15756 Randomized Algorithms

Rohan Jain

Contents

Chapter 1

100 Prisoners Problem

This is a problem based on a riddle:

Question

There is a jail and a warden. However, the warden is very lazy and says if they win a game, they're all set free, otherwise they all die. There is 100 boxes in a room and each box has a hidden number from 1 to 100. They are assorted in a random permutation. Each prisoner is also given a number from 1 to 100. Each prisoner is allowed to enter the room and will be allowed to open 50 boxes and if they find their number, they win. They will only live if everyone finders their number. The prisoners can't communicate with each other. What strategy should they use to maximize their chances of winning?

Naive strategy: Each prisoner opens 50 boxes at random. The probability of winning is $\dfrac{1}{2^{100}}$.

Theorem 1.0.1

There exists a strategy such that the probability of winning is $\geqslant \frac{30}{100}$.

Proof. The algorithm is as follows:

1. Each prisoner opens the box with their number.
2. If they find their number, they open the box with the number they found.
3. They continue this process until they find their number or they open 50 boxes.

Cycle notation: Consider the following arrangement, where the ordinal is the box number:

1. 7
2. 6
3. 4
4. 3
5. 8
6. 1
7. 2
8. 0
9. 5
10. 10

This can be represented as 4 individual cycles: $(1, 7, 2, 6)$, $(3, 4)$, $(5, 8, 9)$, (10).

A critical observation to make is that prisoner i will win if and only if the cycle containing i is of length $\leqslant 50$. That is, everyone wins iff there are no cycles of length > 50.

We warm up by considering opening 75 boxes instead of 50. The observation is that if anyone loses, at least 75 people will lose. Let X be a random variable representing the number of people who lose. Then, $\mathrm{E}[X] = 25$ as each player has a $\frac{1}{4}$ chance of losing. Recall Markov's inequality:

Theorem 1.0.2

Let X be a non-negative random variable. Then, for any $t > 0$, $\Pr[X \geqslant t] \leqslant \frac{\mathrm{E}[X]}{t}$.

Applying Markov's inequality, we have $\Pr[X \geqslant 75] \leqslant \frac{25}{75} = \frac{1}{3}$. Thus, the probability of winning is $\geqslant \frac{2}{3}$. Moving back to the original problem, we can apply the same logic.

$$\Pr(\text{anyone loses}) = \Pr(\text{at least one cycle of length} > 50).$$

Recall how we can count the number of cycles of length $\geqslant 50$. For a cycle to be exactly length $n > \frac{100}{n} = 50$, we need to choose n boxes and then permute them. Thus, the number of cycles of length n is $\frac{100!}{n \cdot (100-n)!}$. Now we have to worry about the number of arrangements of the other $100 - n$ boxes, which is $(100 - n)!$. Multiplying these two quantities, we get:

$$\frac{100!}{n \cdot (100-n)!} \cdot (100-n)! = \frac{100!}{n}.$$

As there are 100! total permutations, we realize that exactly $\frac{1}{n}$ of the permutations contain a cycle of length n. Thus, the probability of losing is:

$$\Pr(\text{anyone loses}) = \sum_{n=51}^{100} \frac{1}{n} = \sum_{n=1}^{100} \frac{1}{n} - \sum_{n=1}^{50} \frac{1}{n} = H_{100} - H_{50} \approx \ln 100 - \ln 50 = \ln 2 \approx 0.693.$$

Thus, the probability of winning is $\geqslant 1 - 0.693 \approx 0.307$ as desired.

☺

We move on to proving that this algorithm is actually optimal:

Theorem 1.0.3

The algorithm described above is optimal.

Proof. We consider a second version of the game, where any box that a previous prisoner has opened stays open. So if prisoner i walks in to see that i has been revealed, he just leaves. Otherwise, he opens boxes until he finds i or has opened 50 boxes. Boxes are never closed.

Lenma 1.0.1

Cycle algorithm is no better at version 2 than version 1.

Proof. Option 1: Someone from my cycle in the past lost.

Option 2: I am the first person in my cycle to enter.

In both versions, the boxes in my cycle are closed. This means that in both versions, I win iff the cycle is of length $\leqslant 50$.

Option 3: Someone from my cycle has already entered, and they won. This means the cycle is of length $\leqslant 50$, so I win in both versions. ☺

Lenma 1.0.2

All algorithms are equally good in version 2.

Proof. By symmetry: If I'm opening a box, I have to pick out of the remaining boxes, and they all have the same probability of containing my number. ☺

Together, these lemmas show that the cycle algorithm is optimal in version 1. ☺

Chapter 2

Graph algorithms

*Note: all graphs are going to be undirected and unweighted.

2.1 Graph coloring

Suppose we are given a graph $G = (V, E)$. Suppose G is 3-colorable. Recall that a 3-coloring is a function $c : V \longrightarrow \{1, 2, 3\}$ such that $c(u) \neq c(v)$ for all $(u, v) \in E$. Simply put, you can color the vertices such that no two adjacent vertices have the same color.

Question

Can we efficiently find a 3-coloring of G?

Answer: No, this is NP-hard.
Observation: It's possible to partition the vertices in the graph into two parts that are each $\triangle$-free. This is easy because we can create the partition of one set being the odd vertices (1 and 3) and the other set being the even vertices (2).

> **Theorem 2.1.1** McDiarmid (1993)
>
> Given a 3-colorable graph, you can construct a triangle-free partition of the vertices in poly(n) expected time.

Proof. Consider the following algorithm: in each round, find a triangle in some side, pick a random vertex from the triangle, and move it to the other side. Repeat this process until there are no more triangles.

Now we analyze the algorithm. Fix a 3-coloring of G. Let X be the number of vertices of color 1 on the left side. Let Y be the number of vertices of color 2 on the right side. Now we analyze what happens to $X + Y$ during a round:

- If we move a vertex from the left to the right, X can decrease by 1 with probability $\frac{1}{3}$, Y can increase by 1 with probability $\frac{1}{3}$, or both can stay the same with probability $\frac{1}{3}$.
- If we move a vertex from the right to the left, Y can decraese by 1, X can increase by 1, or both can stay the same all with probability $\frac{1}{3}$.

So overall, $X + Y$ increases by 1 with probability $\frac{1}{3}$, decreases by 1 with probability $\frac{1}{3}$, and stays the same with probability $\frac{1}{3}$. This is just a random walk.

The random walk starts at an unknown value, but if it ever leaves $[0, n]$, the algorithm terminates. It can look something like $+1, -1, 0, 0, -1, +1, +1, +1$, and at some point t_1 the algorithm finishes. For sake of discussion, assume this hypothetical random walk keeps going. At some point, the walk may leave $[0, n]$ at time t_2. Note that t_2 can exist only if t_1 exists, and $t_1 < t_2$. So $\mathrm{E}[t_2] \geqslant \mathrm{E}[t_1]$.

Goal: prove that $\mathrm{E}[t_2] \in O(n^2)$. This will imply that $\mathrm{E}[t_1] \in O(n^2)$, which will imply that the algorithm runs in poly(n) time.

We step away from the graph problem to analyze random walks further. Consider a random walk on $\mathbb{Z}$ that starts at 0. Let T be the first time the walk arrives at n or $-n$. We want to show that $\mathrm{E}[T] \in O(n^2)$.

