# Chapter 3: The Second Moment

The Probabilistic Method

Summer 2020

Freie Universität Berlin

# Chapter Overview

- Introduce the second moment method
- Survey applications in graph theory and number theory

§1 Concentration Inequalities

Chapter 3: The Second Moment
The Probabilistic Method

§2 Thresholds

Chapter 3: The Second Moment
The Probabilistic Method

§3 Subgraphs of $G(n,p)$

Chapter 3: The Second Moment
The Probabilistic Method

§4 Prime Factors

Chapter 3: The Second Moment
The Probabilistic Method

§5 Distinct Sums

Chapter 3: The Second Moment
The Probabilistic Method

# §1 Concentration Inequalities

Chapter 3: The Second Moment

The Probabilistic Method

# What Does the Expectation Mean?

## Basic fact

- $\{X \leq \mathbb{E}[X]\}$ and $\{X \geq \mathbb{E}[X]\}$ have positive probability
- Often want more quantitative information
  - What are these positive probabilities?
  - How much below/above the expectation can the random variable be?

## Limit laws

- Law of large numbers
  - Average of independent trials will tend to the expectation
- Central limit theorem
  - Average will be normally distributed

## Not always applicable

- We often only have a single instance, or lack independence
- Can still make use of more general bounds

# Markov's Inequality

## Theorem 3.1.1 (Markov's Inequality)

Let $X$ be a non-negative random variable, and let $a > 0$. Then

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a}.$$

## Proof

- Let $f$ be the density function for the distribution of $X$
- $\mathbb{E}[X] = \int_0^\infty xf(x)\,dx = \int_0^a xf(x)\,dx + \int_a^\infty xf(x)\,dx$

$$\geq \int_a^\infty xf(x)\,dx \geq \int_a^\infty af(x)\,dx = a\int_a^\infty f(x)\,dx = a\mathbb{P}(X \geq a) \quad \blacksquare$$

Moral: $\mathbb{E}[X]$ small $\Rightarrow$ $X$ typically small

# Chebyshev's Inequality

Converse? Does $\mathbb{E}[X]$ large $\Rightarrow$ $X$ typically large?

- Not necessarily; e.g. $X = n^2$ with probability $n^{-1}$, 0 otherwise
- But such random variables have large variance…

Theorem 3.1.2 (Chebyshev's Inequality)

Let $X$ be a random variable, and let $a > 0$. Then

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

Proof

- $\{|X - \mathbb{E}[X]| \geq a\} = \{(X - \mathbb{E}[X])^2 \geq a^2\}$
- Let $Y = (X - \mathbb{E}[X])^2$
- Then $\mathbb{E}[Y] = \mathrm{Var}(X)$
- Apply Markov's Inequality ∎

# Using Chebyshev

## Moral

- $\mathbb{E}[X]$ large *and* $\mathrm{Var}(X)$ small $\Rightarrow$ $X$ typically large
- Special case: showing $X$ nonzero

## Corollary 3.1.3

If $\mathrm{Var}(X) = o(\mathbb{E}[X]^2)$, then $\mathbb{P}(X = 0) = o(1)$.

## Proof

- $\{X = 0\} \subseteq \{|X - \mathbb{E}[X]| \geq |\mathbb{E}[X]|\}$
- Chebyshev $\Rightarrow$ $\mathbb{P}(|X - \mathbb{E}[X]| \geq |\mathbb{E}[X]|) \leq \frac{\mathrm{Var}(x)}{\mathbb{E}[X]^2} = o(1)$ ■

- In fact, in this case $X = \big(1 + o(1)\big)\mathbb{E}[X]$ with high probability

# Typical application

## Set-up

- $E_i$ events, occurring with probability $p_i$
- $X_i = 1_{E_i}$ their indicator random variables
- $X = \sum_i X_i$ their sum, the number of occurring events

## Goal

- Show that with high probability, some event occurs

## Applying Chebyshev

- Need to show $\mathrm{Var}(X) = o(\mathbb{E}[X]^2)$

## Expand the variance

- $\mathrm{Var}(X) = \mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$

# Some Simplification

## Estimating the summands

- $\mathrm{Var}(X) = \mathrm{Var}(\sum_i X_i) = \sum_i \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j)$
- $\mathrm{Var}(X_i) = p_i(1 - p_i) \leq p_i$
  - $\therefore \sum_i \mathrm{Var}(X_i) \leq \sum_i p_i = \sum_i \mathbb{E}[X_i] = \mathbb{E}[X]$
- $\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$
  - $\mathrm{Cov}(X, Y) = 0$ if $X$ and $Y$ are independent
  - Otherwise $\mathrm{Cov}(X_i, X_j) \leq \mathbb{E}[X_i X_j] = \mathbb{P}(E_i \wedge E_j)$

## Corollary 3.1.4

Let $\{E_i\}$ be a sequence of events with probabilities $p_i$, and let $X$ count the number of events that occur. Write $i \sim j$ if the events $E_i$ and $E_j$ are not independent, and let $\Delta = \sum_{i \sim j} \mathbb{P}(E_i \wedge E_j)$. If $\mathbb{E}[X] \to \infty$ and $\Delta = o(\mathbb{E}[X]^2)$, then $P(X = 0) = o(1)$.

