < c_2$ such that the quotient c_2/c_1 arbitrarily close to 1, and find an asymptotic basis A such that

$$c_1 \log n < r_2(A,n) < c_2 \log n \qquad (21.23)$$

for all $n \gg 0$.

We now come to the last part of the course, which is a quick introduction to the subject of *Generalised Ramsey Theory*, in particular the part of the subject which connects to number theory. For an introduction to the subject in general see, for example, the book

R. Graham, B. Rothschild and J. Spencer, *Ramsey Theory*, 2nd edition, Wiley, New York (1990).

We have already had a glimpse into this subject via the original work of Gordon Ramsey, which we presented in the now standard language of coloring the edges of graphs. As an example of a result of a certain type, which we can use as a launch pad to construct a general ‘theory’, the way to think about Theorem 19.1 is as follows : it says that, among objects of a certain class (graphs), if we pick some such very large object then, no matter how random it looks on the whole, it must have certain regularities in places. Here the ‘randomness’ is in the mish-mash of red and blue edges we observe when the edges of K_n are colored at random, whereas the ‘regularity’ is in the fact that, in places, the graph is entirely monochromatic.

Around the same time as Ramsey was at work, a Belgian mathematician called Bartel van der Waerden quite independently proved a result about numbers, rather than graphs, which sounds eerily similar to Ramsey's theorem. In modern terminology, his theorem is as follows :

Theorem 21.3. (van der Waerden 1927) *Let $k, l \in \mathbb{N}$. Then for all $n \gg 0$, depending on k and l, the following is true : if each of the numbers $1, 2, ..., n$ is given one of a selection of l colors then, no matter how the coloring is done, there must be a monochromatic k-term arithmetic progression (AP).*

Definition. For each pair k, l of positive integers, let $W(k, l)$ denote the smallest positive integer n such that for any l-coloring of $\{1, 2, ..., n\}$ there must exist a monochromatic k-term AP. These are called the *van der Waerden numbers* and Theorem 21.3 asserts that they all exist.

I gave some preliminary remarks about the proof of this theorem, but will present more details next day ...

22. Twenty-Second Lecture : 12/12

The main purpose of this lecture is to outline the proof of Theorem 21.3. I will not give the argument in full generality, in order to avoid getting bogged down in complicated notation and technicalities. Instead I will illustrate the key ideas by means of examples. First of all, note that it is trivial that

$$W(1,l)=1, \quad W(2,l)=l+1, \quad \text{for any } l. \tag{22.1}$$

The proof of the theorem proceeds by the following induction procedure : at step k, we assume that the numbers $W(k,l)$ exist for all l, and then deduce that the numbers $W(k+1,l)$ exist for all l. Eq. (22.1) allows us to get started. Below, I will prove three special cases of the theorem :

$$W(3,2) \leq 325, \tag{22.2}$$

$$W(3,3) \leq 7(2 \cdot 3^7+1)(2 \cdot 3^{30625}+1), \tag{22.3}$$

$$W(4,2) \leq [2 \cdot W(3,2)-1]\left[2 \cdot W(3,2^{2 \cdot W(3,2)-1})+1\right]. \tag{22.4}$$

In the proofs of (22.2) and (22.3), I will assume (22.1), while in the proof of (22.4), I will assume the existence of $W(3,l)$ for every l, a fact which I hope to convince you of by means of the first two proofs.

In going from (22.2) to (22.3), I wish to illustrate how the proof by induction proceeds with k fixed and l increasing. With (22.4), I illustrate the second key idea, namely what to do when k is increased.

In the lectures, I drew pictures to better show what was going on. I will not draw any pictures here, but it really is a lot easier to grasp the ideas with pictures, so please consult your lecture notes.