Fact: after t steps, the middle 90% of possible positions would be $O(\sqrt{t})$ from the origin. Additionally, within this 90%, the walk is about uniformly distributed.

Using the fact above, we want $t = n^2$ to be the time it takes to reach n or $-n$.

Now we return to the original problem.

> **Lenma 2.1.1**
>
> For any power of 2, the expected time to get from the origin to n or $-n$ is n^2.

Proof. If $n = 1$, the expected time is 1. Define T_n as the expected time to get from the origin to n or $-n$. Pretend we split the number line into $-n, -n/2, 0, n/2, n$. Then we can write T_n as:

$$T(n) = T(n/2) + T(n/2) + \frac{1}{2}T(n).$$

Solving, we get $T(n) = 4T(n/2) \implies T(n) = n^2$. ☺

This means that our algorithm is $O(n^2)$ rounds, meaning the algorithm runs in $\mathrm{poly}(n)$ time. ☺

2.2 Shortest paths

Given $G = (V, E)$, for each $x, y \in V$, we want $d(x, y) \pm O(1)$ in time $O(n^{2.5} \log n)$.

Naive algorithm: $O(|V| \cdot |E|) = O(n^3)$.

Now consider a shortest path from x to y.

- Option 1: Every vertex on the path has degree $< \sqrt{n}$.

  Then for each vertex, we can BFS, but only on low degree vertices. This takes $O(n^{2.5})$ time. The result is that we get exact distances for all vertices on the path.

- Option 2: There is a vertex on the path with degree $\geqslant \sqrt{n}$.

  Now we sample $100\sqrt{n}\log n$ iid uniformly random vertices and call this set S. We claim that with high probability, every high degree vertex w will be incident to a vertex $s \in S$.

  If we take this claim for granted, observe that for x, y in option 2, $d(x, y) = d(x, s) + d(s, y) \pm 2$ for some $s \in S$. So for each $s \in S$, do a BFS which takes $O(n^{2.5} \log n)$ time. For each pair (x, y), compute the minimum $d(x, s) + d(s, y)$ over all $s \in S$. This takes $O(n^{2.5} \log n)$ time.

Now we prove the claim made in option 2.

Proof. Let w be a high degree vertex. We want to find $P(\text{none of } w\text{'s neighbors are sampled})$.

$$\begin{aligned} P(\text{none of } w\text{'s neighbors are sampled}) &\leqslant P(\text{given sample misses})^{|S|} \\ &= \left(1 - \frac{1}{\sqrt{n}}\right)^{100\sqrt{n}\log n}. \end{aligned}$$

Now we use the fact that $\left(1 - \frac{1}{n}\right)^n \leqslant \frac{1}{e}$. So now,

$$\left(1 - \frac{1}{\sqrt{n}}\right)^{100\sqrt{n}\log n} \leqslant \frac{1}{e^{100\log n}} = \frac{1}{n^{100}}.$$

This is the probability of w screwing up. The probability of any high degree vertex screwing up is $\leqslant \frac{n}{n^{100}} = \frac{1}{n^{99}}$ by union bound. ☺

2.3 Global Min-Cut

We start by stating the question we are trying to solve.

Question: Global Min-Cut Problem

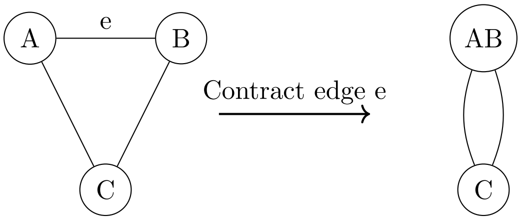
We are given an unweighted and undirected graph $G = (V, E)$. A "cut" of the graph is a partition of the vertices into two sets such that each set is non-empty. The "cost" of the cut (also called the cut "size") is the number of edges that have one endpoint in each set. The goal is to find the cut with the smallest cost.

Let $n := |V|$. We consider two randomized algorithms:

- Karger's Algorithm (1995) - poly(n)-time.
- Karger-Stein Algorithm (1996) - $\tilde{O}(n^2)$ time.

2.3.1 Edge Contraction

The key idea in both algorithms is edge contraction. Given an edge (u, v), we can contract it by removing the edge and merging the two vertices into a single vertex. The new vertex will have all the edges that u and v had. For example:



Each contraction reduces the number of vertices by 1, and we want to stop when we have two vertices left. This leaves us with two vertices and many edges between them.

Note:
We allow multi-edges.

Question: Should we keep self-edges?
Answer: No. Delete them.
The observation we make is that we win iff edges at the end are a min-cut.

2.3.2 Karger's Algorithm

The algorithm is as follows:

- Randomly pick an edge (u, v).
- Contract (u, v).
- Repeat until 2 vertices are left.

Lenma 2.3.1
For any min-cut C, the probability that Karger's algorithm finds C is $\geqslant \frac{1}{\text{poly}(n)}$.

Proof. Let $e_1, e_2, \ldots$ be the edges we contract. Let's start by considering $P(e_1 \in C)$.

We first have $P(e_1 \in C) = \frac{|C|}{|E|}$, but there's something better to notice. Let d be the minimum degree of any vertex in G. Then $|C| \leqslant d$ and $|E| \geqslant \frac{dn}{2}$. So $P(e_1 \in C) \leqslant \frac{2}{n}$ and $P(e_1 \notin C) \geqslant 1 - \frac{2}{n}$.

Now consider $P(e_2 \notin C \mid e_1 \notin C)$. By the same analysis, we have $P(e_2 \notin C \mid e_1 \notin C) \geqslant 1 - \frac{2}{n-1}$. This is because after we contract e_1, we have $n-1$ vertices left. So, C is still a min-cut with respect to the remaining vertices. Set d as the new minimum degree, and we have $|C| \leqslant d$ and $|\text{remaining edges}| \geqslant \frac{d(n-1)}{2}$. Then:

$$P(e_2 \in C \mid e_1 \notin C) = \frac{|C|}{|\text{remaining edges}|} \leqslant \frac{d}{\frac{d(n-1)}{2}} = \frac{2}{n-1}.$$

Now:

$$\begin{aligned} P(e_3 \notin C | e_1, e_2 \notin C) &\geqslant 1 - \frac{2}{n-2} \\ &\vdots \\ P(e_k \notin C | e_1, \ldots, e_{k-1} \notin C) &\geqslant 1 - \frac{2}{n-k+1}. \end{aligned}$$

So,

$$\begin{aligned} P(\text{we find } C) = P(e_1, \ldots, e_{n-2} \notin C) &\geqslant \left(\frac{n-2}{n}\right)\left(\frac{n-3}{n-1}\right)\cdots\left(\frac{1}{3}\right) \\ &\approx \frac{2}{n(n-1)}. \end{aligned}$$

☺

Conclusion: The algorithm succeeds with probability $\geqslant \frac{2}{n(n-1)}$.
New Goal: Find min-cut with $\geqslant 1 - \frac{1}{n^2}$ probability.
Idea: Repeat t times and return the best cut we find. (We will determine the value for t below.)