# Any questions?

# §2 Thresholds

Chapter 3: The Second Moment

The Probabilistic Method

# Monotone properties

## Graph properties

- Say a graph $\mathcal{P}$ is *monotone (increasing)* if adding edges preserves $\mathcal{P}$
- e.g.: containing a subgraph $H \subseteq G$, having $\alpha(G) < k$, connectivity, …

## Lemma 3.2.1

If $\mathcal{P}$ is a monotone increasing graph property, then $\mathbb{P}(G(n,p) \in \mathcal{P})$ is monotone increasing in $p$.

## Proof (Coupling)

- Sampling $G(n,p)$
  - Assign to each pair of vertices $\{u,v\}$ an independent uniform $Y_{u,v} \sim \mathrm{Unif}([0,1])$
  - Add edge $\{u,v\}$ to $G$ iff $Y_{u,v} \le p$
  - Each edge appears independently with probability $p$
- If $p \le p'$, then $G(n,p) \subseteq G(n,p') \Rightarrow$ if $G(n,p) \in \mathcal{P}$, then $G(n,p') \in \mathcal{P}$ ∎

# Thresholds

## Transitions

- A monotone property $\mathcal{P}$ is *nontrivial* if it is not satisfied by the edgeless graph, and is satisfied by the complete graph
  - $\Rightarrow \mathbb{P}(G(n,0) \in \mathcal{P}) = 0$ and $\mathbb{P}(G(n,1) \in \mathcal{P}) = 1$
- Lemma 3.2.1 $\Rightarrow \mathbb{P}(G(n,p) \in \mathcal{P})$ increases from 0 to 1 as $p$ does
- How quickly does this increase happen?

### Definition 3.2.2 (Thresholds)

Given a nontrivial monotone graph property $\mathcal{P}$, $p_0(n)$ is a threshold for $\mathcal{P}$ if

$$\mathbb{P}(G(n,p) \in \mathcal{P}) \to \begin{cases} 0 \text{ if } \mathrm{p} \ll p_0(n), \\ 1 \text{ if } p \gg p_0(n). \end{cases}$$

# A Cyclic Example

## Proposition 3.2.3

The threshold for $G(n,p)$ to contain a cycle is $p_0(n) = \frac{1}{n}$.

## Proof (lower bound)

- Let $X = \#$ cycles in $G(n,p)$
- For $\ell \geq 3$, let $X_\ell = \#\{C_\ell \subseteq G(n,p)\}$
  - $\Rightarrow X = \sum_{\ell=3}^{n} X_\ell$
- Linearity of expectation: $\mathbb{E}[X_\ell] \leq n^\ell p^\ell$
- $\Rightarrow \mathbb{E}[X] \leq \sum_{\ell=3}^{n} (np)^\ell < (np)^3 \sum_{\ell=0}^{\infty} (np)^\ell = \frac{(np)^3}{1-np}$
  - $\Rightarrow \mathbb{E}[X] = o(1)$ if $p \ll \frac{1}{n}$
- Markov: $\mathbb{P}(G(n,p)$ has a cycle$) = \mathbb{P}(X \geq 1) \leq \mathbb{E}[X] \to 0$ ■

# Cycles Continued

## Proposition 3.2.3

The threshold for $G(n,p)$ to contain a cycle is $p_0(n) = \frac{1}{n}$.

**Proof** (upper bound)

- Let $p = \frac{4}{n-1}$ and set $Y = e\big(G(n,p)\big)$
- Then $Y \sim \operatorname{Bin}\left(\binom{n}{2}, p\right)$
  - $\Rightarrow \mathbb{E}[Y] = \binom{n}{2}p = 2n$
  - $\Rightarrow \operatorname{Var}(Y) = \binom{n}{2}p(1-p) < 2n$
- $\therefore \operatorname{Var}(Y) = o(\mathbb{E}[Y]^2)$
- Chebyshev: $\mathbb{P}(Y < n) \to 0$
- $\mathbb{P}(G(n,p) \text{ has a cycle}) \geq \mathbb{P}\Big(e\big(G(n,p)\big) \geq n\Big) \to 1$ ■

# Existence of Thresholds

**Theorem 3.2.4 (Bollobás-Thomason, 1987)**

Every nontrivial monotone graph property has a threshold.