Proof of (22.2). We have to show that any 2-coloring of the numbers $1,2,...,325$ yields a monochromnatic 3-term AP. Let the colors be red and blue, and consider an arbitrary coloring. Divide the 325 numbers into 65 blocks of 5 consecutive numbers called $B_1,...,B_{65}$. Thus $B_1=\{1,2,3,4,5\}$, $B_2=\{6,7,8,9,10\}$ and so on. There are $2^5=32$ possible ways to color a block. Thus, amongst the first 33 blocks, two must be colored in exactly the same way, reading from left to right. Let B_i and B_{i+j} be any two such blocks. Since $i+j \leq 33$, the block B_{i+2j} exists. We now consider the three blocks B_i, B_{i+j} and B_{i+2j}, so

$$B_i=\{5i-4,5i-3,5i-2,5i-1,5i\}, \tag{22.5}$$

$$B_{i+j}=\{5(i+j)-4,5(i+j)-3,5(i+j)-2,5(i+j)-1,5(i+j)\}, \tag{22.6}$$

$$B_{i+2j}=\{5(i+2j)-4,5(i+2j)-3,5(i+2j)-2,5(i+2j)-1,5(i+2j)\} \tag{22.7}$$

Amongst the first three elements of B_i, two must have the same color (this is another way of saying that $W(2,2)=3$, so it is at this point that we are using an induction hypothesis). Let us suppose that $5i-4$ and $5i-2$ are each colored red - the argument is similar in all other cases and will not be repeated. If $5i$ is also red, then $(5i-4,5i-2,5i)$ is a red AP and we're done. Thus we may suppose that $5i$ is colored blue. Since B_{i+j} is colored in exactly the same way as B_i, we have that both $5(i+j)-4$ and $5(i+j)-2$ are red, whereas $5(i+j)$ is blue. Now we focus on the number $5(i+2j)$. If it is red, then $(5i-4,5(i+j)-2,5(i+2j))$ is a red AP. If it is blue, then $(5i,5(i+j),5(i+2j))$ is a blue AP. So we are done in either case.

Proof of (22.3). Let

$$n = 7(2\cdot 3^7+1)(2\cdot 3^{30625}+1) = 7\cdot 4375\cdot(2\cdot 3^{30625}+1) = 30625(2\cdot 3^{30625}+1). \quad (22.8)$$

We must show that any 3-coloring of the numbers $1, 2, ..., n$ yields a monochromatic 3-term AP. Let the colors be red, blue and green and consider an arbitrary coloring. First divide the numbers $1, 2, ..., n$ into $2 \cdot 3^{30625} + 1$ blocks B_ξ, each of length 30625. There are 3^{30625} ways to color a block, so amongst the first $3^{30625} + 1$ blocks, two must be colored in exactly the same way. Let B_i and B_{i+j} be any two such blocks. The block B_{i+2j} exists, and so we may consider the triple (B_i, B_{i+j}, B_{i+2j}) of blocks.

This time we first need to take the argument to another level, by looking inside the block B_i. It has length 30625, so we subdivide it into 4375 subblocks $C_{i,\xi}$, each of length 7. There are 3^7 ways to color a subblock, so amongst the first $3^7 + 1$ subblocks, two have exactly the same coloring. Let $C_{i,r}$ and $C_{i,r+s}$ be any two such subblocks. The subblock $C_{i,r+2s}$ exists. Let $C_{i+j,r}$ and so on denote the corresponding subblocks inside B_{i+j} and B_{i+2j}, and let $c_{i,r,\xi}$ and so on denote the elements of all these subblocks, where $\xi \in \{1, ..., 7\}$.