The probability that all t trials fail is $\leqslant \left(1 - \frac{2}{n(n-1)}\right)^t$. We want this to be $\leqslant \frac{1}{n^2}$. We utilize the inequality below:

$$\left(1 - \frac{1}{k}\right)^k \leqslant \frac{1}{e} \leqslant \left(1 - \frac{1}{k}\right)^{k-1}.$$

So,

$$\begin{aligned} \left(1 - \frac{2}{n(n-1)}\right)^t &\leqslant \frac{1}{e^{t \cdot \frac{2}{n(n-1)}}} \implies t = 2\binom{n}{2}\log n \\ &\leqslant \frac{1}{e^{2\log n}} = \frac{1}{n^2}, \end{aligned}$$

as desired.

2.3.3 Karger-Stein Algorithm

We start with the following motivation. We have:

$$P(\text{first } n/2 \text{ contractions succeed}) \geqslant \frac{n-2}{n}\frac{n-3}{n-1}\cdots\frac{n/2-1}{n/2+1} \approx \frac{1}{4}.$$

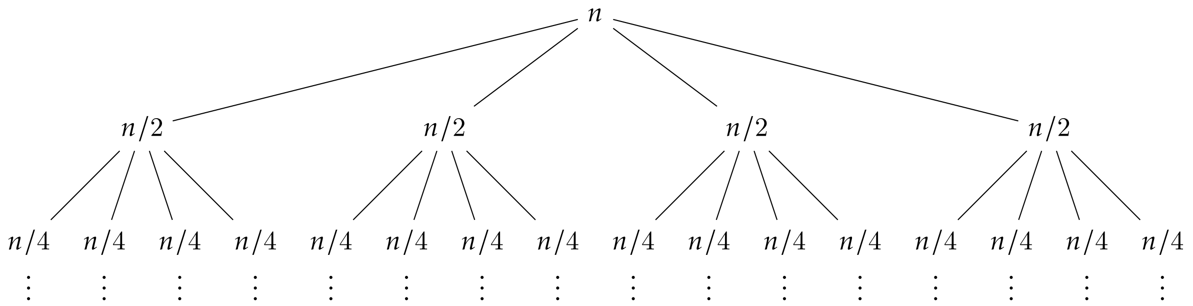
This is because after we telescope, the bottom two terms are both near n while the top two terms are near $n/2$. So now, if the first $n/2$ contractions succeed, what's the probability that the next $n/4$ succeed?

$$P(\text{next } n/4 \text{ contractions succeed}) \geqslant \frac{n/2-2}{n/2}\frac{n/2-3}{n/2-1}\cdots\frac{n/4-1}{n/4+1} \approx \frac{1}{4}.$$

And this pattern continues.

We now go over the Karger-Stein algorithm:

- Represent the problem as a tree.
- For the first $n/2$ contractions, run Karger's algorithm 4 times.
- For the next $n/4$ contractions, run Karger's algorithm 16 times.
- So on, until the base case of the algorithm which is when there are two vertices left.


n
$n/2$
$n/2$
$n/2$
$n/2$
$n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$ $n/4$

Time Analysis: The first level takes $\tilde{O}(n^2)$ time. The second level takes $\tilde{O}\left(\frac{n^2}{2}\right)$ time. The third level takes $\tilde{O}\left(\frac{n^2}{4}\right)$ time. So the total time is $\tilde{O}(n^2)$.
So, at level i:

- There are 4^i edges.
- The work per edge $\leqslant$ $(|\text{remaining vertices}|)^2 = \tilde{O}\left(\frac{n}{2^i}\right)^2 = \tilde{O}\left(\frac{n^2}{4^i}\right)$.

This implies that the cost of the ith level is less than $\tilde{O}(n^2)$. Summing, we get a total cost $\tilde{O}(n^2)$.

Now we consider the probability that our algorithm succeeds. Again, we have a 4-ary tree with a depth of $t = O(\log n)$. Each edge flips a coin where $P(\text{heads}) = \frac{1}{4}$. So what we want is a t length chain of all heads. We can show the following result:

$$P(\text{good path}) \geqslant \frac{1}{t} \geqslant \Omega\left(\frac{1}{\log n}\right).$$

The goal is an algorithm that succeeds with probability greater than $1 - \frac{1}{n^2}$.
Exercise: $O(\log^2 n)$ repetitions suffice.

2.4 Galton Walton Tree

Building on the previous problem, we introduce the concept of a Galton Walton tree. This is a 4-ary tree with t levels where each edge flips a biased coin. The probability of heads is $1/4$. A root-to-leaf path in this tree is "successful" if all edges are heads.
Claim: $P(\text{successful path}) \geqslant \frac{1}{t}$.

Proof. Pretend the paths are independent. Then we have that

$$P(\text{No successful path}) = \left(1 - \frac{1}{4^t}\right)^{4^t} \leqslant \frac{1}{e}.$$

This is not a very interesting result. It is worth noting that the paths are extremely dependent on each other. Consider a path of length $t-2$ and creating two separate paths by diverging at the last edge. This should illustrate the dependence of the paths.

Now let f_t the probability of tailing in a tree with t levels. We can define this as a recurrence on f_{t-1}. We have:

$$f_t = \left(\frac{1}{4}f_{t-1} + \frac{3}{4}\right)^4.$$

To proceed with induction on t, we first prove the following identity:

$$1 - 1/k < \frac{1}{e^{1/k}} < 1 - 1/(k+1).$$

This is a useful identity to have in our back pocket. We have:

$$\begin{aligned}
&(1-1/k)^k < 1/e < (1-1/k)^{k-1}\\
\implies &1 - 1/k < 1/e^{1/k}\\
&1/e^{1/(k-1)} < 1 - 1/k \implies 1/e^{1/k} < 1 - 1/(k+1).
\end{aligned}$$

Putting it all together, we get the desired result. We can now proceed with induction.

Our hypothesis: $f_{t-1} \leqslant 1 - \frac{1}{t-1}$. What we want to show is that $f_t \leqslant 1 - \frac{1}{t}$.

We have:

$$\begin{aligned}
f_t &\leqslant \left(\frac{3}{4} + \frac{1}{4}\left(1 - \frac{1}{t-1}\right)\right)^4\\
&= \left(1 - \frac{1}{4(t-1)}\right)^4\\
&\leqslant \left(\frac{1}{e^{\frac{1}{4(t-1)}}}\right)^4\\
&= \frac{1}{e^{1/(t-1)}} && \text{(by the above identity)}\\
&< 1 - \frac{1}{t}. && \text{(by the above identity)}
\end{aligned}$$

This completes the induction, as well as the entire proof. ☺

Chapter 3

Lovasz Local Lemma

Lemma 3.0.1 Lovasz Local Lemma

Let $A_1, A_2, \ldots, A_n$ be "bad events." Suppose:

- $P(A_i) < p$ for some p.
- Each A_i has a "dependency set" $D_i \subseteq \{A_1, A_2, \ldots, A_n\}$ such that A_i is independent of $\{A_1, A_2, \ldots, A_n\} \setminus D_i$ and $|D_i| \leqslant d$.

If $p < \frac{1}{ed}$, then there exists an outcome where none of the A_i occur. That is, $P(\text{no bad events}) \geqslant \frac{1}{d^n}$.