**Proof** (upper bound)

- Let $p(n) = p_0$ be such that $\mathbb{P}(G(n, p_0) \in \mathcal{P}) = \frac{1}{2}$
- Let $G \sim G_1 \cup G_2 \cup \cdots \cup G_m$, where each $G_i \sim G(n, p_0)$ is independent
  - $\Rightarrow G \sim G(n, p)$ for $p \coloneqq 1 - (1 - p_0)^m \leq mp_0$
- Property is monotone:
  - $\mathbb{P}(G \in \mathcal{P}) \geq \mathbb{P}(\cup_i \{G_i \in \mathcal{P}\}) = 1 - \mathbb{P}(\cap_i \{G_i \notin \mathcal{P}\})$
- Graphs are independent:
  - $\mathbb{P}(\cap_i \{G_i \notin \mathcal{P}\}) = \prod_i \mathbb{P}(G_i \notin \mathcal{P})$
- Since $G_i \sim G(n, p_0)$, $\mathbb{P}(G_i \notin \mathcal{P}) = \frac{1}{2}$
- $\therefore \mathbb{P}(G \in \mathcal{P}) \geq 1 - 2^{-m} \to 1$ if $m \to \infty$ (or if $p \gg p_0$) ■

# Below the Threshold

**Theorem 3.2.4 (Bollobás-Thomason, 1987)**

Every nontrivial monotone graph property has a threshold.

**Proof** (lower bound)

- Let $G \sim G_1 \cup G_2 \cup \cdots \cup G_m$ as before, but with $G_i \sim G(n,p)$ for $p = \frac{p_0}{m}$
- $\Rightarrow G \sim G(n,q)$ for $q = 1 - (1-p)^m \leq mp = p_0$
- $\Rightarrow \mathbb{P}(G \notin \mathcal{P}) \geq \frac{1}{2}$
- As before, $\mathbb{P}(G \notin \mathcal{P}) \leq \mathbb{P}(G(n,p) \notin \mathcal{P})^m$
- $\Rightarrow \mathbb{P}(G(n,p) \notin \mathcal{P}) \geq \left(\frac{1}{2}\right)^{1/\mathrm{m}}$
- $\Rightarrow \mathbb{P}(G(n,p) \in \mathcal{P}) \leq 1 - \left(\frac{1}{2}\right)^{1/m} \to 0$ if $m \to \infty$ (or if $p \ll p_0$) ■

# Closing Remarks

## Random graph theory

- Fundamental problem: given a graph property $\mathcal{P}$, what is its threshold?

## At the threshold

- We showed what happens for probabilities much smaller than the threshold, and much larger than the threshold
- What if $p = \Theta\big(p_0(n)\big)$? Some properties have a much quicker transition

## Definition 3.2.5 (Sharp thresholds)

We say $p_0(n)$ is a *sharp* threshold for $\mathcal{P}$ if there are positive constants $c_1, c_2$ such that

$$\mathbb{P}(G(n,p) \in \mathcal{P}) \to \begin{cases} 0 \text{ if } p \leq c_1 p_0(n), \\ 1 \text{ if } p \geq c_2 p_0(n). \end{cases}$$

# Any questions?

# §3 Subgraphs of $G(n, p)$

Chapter 3: The Second Moment

The Probabilistic Method

# Returning to Ramsey

**Theorem 1.5.7**

Given $\ell, k, n \in \mathbb{N}$ and $p \in [0,1]$, if

$$\binom{n}{\ell} p^{\binom{\ell}{2}} + \binom{n}{k} (1-p)^{\binom{k}{2}} < 1,$$

then $R(\ell, k) > n$.

**Choosing parameters**

- Want to choose $n$ as large as possible
- Need to avoid large independent sets
  - $\Rightarrow$ would like to make edge probability $p$ large
- Limitation: need to avoid $K_\ell$

**Question:** What is the threshold for $\mathrm{K}_\ell \subseteq G(n,p)$?

# A Lower Bound

## Goal

- Let $X$ count the number of $K_\ell$ in $G(n, p)$
- For which $p$ do we have $\mathbb{P}(X \geq 1) = o(1)$?

## First moment

- $\mathbb{E}[X] = \binom{n}{\ell} p^{\binom{\ell}{2}} = \Theta\left(n^\ell p^{\binom{\ell}{2}}\right)$
- Markov's Inequality: $\mathbb{P}(X \geq 1) \leq \mathbb{E}[X]$

## Threshold bound

- $\mathbb{E}[X] = \Theta\left(n^\ell p^{\binom{\ell}{2}}\right) \ll 1$
- $\Leftrightarrow p^{\binom{\ell}{2}} \ll n^{-\ell} \Leftrightarrow p \ll n^{-2/(\ell-1)}$
- $\Rightarrow p_0(n) \geq n^{-2/(\ell-1)}$

# An Upper Bound

## Goal

- For which $p$ do we have $\mathbb{P}(X = 0) = o(1)$?

## Corollary 3.1.4

Let $\{E_i\}$ be a sequence of events with probabilities $p_i$, and let $X$ count the number of events that occur. Write $i \sim j$ if the events $E_i$ and $E_j$ are not independent, and let $\Delta = \sum_{i \sim j} \mathbb{P}\left(E_i \wedge E_j\right)$. If $\mathbb{E}[X] \to \infty$ and $\Delta = o(\mathbb{E}[X]^2)$, then $P(X = 0) = o(1)$.