Now let's look at $C_{i,r}$. It has seven elements, so amongst the first four there must be two with the same color. Let us suppose that the first and fourth elements of $C_{i,r}$ are both red - the argument is similar in all other cases and will not be repeated. Thus $c_{i,r,1}$ and $c_{i,r,4}$ are both red. If $c_{i,r,7}$ were red, then we'd have a red 3-term AP inside $C_{i,r}$ and be done already. Without loss of generality, we may thus assume that $c_{i,r,7}$ is blue. Since $C_{i,r}$ and $C_{i,r+s}$ have the same coloring, we have $c_{i,r+s,1}$ and $c_{i,r+s,4}$ both red, whereas $c_{i,r+s,7}$ is blue. Now focus on the element $c_{i,r+2s,7}$. If it were red, then $(c_{i,r,1}, c_{i,r+s,4}, c_{i,r+2s,7})$ woul be a red AP. If it were blue, then $(c_{i,r,7}, c_{i,r+s,7}, c_{i,r+2s,7})$ would be a blue AP. We may thus assume that $c_{i,r+2s,7}$ is green.

Next, since B_i and B_{i+j} have the same coloring, we can deduce that
(i) $c_{i+j,r,1}, c_{i+j,r,4}, c_{i+j,r+s,1}$ and $c_{i+j,r+s,4}$ are all red,
(ii) $c_{i+j,r,7}$ and $c_{i+j,r+s,7}$ are blue,
(iii) $c_{i+j,r+2s,7}$ is green.

Finally, then, zoom in on the number $c_{i+2j,r+2s,7}$. If it is red, then $(c_{i,r,1}, c_{i+j,r+s,4}, c_{i+2j,r+2s,7})$ is a red AP. If it is blue, then $(c_{i,r,7}, c_{i+j,r+s,7}, c_{i+2j,r+2s,7})$ is a blue AP. If it is green, then $(c_{i,r+2s,7}, c_{i+j,r+2s,7}, c_{i+2j,r+2s,7})$ is a green AP. So we are done in all three cases.

Proof of (22.4). Put $a := 2 \cdot W(3, 2) - 1$, $b := W(3, 2^a)$ and

$$n := a(2b + 1). \quad (22.9)$$

We must show that any 2-coloring of the numbers $1, 2, ..., n$ yields a monochromatic 4-term AP. Divide all the numbers into $2b + 1$ blocks of length a. There are 2^a possible colorings of a block. Hence, amongst the first $W(3, 2^a)$ blocks, there must be a 3-term AP of blocks, all with the same coloring. Call these blocks B_i, B_{i+j} and B_{i+2j}. The block B_{i+3j} exists, so we may consider the 4-tuple $(B_i, B_{i+j}, B_{i+2j}, B_{i+3j})$ of blocks.

Since B_i has length $2 \cdot W(3, 2) - 1$, amongst the first $W(3, 2)$ elements of B_i there must exist a monochromatic 3-term AP. Suppose, without loss of generality, that $c_{i,r}, c_{i,r+s}, c_{i,r+2s}$

is a red AP inside B_i. The element $c_{i,r+3s}$ exists inside B_i. If it were red then we'd already have a 4-term red AP, so we may assume this element is blue. Since both B_{i+j} and B_{i+2j} are colored in exactly the same way as B_i, we now know that

(i) $c_{i+j,r}, c_{i+j,r+s}, c_{i+j,r+2s}, c_{i+2j,r}, c_{i+2j,r+s}$ and $c_{i+2j,r+2s}$ are all red,

(ii) $c_{i+j,r+3s}$ and $c_{i+2j,r+3s}$ are blue.

Now focus on the element $c_{i+3j,r+3s}$. If it is red, then $(c_{i,r}, c_{i+j,r+s}, c_{i+2j,r+2s}, c_{i+3j,r+3s})$ is a red AP. If it is blue, then $(c_{i,r+3s}, c_{i+j,r+3s}, c_{i+2j,r+3s}, c_{i+3j,r+3s})$ is a blue AP. So in both cases, we are done. This completes the proof of (22.4), and thus our outline of the proof of Theorem 21.3.