Application: KSAT. In KSAT, we have a bunch of boolean variables $x_1, x_2, \ldots$. A K-SAT formula is a conjunction of n clauses, where each clause selects k unique variables that cannot take a specific form. For example, if we have $k = 3$, we might have a clause that says $(x_1, x_2, x_3) \neq (0, 1, 0)$. The question is whether we can find a satisfying assignment of these variables.

Theorem 3.0.1

Suppose each clause overlaps in variables $\leqslant d - 1$ other clauses. Also suppose $\frac{1}{2^k} < \frac{1}{ed}$. Then, there exists a satisfying assignment.

Proof. $x_1, x_2, \ldots$ are random $0, 1$ variables. Let A_i be the event that clause i is not satisfied. Then $P(A_i) \leqslant \frac{1}{2^k}$. By the LLL, there exists a satisfying assignment. 😄

3.1 Algorithmic Lovasz Local Lemma

3.1.1 Algorithmic Setting

We have random variables $X_1, X_2, \ldots, X_n$ and events A_i that are determined by some subset of the X_is. We can think about this as a dependency graph where all the X_i and A_i are nodes, and edges are drawn between X_i and A_j if X_i determines A_j.

Lemma 3.1.1 Algorithmic Lovasz Local Lemma

Suppose each A_i overlaps with at most $d - 1$ other A_js on which X_is they use. Suppose $P(A_i) < \frac{1}{4d}$. Then there exists a polynomial time algorithm* that finds an outcome of X_is such that none of the A_is occur.

3.1.2 Moser's Fix-it Algorithm

Algorithm:

- Sample $X_1, X_2, \ldots, X_m$.
- While $\exists$ bad event A_i:
  - Pick A_i arbitrarily.
  - Resample the X_i's that determine A_i.

> **Theorem 3.1.1** Moser's Theorem (KSAT)
>
> Consider KSAT:
>
> - m vars, $X_1, \ldots, X_m$.
> - n clauses $C_1, \ldots, C_n$.
> - Each C_i depends on $\leqslant d$ other C_i's.
>
> If $8d < 2^k$, then Moser's algorithm performs $t = O(n)$ (really $O\left(\frac{n}{d}\right)$) fix-its in expectation.

Proof Idea: Use random bits, for example $R = 0010010100101\ldots$. Construct a "transcript" T such that if you tell me T, I can recover all the random bits you used. The interesting property of T is that:

$$|T| \leqslant (\text{number of fix-its}) \cdot (2 \lg d) + O(n).$$

The number of random bits we use is less than the number of fix-its times k. We know that $k > (2 + \lg d) \cdot 100$.

If Moser's algorithm runs for a long time, say $> 100n$ fix-its, then

$$|T| \leqslant 100n \cdot (2 \lg d) + O(n).$$

Then,

$$\begin{aligned} \text{number of random bits} &= 100n \cdot k \\ &> 100n \cdot (2 + \lg d) \cdot 100. \end{aligned}$$

This means $|T| << $ number of random bits. So if we can show that such a T exists, then we can show that the expected number of fix-its to be pretty small.

3.1.3 Proof of Moser's Theorem

Proof. We start by defining variables we are going to use:

- R = random bits used by the algorithm.
- M_1 = final values of X_i's.
- M_2 = sequence of which clauses we fix.

> **Lenma 3.1.2**
>
> M_1 and M_2 fully determine R.

Proof. Work backwards! Consider the last C_i that was fixed. We know its variables after the fix and the variables before the fix. So we can "undo" the fix. We can keep doing this until we get to the beginning. ☺

Lemma 3.1.3

M_2 can be compressed to $n + t(2 + \lg d)$ bits.

We'll proceed by assuming the lemma first. So now we have $|M_1| = m$ and $|R| = m + tk$. Also since $8d < 2^k$¡ we have $3 + \lg d \leqslant k$. So, $|R| \geqslant m + t(3 + \lg d)$.

Now,

$$\begin{aligned} |M_1| + |M_2| &= n + m + t(2 + \lg d) \\ &\leqslant |R| + n - t. \end{aligned}$$

If $E[t] > n$, then $E[|M_1 \circ M_2|] < E[|R|]$. This is a contradiction by the incompressibility of random bits. So, $E[t] \leqslant n$ as desired. ☺

Now we return to proving the second lemma.

Proof. We start by defining processes.
Define process(C_i):

```
• Resample $C_i$'s variables.
• While ∃ another $C_j$ that is broken and depends on $C_i$:
    – Select such a $C_j$.
    – process($C_j$).
```

Define Algorithm:

```
• For $i = 1, \dots, n$:
    – If $C_i$ is broken, process($C_i$).
```

Claim: When the algorithm finishes, all clauses holds.
Claim: Algorithm terminates with probability 1.
We now look towards constructing M_2:

- A = Part 1: n bits, which C_i's get fixed in for loop.

Now recall the process algorithm. Notice here that for each C_j, there are only $d + 1$ options. This means that we can communicate C_j with $\lg(d + 1)$ bits, which we'll call q. We now rewrite process considering a string B.
Define process(C_i):

```
• Resample $C_i$'s variables.
• While ∃ another $C_j$ that is broken and depends on $C_i$:
    – Select such a $C_j$.
    – $B = B +$ "1" $+ q$.
    – process($C_j$).
• $B = B +$ "0".
```

So given M_2, A lets us recover which clauses the for loop fixes. B lets us figure out which clauses are fixed within each process subroutine. This means we can recover the entire sequence of clause fixes that the algorithm performs.

So, $|A| = n$ and $|B| = t(2 + \lg d)$. This means $|A \circ B| = n + t(2 + \lg d)$, which completes the proof. ☺

Chapter 4

Chernoff Bounds

We start with an example. Suppose I flip 100 fair coins. What is the probability that the number of heads is greater than or equal to 75? Turns out,

$$P(\text{heads} \geqslant 75) \approx \frac{1}{3.5 \times 10^6}.$$

Theorem 4.0.1

Consider n fair coin flips, $X_1, X_2, \ldots, X_n$ iid with $X_i = \begin{cases} 0 & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/2 \end{cases}$. Let $X = \sum X_i$. Then,

$$P\left(X > \frac{n}{2} + K\sqrt{n}\right) \leqslant e^{-\Omega(K^2)}.$$

Corollary 4.0.1

$$P\left(X \geqslant \frac{3}{4}n\right) \leqslant e^{-n/32}$$

You don't actually need $X_1, \ldots, X_n$ to be independent. It suffices to have that, for any i and any outcomes of $X_1, \ldots, X_{i-1}$,

$$P(X_i = 1 \mid X_1, \ldots, X_{i-1}) \leqslant 1/2.$$

Note:

"Azuma's Inequality" is a citable Chernoff bound. Alternatives include "multiplicative version of Azuma's inequality" and "Freedman's inequality."

We turn to the proof of the theorem.

Proof. Idea: suppose we have a function f such that:

- f is non-negative.
- f is increasing on $[n/2, n]$.