## Our parameters

- Let $G \sim G(n, p)$ and, for $S \in \binom{[n]}{\ell}$, let $E_S = \{G[S] \cong K_\ell\}$
- $\mathbb{E}[X] = \binom{n}{\ell} p^{\binom{\ell}{2}} \to \infty$ for $p \gg n^{-2/(\ell-1)}$
- Suffices to show $\Delta = o(\mathbb{E}[X]^2)$

# Clique Dependencies

## Independent events

- $E_i$ occurs $\Leftrightarrow$ all edges in $i$th clique present
- Edges appear independently
- $\therefore\ |S \cap T| \leq 1 \Rightarrow E_S, E_T$ independent

## Dependent events

- Suppose $|S \cap T| = s \geq 2$
  - $\Rightarrow S \sim T$
- $E_S \wedge E_T$: $G[S], G[T]$ both $\ell$-cliques, sharing $s$ vertices
- Number of prescribed edges: $2\binom{\ell}{2} - \binom{s}{2}$
- $\Rightarrow \mathbb{P}(E_S \wedge E_T) = p^{2\binom{\ell}{2} - \binom{s}{2}}$

# Computing Δ

## Recall

- $S \sim T \Leftrightarrow s := |S \cap T| \geq 2$
- $\mathbb{P}(E_S \wedge E_T) = p^{2\binom{\ell}{2}-\binom{s}{2}}$

## Substituting terms

$$\begin{aligned}\Delta &= \textstyle\sum_{S\sim T} \mathbb{P}(E_S \wedge E_T) = \sum_{|S\cap T|\geq 2} \mathbb{P}(E_S \wedge E_T) \\ &= \textstyle\sum_S \sum_{T:|S\cap T|\geq 2} \mathbb{P}(E_S \wedge E_T) \\ &= \textstyle\sum_S \sum_{s=2}^{\ell-1} \sum_{T:|S\cap T|=s} \mathbb{P}(E_S \wedge E_T) \\ &= \textstyle\sum_S \sum_{s=2}^{\ell-1} \sum_{T:|S\cap T|=s} p^{2\binom{\ell}{2}-\binom{s}{2}}\end{aligned}$$

$$\Rightarrow \Delta = \binom{n}{\ell} \textstyle\sum_{s=2}^{\ell-1} \binom{\ell}{s}\binom{n-\ell}{\ell-s} p^{2\binom{\ell}{2}-\binom{s}{2}}$$

# Bounding $\Delta$

## Recall

- $\Delta = \binom{n}{\ell} \sum_{s=2}^{\ell-1} \binom{\ell}{s} \binom{n-\ell}{\ell-s} p^{2\binom{\ell}{2}-\binom{s}{2}}$

## Goal

- Show $\Delta = o(\mathbb{E}[X]^2)$

## Estimates

- $\binom{\ell}{s} \leq 2^{\ell}$
- $\binom{n-\ell}{\ell-s} \leq n^{\ell-s} = \Theta\left(\binom{n}{\ell} n^{-s}\right)$

## Bound

$$\Delta \leq \binom{n}{\ell} \sum_{s=2}^{\ell-1} 2^{\ell} \Theta\left(\binom{n}{\ell} n^{-s}\right) p^{2\binom{\ell}{2}-\binom{s}{2}}$$

$$= \binom{n}{\ell}^2 p^{2\binom{\ell}{2}} \sum_{s=2}^{\ell-1} \Theta\left(n^{-s} p^{-\binom{s}{2}}\right) = \mathbb{E}[X]^2 \sum_{s=2}^{\ell-1} \Theta\left(n^{-s} p^{-\binom{s}{2}}\right)$$

# Completing the Calculation

## Recall

- $\Delta = \mathbb{E}[X]^2 \sum_{s=2}^{\ell-1} \Theta\left(n^{-s} p^{-\binom{s}{2}}\right)$

## Substituting $p$

- $n^{-s} p^{-\binom{s}{2}} = \left(np^{(s-1)/2}\right)^{-s}$
- We took $p \gg n^{-2/(\ell-1)}$
- $\Rightarrow n^{-s} p^{-\binom{s}{2}} \ll \left(\mathrm{n}^{1-(s-1)/(\ell-1)}\right)^{-s}$
- For $2 \le s \le \ell - 1$, this is $o(1)$
- $\Rightarrow \Delta = o(1)$

## Theorem 3.3.1

For $\ell \ge 2$, the threshold for $\mathrm{K}_\ell \subseteq G(n,p)$ is $p_0(n) = n^{-2/(\ell-1)}$.

# An Incomplete Result

## Problem

Given a graph $H$, what is the threshold $p_0^H(n)$ for $H \subseteq G(n,p)$?