We finish off the course with some remarks, for anyone interested in pursuing further studies in this area. The following is an equivalent formulation of van der Waerden's theorem :

Theorem 22.1. *Suppose the natural numbers are colored with finitely many colors. Then there exist arbitrarily long monochromatic AP:s.*

Proof. It is very easy to see that Theorem 21.3 implies this result. The converse requires a bit more effort and is left as an exercise. □

If we color $\mathbb{N}$ with finitely many colors, then at least one color must be used on a 'positive proportion' (more precise definitions to follow) of all numbers. It is natural to expect that the color which is used 'most often' will yield arbitrarily long AP:s. Erdős and Turán conjectured that this is the case. To state their conjecture precisely, we need a definition :

Definition. Let $A \subseteq \mathbb{N}$. The *upper asymptotic density* of A, denoted $\overline{d}(A)$, is the quantity

$$\overline{d}(A) := \limsup_{n\to\infty} \frac{|A \cap \{1, ..., n\}|}{n}. \tag{22.10}$$

Erdős and Turán conjectured that if $\overline{d}(A) > 0$ then A contains arbitrarily long AP:s. This turns out to be true. The first progress was made by Roth (1952) who proved that a set of positive upper density must contain a 3-term AP. His proof used Fourier analysis. Szemerédi (1975) completely solved the problem. His method is purely combinatorial, but anything but easy, and the argument is considered one of the classics of combinatorics. Furstenberg (1977) gave a completely different proof of what has become known as *Szemerédi's Theorem*, using ergodic theory. Yet another entirely different proof was provided by Gowers (1998). His methods build on those of Roth, but essentially involved the development of a whole new kind of Fourier analysis. In 2004, Green and Tao synthesised the ideas of all these authors to solve a very old problem in classical number theory : they proved that the set of primes contains arbitrarily long arithmetic progressions. The field is currently a very hot area of research. There are two outstanding open problems, of which the first is :

Extended Erdős-Turán Conjecture. *Let $A \subseteq \mathbb{N}$ satisfy that*

$$\sum_{a\in A} \frac{1}{a} = +\infty. \tag{22.11}$$

Then A contains arbitrarily long AP:s.

Observe that the set of primes satisfies (22.11), hence the Green-Tao result established a special case of this conjecture. It is at this point unclear, however, whether their methods can shed any light on the general conjecture. Erdős offered $\$3000$ for a proof or disproof of this conjecture. No Erdős problem worth more than $\$1000$ has yet been solved.

The second outstanding open problem is the so-called *Hardy-Littlewood k-tuple conjecture*. It is somewhat technical to state precisely, but informally it states that any possible constellation of prime numbers should appear infinitely often, unless it is ruled out by some simple congrence obstruction. The *Twin Primes Conjecture* is a special case of this more general conjecture. An example of a constellation which is ruled out by a congruence obstruction is $\{n, n+2, n+4\}$. Since, for any n, one of these three numbers is a multiple of 3, the constellation appears only once among the primes, namely $\{3, 5, 7\}$. The work of Green-Tao may be able to be pushed further towards establishing some cases of the Hardy-Littlewood conjecture. Their published work in the last few years is certainly motivated by this application. Possibly in the foreseeable future, the ideas which grew out of van der Waerden's strange little result on coloring numbers may lead to a proof of the Twin Primes Conjecture !

One final remark. The upper bounds on van der Waerden numbers $W(k, l)$ obtained by the kind of *color focusing* argument we outlined above are eeeeeenoooooorrrrrrmoooooouussss[4]. Even for $W(3, 2)$, the actual value is known to be 9, which can be checked by exhaustive computer search. The best-known upper bounds on VdW numbers have been obtained by the Fourier analysis methods of Gowers et al. It is known that

$$W(k, l) \leq e^{l^{2^{2^{k+9}}}}. \tag{22.12}$$

Regarding lower bounds, a probabilistic argument similar to that used in the proof of Theorem 19.3 can be used to prove that

$$W(k, l) > \sqrt{2(k-1)}\, l^{(k-1)/2}. \tag{22.13}$$

The huge gap between (22.12) and (22.13) illustrates that there is still plenty of room for new ideas in this area.

[4] More precisely, they are not *primitive recursive*, for those of you who know what that means.