Then,

$$P\left(x \geqslant \frac{n}{2} + K\sqrt{n}\right) \leqslant P\left(f(x) \geqslant f\left(\frac{n}{2} + K\sqrt{n}\right)\right) \leqslant \frac{E[f(x)]}{f\left(\frac{n}{2} + K\sqrt{n}\right)}.$$

Example 4.0.1 (Trying f's)

$f(x) = \left(x - \frac{n}{2}\right)^2$. This clearly satisfies the requirements. This tells us that:

$$\begin{aligned} P\left(x > \frac{n}{2} + K\sqrt{n}\right) &\leqslant \frac{E[f(x)]}{K^2 n} \\ &= \frac{\operatorname{Var}(X)}{K^2 n}. \end{aligned}$$

We use the linearity of variance. We have $\operatorname{Var}(X_i) = \frac{1}{4}$. So, $\operatorname{Var}(X) = \frac{n}{4}$. Thus,

$$P\left(x > \frac{n}{2} + K\sqrt{n}\right) \leqslant \frac{1}{4k^2}.$$

This is exactly Chebyshev's inequality.

We now find an f that proves our result. Try $f(x) = e^{\lambda x}$ for some $\lambda \in (0, 1)$. This is approximately $(1+\lambda)^x$. We first find the expectation:

$$\begin{aligned} E[f(x)] = E[e^{\lambda \sum_i X_i}] &= E\left[\prod_i e^{\lambda X_i}\right] \\ &= \prod_i E[e^{\lambda X_i}] && \text{(by independence)} \\ &= \left(\frac{1+e^\lambda}{2}\right)^n. \end{aligned}$$

Fact:

$$\begin{aligned} e^\lambda &= 1 + \lambda + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots && (4.1) \\ &\leqslant 1 + \lambda + \lambda^2 \cdot \left(\frac{1}{2!} + \frac{1}{3!} + \cdots\right) && (4.2) \\ &\leqslant 1 + \lambda + \lambda^2. && (4.3) \end{aligned}$$

It follows that

$$E[e^{\lambda X_i}] \leqslant 1 + \frac{1}{2}\lambda + \frac{1}{2}\lambda^2.$$

Now we make use of the other fact that $1 + x \leqslant e^x$. So

$$\begin{aligned} E[e^{\lambda X_i}] &\leqslant 1 + \frac{1}{2}\lambda + \frac{1}{2}\lambda^2 \\ &\leqslant e^{\lambda/2 + \lambda^2/2}. \end{aligned}$$

Thus,

$$E[f(x)] \leqslant e^{n(\lambda/2 + \lambda^2/2)}.$$

So now,

$$\begin{aligned} P\left(X > \frac{n}{2} + K\sqrt{n}\right) &\leqslant \frac{E[f(x)]}{f\left(\frac{n}{2} + K\sqrt{n}\right)} \\ &\leqslant \frac{e^{n(\lambda/2 + \lambda^2/2)}}{e^{n\lambda/2 + \lambda K\sqrt{n}}} \\ &= e^{n\lambda^2/2 - \lambda K\sqrt{n}}. \end{aligned}$$

Now we try to find an appropriate value for λ. We try:

$$\lambda K\sqrt{n} = n\lambda^2$$
$$\frac{K}{\sqrt{n}} = \lambda.$$

This yields:

$$e^{n\lambda^2/2-\lambda K\sqrt{n}} = e^{-\lambda K\sqrt{n}/2} = e^{-k^2/2},$$

as desired. ☺

4.1 Quicksort Algorithm (1961)

We review the algorithm for quicksort:

- Pick random pivot P.
- Partition elements into 2 pieces: elements smaller/larger than P.
- Recurse on the pieces.

Theorem 4.1.1

The running time of quicksort satisfies $T = O(n \lg n)$ with probably at least $1 - \frac{1}{n^2}$.

Proof. Note that the time is $\leqslant \#\text{levels} \cdot O(n)$. So what we have to prove is that with high probability, the depth of the algorithm doesn't exceed $\lg n$. So,

$$\begin{aligned} P(\text{max depth} \geqslant 1000 \lg n) &\leqslant P(\exists \text{element with max depth} \geqslant 1000 \lg n) \\ &\leqslant n \cdot P(\text{a given } v \text{ has depth} \geqslant 1000 \lg n). \end{aligned}$$

We want to show that the probability on the last line is $\leqslant \frac{1}{n^3}$. We have that

$$P(\text{next subproblem} \leqslant 3/4\text{current size}) \geqslant \frac{1}{2}.$$

Let $P_1, P_2, \ldots$ be subproblems in v's path. Let

$$X_i = \begin{cases} 1 & P_{i+1}\text{exists and } |P_{i+1}| > \frac{3}{4}|P_i| \\ 0 & \text{otherwise} \end{cases}$$

We observe that $P(X_i = 1 \mid X_1, \ldots, X_{i-1}) \leqslant 1/2$. So by Chernoff bound, if we look at $X = X_1 + X_2 + \cdots + X_{1000 \lg n =: m}$, we get:

$$P\left(X \geqslant \frac{3}{4}m\right) \leqslant e^{-m/32} <<< \frac{1}{n^3}.$$

If $X \leqslant \frac{3}{4}m$, either P_m doesn't exist or $|P_m| \leqslant n \cdot \left(\frac{3}{4}\right)^{m/4} = n \cdot \left(\frac{3}{4}\right)^{250 \log n} << 1$, which is a contradiction. This means P_m in fact does not exist. As such, we are able to conclude that we have a running time of $O(n \lg n)$ with the desired probability. ☺

4.2 Revisiting Theorem 4.0.1

We start with warm-ups to give an alternative proof to Theorem 4.0.1.

4.2.1 Poor Man's Chernoff Bound

Theorem 4.2.1

$$P(X \geqslant 2K\sqrt{n}) \leqslant \frac{1}{2^K}.$$

Proof. We make use of the core facts below:
Core fact 1: $P(X \geqslant 2\sqrt{n}) \leqslant \frac{1}{4}$, by Chebyshev's inequality.
Core fact 1': $P(\text{any prefix has sum} \geqslant 2\sqrt{n}) \leqslant \frac{1}{2}$. This is a direct result of **Core fact 1**, and the implication will be shown in the homework.

We can graph the running sum as a function of time and track the times that we hit the values $2\sqrt{n}$, $4\sqrt{n}$, and $6\sqrt{n}$ and call them $t_1, t_2, t_3,$ et cetera. What the core fact is saying is that $P(t_1 \text{ exists}) \leqslant \frac{1}{2}$. Then, we can build the the following chain:

$$\begin{aligned} P(t_1 \text{ exists}) &\leqslant \frac{1}{2} \\ P(t_2 \text{ exists} \mid t_1 \text{ exists}) &\leqslant \frac{1}{2} \\ &\vdots \\ P(t_i \text{ exists} \mid t_{i-1} \text{ exists}) &\leqslant \frac{1}{2}. \end{aligned}$$

So,

$$P(X \geqslant 2K\sqrt{n}) \leqslant P(t_K \text{ exists}) \leqslant \frac{1}{2^K},$$

as desired. ☺

4.2.2 Chernoff Bound for Geometric Random Variables

Theorem 4.2.2
Let $X_1, \ldots, X_n$ be non-negative independent random variables such that for all $j \in \mathbb{N}$,

$$P(X_i \geqslant j) \leqslant p^j.$$

Then $X = \sum X_i$ satisfies

$$P(X \geqslant 2n) \leqslant (4p)^n.$$

Proof. $X \geqslant 2n \implies \sum \lfloor X_i \rfloor \geqslant n \implies \exists \langle Y_1, \ldots, Y_n \rangle =: \vec{Y}$ such that

- $\sum Y_i = n$.
- $X_i \geqslant Y_i \, \forall i$.