## Lower bound

- Let $X$ be the number of copies of $H$ in $G(n,p)$
- Markov: $\mathbb{E}[X] = o(1) \Rightarrow p_0(n) \gg p$

## Expectation

- Number of possible copies
  - Specify vertices of $H$ – at most $n^{v(H)}$ possibilities
- Probability of appearance
  - Each edge of $H$ must be present – probability is $p^{e(H)}$
- $\Rightarrow \mathbb{E}[X] \leq n^{v(H)} p^{e(H)}$

## Conclusion: $p_0(n) \geq n^{-v(H)/e(H)}$

# An Illustrated Example

## Graph statistics

- Let $H$ be $K_4$ with a pendant edge
- Statistics:
  - $v(H) = 5$
  - $e(H) = 7$
- $\Rightarrow p_0^H(n) \geq n^{-5/7}$

## An issue

- $K_4 \subseteq H$
- $\Rightarrow$ if $H \subseteq G(n,p)$, then $K_4 \subseteq G(n,p)$
- $\Rightarrow p_0^{K_4}(n) \leq p_0^H(n)$
- But we showed $p_0^{K_4}(n) = n^{-2/3} \gg n^{-5/7}$

# Monotonicity and Density

## General lower bound

- $p_0^H(n) \geq \max\{p_0^F(n): F \subseteq H\}$
- Can substitute first moment bound
- $\Rightarrow p_0^H(n) \geq \max\left\{n^{-v(F)/e(F)}: F \subseteq H, e(F) \geq 1\right\}$

## Definition 3.3.2 (Maximum density)

Given a graph $H$, define $\mathrm{d}(H) = \frac{e(H)}{v(H)}$, and let $m(H) = \max\{d(F): F \subseteq H\}$.

## Remarks

- We have $p_0^H(n) \geq n^{-1/m(H)}$
- Say $H$ is *balanced* if $\mathrm{d}(H) = m(H)$
- $H$ is *strictly balanced* if $d(F) < m(H)$ for all $F \subset H$

# Expected Subgraph Counts

## Boundless expectations

- Let $X_H$ be the number of copies of $H$ in $G(n, p)$
- Total # possible copies $= \Theta\left(n^{v(H)}\right)$
- Probability of each copy: $p^{e(H)}$
- $\Rightarrow \mathbb{E}[X_H] = \Theta\left(n^{v(H)} p^{e(H)}\right)$
- $\therefore \mathbb{E}[X_H] \to \infty$ when $p \gg n^{-v(H)/e(H)}$

## Guaranteeing subgraph existence

- Goal: to show $\mathbb{P}(X_H = 0) = o(1)$ for $p \gg p_0^H(n)$
- Apply second moment: need to show $\Delta = \mathrm{o}(\mathbb{E}[X_H]^2)$
- Edge-disjoint copies are independent

# Dependent Subgraphs

## Common subgraphs

- Let $H_1, H_2$ be two copies of $H$ sharing an edge
  - $E_{H_1} \wedge E_{H_2} = \{H_1 \cup H_2 \subseteq G(n,p)\}$
- Let $F := H_1 \cap H_2$ be the common subgraph
  - $v(H_1 \cup H_2) = 2v(H) - v(F)$
  - $e(H_1 \cup H_2) = 2e(H) - e(F)$

## Counting pairs

- Group dependent pairs $(H_1, H_2)$ by common subgraphs $F = H_1 \cap H_2$
- At most $2^{e(H)}$ possible subgraphs $F$
- For each $J$, $\mathrm{O}\left(n^{2v(H)-v(F)}\right)$ pairs $(H_1, \mathrm{H}_2)$
- For each such pair, $\mathbb{P}\left(E_{H_1} \wedge E_{H_2}\right) = p^{2e(H)-e(F)}$

# Bounding Δ

## Recall

$$\Delta = \sum_{i \sim j} \mathbb{P}\left(E_{H_i} \wedge E_{H_j}\right)$$

## Group by common subgraph

$$\Delta = \sum_{i \sim j} \mathbb{P}\left(E_{H_i} \wedge E_{H_j}\right) = \sum_{F \subset H} \sum_{(i,j):H_i \cap H_j = F} \mathbb{P}\left(E_{H_i} \wedge E_{H_j}\right)$$

## Substitute estimates

$$\Delta = \sum_{F \subset H} O\left(n^{2v(H)-v(F)} p^{2e(H)-e(F)}\right)$$

$$\Rightarrow \Delta = \left(n^{v(H)} p^{e(H)}\right)^2 \sum_{F \subset H} O\left(n^{-v(F)} p^{-e(F)}\right)$$

$$\Rightarrow \Delta = \mathbb{E}[X_H]^2 \sum_{F \subset H} O\left(n^{-v(F)} p^{-e(F)}\right)$$

# A Complete Solution

## Recall

- $\Delta = \mathbb{E}[X_H]^2 \sum_{F \subset H} O\left(n^{-v(F)} p^{-e(F)}\right)$

## Choice of $p$

- We have $p \gg n^{-1/m(H)}$
- $\Rightarrow p \gg n^{-v(F)/e(F)}$ for all nonempty $F \subset H$
- $\Rightarrow n^{-v(F)} p^{-e(F)} = o(1)$
- $\Rightarrow \Delta = o(1)$

## Theorem 3.3.3

Given a graph $H$, the threshold for $H \subseteq G(n,p)$ is $p_0^H(n) = n^{-1/m(H)}$, where

$$m(H) = \max \left\{ \frac{e(F)}{v(F)} : F \subseteq H \right\}.$$

# Any questions?

# §4 Prime Factors

Chapter 3: The Second Moment

The Probabilistic Method

# Time For Primes

## Fun facts

- There are infinitely many primes (Euclid, -300)
- The primes contain arbitrarily long arithmetic progressions (Green–Tao, 2004)
- Infinitely many pairs of primes are at most 70000000 apart (Zhang, 2014)

## Central problem

- How are the primes distributed in $\mathbb{N}$?

**Theorem 3.4.1 (Hadamard, De la Vallée Poussin, 1896)**

The number $\pi(n)$ of prime numbers in $[n]$ satisfies

$$\pi(n) = \big(1 + o(1)\big)\frac{n}{\ln n}.$$

# Prime Factorisation

## The funnest of facts

- Every natural number is the product of primes

## Our goal

- To understand what these factorisations look like

## Definition 3.4.2

Given $x \in \mathbb{N}$, let $\nu(x)$ denote the number of *distinct* prime factors of $x$.

## Examples

- $\nu(19) = ?$
- $\nu(210) = ?$
- $\nu(256) = ?$
- $\nu(2020) = ?$

# The Average Case

## Proposition 3.4.3

The average number of distinct prime factors of a number $x \in [n]$ is $\ln \ln n + O(1)$.

## Proof

- Express $\nu(x)$ in terms of indicator random variables:
  - $\nu(x) = \sum_{p \le n} 1_{\{p|x\}}$
- Exchange order of summation
  - $\frac{1}{n} \sum_{x \in [n]} \nu(x) = \frac{1}{n} \sum_{p \le n} \sum_{x \in [n]} 1_{\{p|x\}}$
- Count multiples
  - $\sum_{x \in [n]} 1_{\{p|x\}} = \left\lfloor \frac{n}{p} \right\rfloor = \frac{\mathrm{n}}{\mathrm{p}} + \mathrm{O}(1)$
- $\Rightarrow \frac{1}{n} \sum_{x \in [n]} \nu(x) = \sum_{p \le n} \frac{1}{p} + O(1) = \ln \ln n + O(1)$ ∎

# A Harmonic Digression

**Theorem 3.4.4 (Mertens, 1874)**

As $n \to \infty$, we have $\sum_{p \le n} \frac{1}{p} = \ln \ln n + O(1)$.

"Proof"

- Let $m = \pi(n) \sim \frac{n}{\ln n}$
  - $\sum_{p \le n} \frac{1}{p} = \sum_{k=1}^{m} \frac{1}{p_k}$
- Prime Number Theorem $\Rightarrow p_k \sim k \ln k$
  - $\Rightarrow \sum_{p \le n} \frac{1}{p} \sim \sum_{k=2}^{m} \frac{1}{k \ln k}$
- Approximate by an integral:
  - $\sum_{k=2}^{m} \frac{1}{k \ln k} \sim \int_{x=2}^{m} \frac{1}{x \ln x} dx \sim \ln \ln m \sim \ln \ln n$ ∎

# The Typical Case

## Variation in $\nu(x)$, $x \in [n]$

- Minimum: $1$
- Average: $\ln \ln n + O(1)$
- Maximum: $\big(1 + o(1)\big)\frac{\ln n}{\ln \ln n}$
  - Product of first $m$ primes $\sim \prod_{k=1}^{m} k \ln k \sim m!\,(\ln m)^m \leq n$ for $m \sim \frac{\ln n}{\ln \ln n}$

## What can we say about the distribution of $\nu(x)$?

**Theorem 3.4.5 (Hardy-Ramanujan, 1920)**

As $n \to \infty$, we have $\nu(x) = \big(1 + o(1)\big) \ln \ln n$ for all but $o(n)$ integers $x \in [n]$.

# The Probabilistic Approach

## Theorem 3.4.5 (Hardy-Ramanujan, 1920)

As $n \to \infty$, we have $\nu(x) = \big(1 + o(1)\big) \ln \ln n$ for all but $o(n)$ integers $x \in [n]$.

## Probabilistic proof (Turán, 1934)

- Choose $x \in [n]$ uniformly at random
- Interested in the random variable $X = \nu(x)$
- Proposition 3.4.3 $\Rightarrow$ $\mathbb{E}[X] = \ln \ln n + O(1)$

## Corollary 3.1.3'

If $\mathrm{Var}(X) = o(\mathbb{E}[X]^2)$, then $X = \big(1 + o(1)\big)\mathbb{E}[X]$ with high probability.