For a given option $\langle Y_1, \dots, Y_n \rangle$ for $\vec{Y}$,

$$\begin{aligned} P(\vec{Y} \text{ occurs}) &= P(X_i \geqslant Y_i \forall i) \\ &\leqslant \prod_i P(X_i \geqslant Y_i) \\ &\leqslant \prod_i p^{Y_i} \\ &= p^{\sum_i Y_i} \\ &= p^n. \end{aligned}$$

Now we bound the number of options for $\vec{Y}$. We can express $\vec{Y}$ as a binary string, for example:

$$\underbrace{000}_{Y_1=3}\ 1\ \underbrace{\ 0\ }_{Y_2=1}\ 1\ \underbrace{}_{Y_3=0}\ 1\dots$$

So the number of 0's is $\sum Y_i = n$ and the number of 1's is n. This means that the number o options satisfying $\sum Y_i = n$ is less than the number of binary strings of length $2n$ which is 4^n.

So,

$$\begin{aligned} P(X \geqslant 2n) \leqslant P(\text{some } \vec{Y} \text{ occurs}) \\ &\leqslant (\# \text{ of options for } \vec{Y}) \cdot P(\vec{Y} \text{ occurs}) \\ &\leqslant 4^n \cdot p^n \\ &= (4p)^n. \end{aligned}$$

😄

4.2.3 Anti-(Chernoff Bound)

Theorem 4.2.3

$$P\left(X \geqslant \frac{K}{4}\sqrt{n}\right) \geqslant \frac{1}{4^{K^2}}$$

Proof. Break the coins into K^2 groups. So each group has $m = \frac{n}{K^2}$ coins.

Core fact: $P\left(X \geqslant \frac{1}{4}\sqrt{n}\right) \geqslant \frac{1}{4}$.

Let C_i be the sum of the i-th group. So by the core fact,

$$P\left(C_i \geqslant \frac{1}{4}\sqrt{m}\right) \geqslant \frac{1}{4}.$$

So,

$$P\left(\text{every } C_i \geqslant \frac{1}{4}\sqrt{m}\right) \geqslant \frac{1}{4^{K^2}}$$

If this occurs, then

$$X = \sum C_i \geqslant K^2 \cdot \left(\frac{1}{4}\sqrt{m}\right) = \frac{K}{4}\sqrt{n}.$$

😄

4.2.4 Real Chernoff Bound

Theorem 4.2.4

$$P(X \geqslant 16k\sqrt{n}) \leqslant \frac{1}{2^{K^2}}.$$

Proof. We start by defining $Y_i = \max\left(0, \frac{C_i}{8\sqrt{m}}\right)$. We can apply the Poor Man's Chernoff bound to each C_i to get

$$\begin{aligned} P(C_i \geqslant 2\ell\sqrt{m}) &\leqslant \frac{1}{2^\ell} \\ \implies P\left(Y_i \geqslant \frac{2\ell\sqrt{m}}{8\sqrt{m}}\right) &\leqslant \frac{1}{2^\ell} \\ \implies P(Y_i \geqslant \ell) &\leqslant \frac{1}{2^{4\ell}} \leqslant \frac{1}{16^\ell}. \end{aligned}$$

This is saying that Y_i are geometric random variables. So by the geometric random variable bound,

$$P\left(\sum Y_i \geqslant 2K^2\right) \leqslant (4 \cdot 1/16)^{K^2} = \left(\frac{1}{4}\right)^{K^2}.$$

But also

$$\begin{aligned} X \geqslant 16K\sqrt{n} &\implies \sum C_i \geqslant 16K\sqrt{n} \\ &\implies \sum Y_i \geqslant \frac{16K\sqrt{n}}{8\sqrt{m}} = 2K^2. \end{aligned}$$

☺

Theorem 4.2.5

Let $X_1, X_2, \ldots, X_n$ be independent $\{0,1\}$ coin flips with $P(X_i = 1) = p$ for some $p \leqslant \frac{1}{2}$. Define $X = \sum X_i$, $\mu = E[X] = np$.

Small-deviation regime: For $K \in \{1, \ldots, \sqrt{\mu}\}$,

$$P(X \geqslant \mu + K\sqrt{\mu}) \leqslant e^{-\Theta(K^2)}.$$

Large-deviation regime: For $J \geqslant 2$ such that $J\mu \leqslant n$,

$$P(X \geqslant J\mu) \leqslant \frac{1}{J^{\Theta(J\mu)}}.$$

Proof. (large-deviation case)

Warmup 1

Theorem 1 (Poor Man's Chernoff Bound) If $\mu \leqslant 1$, then

$$P(X \geqslant K) \leqslant \mu^K.$$

Proof. **Core fact** $P(X \geqslant 1) \leqslant \mu$ by Markov's inequality.

Now repeat the ideas with $t_1, t_2, \ldots$ to say that $P(X \text{ reaches } i) = \mu^i$. ☺

Warmup 2 This was already done prior as it is the exact same.

Warmup 3

Theorem 3

$$P(X \geqslant J) \geqslant \frac{1}{J^{O(J\mu)}}$$

Proof. We'll have $J\mu$ groups where each group has mean $\frac{1}{J}$.
Core fact: If $\mu \leqslant 1$, then $P(X \geqslant 1) \geqslant \Omega(\mu)$.

By the core fact, each C_i is $\geqslant 1$ with probability $\geqslant \Omega\left(\frac{1}{J}\right)$. So,

$$P(\text{every } C_i \geqslant 1) = P(X \geqslant J\mu) \geqslant \Omega\left(\frac{1}{J}\right)^{J\mu}.$$

"Squiggly square to signify almost proof done." ☺

So the Poor Man's Chernoff Bound implies $P(C_i \geqslant K) \leqslant \frac{1}{J^K}$. Then by Chernoff for geometric random variables,

$$P\left(\sum C_i \geqslant 2J\mu\right) \leqslant \left(\frac{4}{J}\right)^{J\mu}$$

☺

4.3 More Chernoff Bounds

Theorem 4.3.1 Chernoff Bound

Let $X_1, X_2, \ldots, X_n \in [0,1]$ be independent random variables. Let $X = \sum X_i$ and $\mu = E[X]$. Then, for $k \in O(\sqrt{\mu})$,

$$P(|X - \mu| \geqslant k\sqrt{\mu}) \leqslant e^{-\Omega(k^2)}.$$

And for $J \geqslant 2$,

$$P(X \geqslant J\mu) \leqslant \left(\frac{1}{J}\right)^{\Omega(J\mu)}.$$

Small-deviation regime:

- There are k^2 chunks.
- So each has expected value $\approx \frac{\mu}{k^2}$.
- If each exceeds expected value by $\sqrt{\frac{\mu}{k^2}}$, then the total sum is $\geqslant k^2 \cdot \sqrt{\frac{\mu}{k^2}} = k\sqrt{\mu}$.