# Expressing the Variance

## Recall

- $x \in [n]$ uniformly random
- $X = \nu(x)$ number of distinct prime factors
- Goal: show $\mathrm{Var}(X) = o(\mathbb{E}[X]^2)$

## Indicator random variables

- For a prime $p$, let $X_p = 1_{\{p|x\}}$, Bernoulli random variable
- $\mathbb{P}(X_p = 1) = \frac{\lfloor n/p \rfloor}{\mathrm{n}} \in \left(\frac{1}{p} - \frac{1}{n}, \frac{1}{p}\right]$
- $X = \sum_{p \leq n} X_p$

## Our friend the variance

- $\mathrm{Var}(X) = \sum_p \mathrm{Var}(X_p) + \sum_{(p,q):p \neq q} \mathrm{Cov}(X_p, X_q)$
- $\sum_p \mathrm{Var}(X_p) \leq \sum_p \mathbb{E}[X_p] = \mathbb{E}[X]$

# Computing Covariances

## Pairs $p \neq q$

- $\mathrm{Cov}(X_p, X_q) = \mathbb{E}[X_p X_q] - \mathbb{E}[X_p]\mathbb{E}[X_q]$
- $\mathbb{E}[X_p] \geq \frac{1}{p} - \frac{1}{n}$, $\mathbb{E}[X_q] \geq \frac{1}{q} - \frac{1}{\mathrm{n}}$
- $\mathbb{E}[X_p X_q] = \mathbb{P}(pq|x) \leq \frac{1}{pq}$
- $\Rightarrow \mathrm{Cov}(X_p, X_q) \leq \frac{1}{pq} - \left(\frac{1}{p} - \frac{1}{n}\right)\left(\frac{1}{q} - \frac{1}{n}\right) \leq \frac{1}{n}\left(\frac{1}{p} + \frac{1}{q}\right)$

## Bounding the sum

- $\Rightarrow \sum_{(p,q):p\neq q} \mathrm{Cov}(X_p, X_q) \leq \frac{1}{n}\sum_{(p,q):p\neq q}\left(\frac{1}{p} + \frac{1}{q}\right) \leq \frac{2\pi(n)}{n}\sum_{p\leq n}\frac{1}{p}$
- $\pi(n) = (1 + o(1))\frac{n}{\ln n}$ and $\sum_{p\leq n}\frac{1}{p} = \ln\ln n + O(1) = \mathbb{E}[X]$
- $\Rightarrow \sum_{(p,q):p\neq q} \mathrm{Cov}(X_p, X_q) = o(\mathbb{E}[X])$

# A Final Flourish

## The variance

- $\mathrm{Var}(X) = \sum_p \mathrm{Var}(X_p) + \sum_{p\neq q} \mathrm{Cov}(X_p, X_q)$
  - $\sum_p \mathrm{Var}(X_p) \leq \mathbb{E}[X]$ and $\sum_{p\neq q} \mathrm{Cov}(X_p, X_q) = o(\mathbb{E}[X])$
- $\Rightarrow \mathrm{Var}(X) = (1 + o(1))\mathbb{E}[X] = (1 + o(1)) \ln\ln n$

## Applying Chebyshev

- $\mathbb{P}\left(|\nu(x) - \ln\ln n| > \lambda\sqrt{\ln\ln n}\right) \leq \frac{\mathrm{Var}(X)}{\lambda^2 \ln\ln n} = \frac{1}{\lambda^2} + o(1)$
- $\Rightarrow \mathbb{P}\left(\nu(x) \neq (1 + o(1)) \ln\ln n\right) = o(1)$
- $x$ uniform in $[n] \Rightarrow o(n)$ such integers ■

## Remark

- Most $x \in [n]$ satisfy $\nu(x) = \ln\ln n + O\left(\sqrt{\ln\ln n}\right)$

# Any questions?

# §5 Distinct Sums

Chapter 3: The Second Moment

The Probabilistic Method

# Mathemagic

## An illusion

- You have a deck of cards, with each card bearing a number
- You invite your friend to select as many cards from the deck as they like
- They add the numbers and only tell you the sum
- The chosen cards are then shuffled back into the deck
- You then go through the deck, and magically pick out your friend's cards

## The secret

- Cards labelled with powers of two: $1,2,4,8,16, \ldots$
- Each number $x \in \mathbb{N}$ has a unique binary expansion, $x = \sum_j 2^{i_j}$
- $\Rightarrow$ given the sum $x$, can recover the labels $2^{i_j}$ of the chosen cards

# A Little Showmanship

## Obstacles

- Mathematician friends will see through the illusion
- Non-mathematician friends may not be able to add well
  - Card labels shouldn't be larger than $n$
- Binary labels $\Rightarrow \log n$ cards
  - Small deck is not so impressive

## Better decks

- Can we replace the binary labels?
- Suppose we have labels $S = \{s_1, s_2, \dots, s_k\}$
- Key property:
  - *distinct sums* – no two subsets should have the same total
- Extremal problem
  - How large can a subset $S \subseteq [n]$ with distinct sums be?