Large-deviation regime:

- $J\mu$ chunks.
- Each chunk has expected value $\approx \frac{1}{J}$.
- If each chunk is at least 1, then the total sum is at least $J\mu$.

Theorem 4.3.2 Bennett's Inequality, Bernstein's Bound

Same bound, but let $v = \mathrm{Var}(X)$.

$$P(X - E[X] \geqslant k\sqrt{v}) \leqslant e^{-\Theta(k)} \text{ for } k \in \{1, 2, \ldots, \sqrt{v}\}$$
$$P(X - E[X] \geqslant Jv) \leqslant \left(\frac{1}{J}\right)^{\Omega(Jv)} \text{ for } J \geqslant 2.$$

4.3.1 Adaptive Version

Theorem 4.3.3 Azuma's Inequality

Let $X_1, X_2, \ldots, X_n \in [-1, 1]$. Suppose $E[X_i|X_1, \ldots, X_{i-1}] = 0$. Then, $X = \sum X_i$ satisfies

$$P(X \geqslant K\sqrt{n}) \leqslant e^{-\Omega(K^2)}.$$

Fancier versions:

Have random variables $X_1, X_2, \ldots, X_n \in [0, 1]$ which are revealed one after the other. After $X_1, \ldots, X_{i-1}$ are revealed, Alice gets to select the distribution P_i for X_i and then X_i is revealed from P_i.

Let $\mu_i = E[X_i \mid P_i]$, $v_i = \mathrm{Var}(X_i \mid P_i)$, $X = \sum X_i$, $\mu = \sum \mu_i$ and $v = \sum v_i$. As a corollary to Azuma's inequality,

$$P(X \geqslant \mu + k\sqrt{n}) \leqslant e^{-\Omega(k^2)}.$$

Now if μ is deterministically at most $\bar{\mu}$, then

$$P(X \geqslant K\sqrt{\bar{\mu}} + \bar{\mu}) \leqslant e^{-\Omega(K^2)}$$
$$P(X \geqslant J\bar{\mu}) \leqslant \left(\frac{1}{J}\right)^{\Omega(J\bar{\mu})}.$$

Theorem 4.3.4 Freedman's Inequality

If v is deterministcally at most $\bar{v}$, then

$$P(X \geqslant K\sqrt{\bar{v}} + \mu) \leqslant e^{-\Omega(K^2)} \text{ for } K \in \{1, 2, \ldots, \sqrt{v}\}$$
$$P(X \geqslant \mu + J\bar{v}) \leqslant \left(\frac{1}{J}\right)^{\Omega(J\bar{v})} \text{ for } J \geqslant 2.$$

Note:

Bill has used the following bound in ~ 90% of his papers.
"Sometimes when you don't need this bound you want to find a way to use it anyway."

Theorem 4.3.5 McDiarmid's Inequality

Let $X_1, X_2, \ldots X_n$ be independent. Let $F : X_1, \ldots, X_n \longrightarrow \mathbb{R}$. Suppose: If I change some X_i to a new value $\bar{X}_i$,

$$|F(X_1, \ldots, X_i, \ldots, X_n) - F(X_1, \ldots, \bar{X}_i, \ldots, X_n)| \leqslant 1.$$

Then,

$$P(|F - E[F]| \geqslant K\sqrt{n}) \leqslant e^{-\Omega(K^2)}.$$

Example 4.3.1

Consider a random graph:

- n vertices.
- each (i, j) is an edge with probability $\frac{1}{2}$.

Consider the chromatic number $\chi(G)$. There's a way to show that

$$P(|\chi(G) - E[\chi(G)]| \geqslant k\sqrt{n}) \leqslant e^{-\Omega(k^2)}.$$

We cannot define the random variables X_i on whether an edge is included or not. This is because we end up getting n^2 random variables, which makes the inequality flop. We proceed by defining $X_1, X_2, \ldots, X_n$ where X_i is the number of edges from vertex i to vertices $> i$. Then the result follows directly from McDiarmid's inequality.

Proof Idea: Let

$$\begin{aligned} Y_0 &= E[F] \\ Y_1 &= E[F \mid X_1] \\ Y_2 &= E[F \mid X_1, X_2] \\ &\vdots \\ Y_n &= E[F \mid X_1, \ldots, X_n] = F. \end{aligned}$$

Then let

$$\begin{aligned} Z_1 &= Y_1 - Y_0 \\ Z_2 &= Y_2 - Y_1 \\ &\vdots \\ Z_n &= Y_n - Y_{n-1}. \end{aligned}$$

Claim 1: $E[Z_i \mid Z_1, \ldots, Z_{i-1}] = 0$.
Claim 2: $|Z_i| \leqslant 1$.

So by Azuma, $P(\sum Z_i > K\sqrt{n}) \leqslant e^{-\Omega(K^2)}$. But $\sum Z_i = Y_n - Y_0 = F - E[F]$.

Chapter 5

Oblivious Routing

We start with an n-node hypercube:

- n vertices: $1, \dots n$.
- edges: For each pair i, j if i, j differ in exactly one bit, then we have edges (i, j) and (j, i).

This graph contains $n \log n$ edges. Each vertex i wants to send a message to another vertex $\pi(i)$, where π is a permutation. Each edge can only send 1 message per time step (may have to wait in queue before goign down edge).
Goal: Complete all the message sending within time $O(\log n)$.
Oblivious Routing: Each message decides its path at the beginning of time.

Theorem 5.0.1 Brebner, Valiant (Leslie) (1981)

With probability $\geqslant 1 - \frac{1}{n^{1000}}$, we can achieve $O(\log n)$ total time.

Proof. Algorithm:

- To send $i \longrightarrow \pi(i)$;
  - Pick random $r_i \in \{1, \dots, n\}$.
  - Phase 1: Send $i \longrightarrow r_i$.
  - Phase 2: Send $r_i \longrightarrow \pi(i)$.

We wait until a preplanned time to begin phase 2. To send $i \longrightarrow r_i$, use bit-fixing where we fix bits from left to right. Let $u_i :=$ message going from $i \longrightarrow r_i$, $P_i =$ path that u_i takes. Focus on a fixed i.
Claim: With high probability, u_i completes p_i within $O(\log n)$ time.

Lemma 5.0.1 Lemma 1

If $i \neq j$, then edges in $P_i \cap P_j$ form a continguous subpath.

Proof. Consider an edge in P_i. For example if we have

$$\begin{array}{c} 010111011001 \\ 010111111001 \end{array}$$

we can say that the 011001 agrees with i and j, while the 0101111 agrees with r_i and r_j.

The values of k where paths agree,

$$\{k \mid r_i, r_j \text{ agree first } k \text{ bits}\} \cap \{k \mid i, j \text{ agree on final } n - k + 1 \text{ bits}\}.$$

Note that the first set is a prefix of $1, \dots n$ and the second set is a suffix of $1, \dots n$. Therefore, their intersection is a continguous interval. 😊

Lenma 5.0.2 Lemma 2

The number of time steps that u_i spends in queues is at most $|S_i|$, where $S_i = \{j \neq i \mid P_i \cap P_j \neq \emptyset\}$.