# The Greedy Magician

## Greedy algorithm

- Start with $S = \emptyset$
- Go through elements in $[n]$ one at a time
- Add to $S$ if they preserve distinct sums property

## Claim 3.5.1

The greedy algorithm returns the set of powers of two.

## Proof

- After the first step, we have $S = \{1\}$
- Suppose we have $S = \{1, 2, \ldots, 2^r\}$ at some stage in the algorithm
- We can write every number up to $2^{r+1} - 1$ as a sum of these elements
  - None of these added to $S$
- Next available number to be added: $2^{r+1}$ ■

# The Extremal Function

## Notation

- Let $f(n) = \max\{|S| : S \subseteq [n] \text{ has distinct sums}\}$

## Lower bound

- Binary set $\Rightarrow f(n) \geq \lfloor \log n \rfloor + 1$
- Is this best possible?

## Counterexamples

- $S = \{11,17,20,22,23,24\}$ has distinct sums
  - $\Rightarrow f(n) \geq \lfloor \log n \rfloor + 2$ for $24 \leq n \leq 31$
- If a set $S$ has distinct sums, so does $S' = 2S \cup \{1\}$
  - Iterating → infinite sequence of counterexamples

# An Upper Bound

**Proposition 3.5.2**

As $n \to \infty$, we have $f(n) \leq \log n + \log\log n + 1$.

**Proof**

- Let $k = f(n)$ and let $S \subseteq [n]$ be a largest set with distinct sums
- For each $T \subseteq S$, we have $0 \leq \sum_{s \in T} s < kn$
- Distinct sums $\Rightarrow$ each of these $2^k$ sums is distinct
- $\Rightarrow 2^{\mathrm{k}} \leq kn$

$$
\begin{aligned}
&\Rightarrow k \leq \log n + \log k \\
&\Rightarrow k \leq \log n + \log(\log n + \log k) \\
&\qquad \leq \log n + \log(2 \log n) \\
&\qquad = \log n + \log\log n + 1
\end{aligned}
$$

∎

# An Improved Upper Bound

## Flawed argument

- Wasteful in estimating range of sums
- Max sum $\sim kn \Rightarrow$ all members of $S \sim n$
- In that case, few small numbers will be sums

## Fix

- Try to find a smaller interval still containing many sums
- Chebyshev $\Rightarrow$ sums may concentrate around the average

### Theorem 3.5.3

As $n \to \infty$, $f(n) \leq \log n + \frac{1}{2} \log \log n + O(1)$.

# Probabilistic Framework

## Random variables

- Let $f(n) = k$, let $S = \{s_1, s_2, \dots, s_k\} \subseteq [n]$ be a largest set with distinct sums
- Let $X$ be a uniformly random sum from $S$
- $\Rightarrow X = \sum_{i=1}^{k} \varepsilon_i s_i$, where each $\varepsilon_i$ is independent, uniform on $\{0,1\}$

## Expectation

- Let $\mu \coloneqq \mathbb{E}[X] = \sum_{i=1}^{k} \mathbb{E}[\varepsilon_i s_i] = \frac{1}{2}\sum_{i=1}^{k} s_i$
- Actual value is unimportant

## Variance

- Variables $\varepsilon_i$ are independent
- $\Rightarrow \mathrm{Var}(X) = \mathrm{Var}\left(\sum_{i=1}^{k} \varepsilon_i s_i\right) = \sum_{i=1}^{k} \mathrm{Var}(\varepsilon_i) s_i^2 = \frac{1}{4}\sum_{i=1}^{k} s_i^2 \leq \frac{1}{4} n^2 k$

# Concentrated Sums

## Recall

- $\mathrm{Var}(X) \le \frac{1}{4}n^2 k$

## Applying Chebyshev

- $\mathbb{P}\left(|X-\mu| \ge n\sqrt{k}\right) \le \frac{\mathrm{Var}(X)}{n^2 k} \le \frac{1}{4}$
- $\Rightarrow \mathbb{P}\left(|X-\mu| < n\sqrt{k}\right) \ge \frac{3}{4}$

## Distinct sums

- Each value comes from at most one sum $\Rightarrow \mathbb{P}(X=x) \in \{0, 2^{-k}\}$
- $\therefore \mathbb{P}\left(|X-\mu| < n\sqrt{k}\right) = \mathbb{P}\left(\mu - n\sqrt{k} < X < \mu + n\sqrt{k}\right) \le 2n\sqrt{k} \cdot 2^{-k}$

## Bounding $k$

- $2^k \le \frac{8}{3}n\sqrt{k} \Rightarrow k \le \log n + \frac{1}{2}\log k + \log\frac{8}{3} \le \log n + \frac{1}{2}\log\log n + O(1)$ $\blacksquare$

# Any questions?