Proof. Say that a message w that is currently on P_i has "delay" d if

- w is on the t-th step of P_i for some t.
- We are in the $t + d$-th time step of the algorithm.

If u_i waits in a queue and its delay goes from d to $d + 1$, give a note with the number d on it to whoever used the edge that u_i wanted to use. Later on, if a message has a note "d", and w waits in a queue:

- if edge $\notin P_i$, keep the note
- otherwise, pass the note to message that's using the edge we're waiting on.

If note d is currently held by a message w and if w is still on P_i, then w's delay is exactly d. This means that no two notes will be held by the same message. This implies that the number of notes is equal to the number of messages hold notes, which is then less or equal to $|S_i|$. ☺

Now let e be an edge. Define $m_e :=$ the number of messages tha tuse e.

Lenma 5.0.3 Lemma 3

$E[m_e] \leqslant 1$.

Proof. What does it mean to use edge e? The first k bits are free variables for j while the last $n - k + 1$ agree with j. So the number of options for $j = 2^{k-1}$. FOr each such j,

$$P(r_j \text{ has correct first } k \text{ bits}) = \frac{1}{2^k}.$$

So,

$$E[\# \ j \text{ that use } e] \leqslant \sum_{\text{valid } j} P(j \text{ uses } e) \leqslant 2^{k-1}\frac{1}{2^k} = \frac{1}{2}.$$

☺

Lenma 5.0.4 Lemma 4

With high probability, $|S_i| \leqslant O(\log n)$.

Proof. First, fix P_i.

$$|S_i| = \sum_{j \neq i} \mathbb{I}[P_i \cap P_j \neq \emptyset],$$

where $\mathbb{I}$ is the indicator function. Note that these are all independent random variables determined by r_j. This is exactly the type of thing we want to apply Chernoff bounds to.

Define $\mu := E[|S_i|] \leqslant \sum_{e \in P_i} E[\# \text{ messages } j \neq i \text{ that use } e] \leqslant \frac{\log n}{2}$. ☺

So let's say $\mu = \frac{\log n}{2}$. So,

$$P(|S_i| \geqslant J\mu) \leqslant \left(\frac{1}{J}\right)^{\Omega(J\mu)} \leqslant \frac{1}{2}^{\Omega(J\mu)} \leqslant \frac{1}{\text{poly} n}.$$

☺

Suppose I flip a fair coin repeatedly:
Question 1: How many flips do I need to get $\geqslant 1$ heads with high probability.
Answer: $O(\log n) \implies P(\text{all tails}) \leqslant \frac{1}{2^{\Theta(\log n)}}$.
Question 2: How man flips do I need to get $\geqslant \log n$ heads with high probability?
Answer: $O(\log n)$. If I flip $1000 \log n$ coins, $P(< \log n \text{ heads}) = P(> 999 \log n \text{ tails})$. This means I exceeded the mean greatly as $E[\# \text{ of tails}] = 500 \log n$. If $\mu = E[T]$, then

$$\begin{aligned} T &> 1.9\mu \\ &> \mu + 0.9\sqrt{\mu}(\sqrt{\mu}). \end{aligned}$$

And the probability of this is $\leqslant e^{-\Omega(\log n)}$.

Chapter 6

Metric Embeddings

Definition 6.0.1: Metric Space

A metric space X is a set equipped with a function $d : X \times X \longrightarrow \mathbb{R}$ such that

1. $d(x, y) \geqslant 0$ for all $x, y \in X$, $d(x, y) = 0$ if and only if $x = y$.
2. $d(x, y) = d(y, x)$ for all $x, y \in X$
3. $d(x, z) \leqslant d(x, y) + d(y, z)$ for all $x, y, z \in X$

Example 6.0.1

- ℓ^p-distance. For $p \geqslant 1$, $X = \mathbb{R}^k$. Then

$$d_{\ell^p}(x, y) = \left(\sum_{i=1}^{k} |x_i - y_i|^p \right)^{1/p}.$$

- Graph-distance. Take any undirected graph with positive weights, then you can talk about the graph distance between two vertices where

$$d(v, u) = \text{graph distance}.$$

Definition 6.0.2: Metric Embeddings

Given two metric spaces, $(X_1, d_1), (X_2, d_2)$, a map $\phi : X_1 \longrightarrow X_2$ is a metric embedding with distortion at most α if $\exists \beta$ scaling parameter such that $\forall x, y \in X_1$, $d_1(x, y)$ is within a factor of α of $\beta d_2(\phi(x), \phi(y))$. That is,

$$\beta d_2(\phi(x), \phi(y)) \leqslant d_1(x, y) \leqslant \alpha \beta d_2(\phi(x), \phi(y)).$$

Theorem 6.0.1 Bourgain's Theorem

Let (X, d) be an n-point metric. Then there exists a function $\phi : (X; d) \longrightarrow (\mathbb{R}^{O(\log^2 n)}, d_{\ell^1})$ such that the distortion is $O(\log n)$.

Proof. We review the algorithm. Let c be a large constant.

- For $i = 1, \ldots, \log n$,
  For $j = 1, \ldots, c \log n$,
  $S_{i,j}$ = sample each element $x \in X$ with probability 2^{-i}.

☺

Lenma 6.0.1 Key Lemma

For $i \in \{1, \ldots, t\}$ and $j \in \{1, \ldots, c \log n\}$, we have with probability $\Omega(1)$ that

$$|d(x, S_{i,j}) - d(yS_{i,j})| \geqslant r_{i+1} - r_i.$$

I LEFT LECTURE HERE, FILLIN LATER

6.1 Day 2

Theorem 6.1.1

Let (X, d) be an n-point metric space. Suppose that $\forall x \neq y \in X$, $d(x, y) \in [1, n^2]$. We will embed X into a tree metric (T, d_T) with the following properties:

- depth of tree is $O(\log n)$.
- For any root-leaf path, edge weights are decreasing powers of 2.
- Points in X get mapped to leaves of the tree.

There exists a randomized embedding $\phi : X \to (T, d_T)$, where T is also a random variable, such that

- $d_T(\phi(x), \phi(y)) \geqslant d(x, y)$.
- $E[d_T(\phi(x), \phi(y))] \leqslant O(\log n) d(x, y)$.

Proof. The algorithm:

1. Pick a random permutation $\pi = \pi_1, \ldots, \pi_n$ of the elements of X.
2. Pick uniformly random $\beta \in [1, 2]$. Define $\beta_i := 2^{i-1} \cdot \beta$.

   For each $x \in X$ and $i \in \{0, \ldots, 2 \log n\}$, define

   $$\text{center}_i(x) = \text{first } \pi_j \text{ in permutation to satisfy } d(x, \pi_j) \leqslant \beta_i.$$

3. Construct ϕ as follows:

   Each node represents a set of elements. Root at level $2 \log n + 1$ = all of X. For any node w at level $i + 1$, if node $>$ 1 element.

   - Group w's elements z by $\text{center}_i(z) \to$ child nodes.

   Two elements $a, b \in w$ are in the same child iff $\text{center}_i(a) = \text{center}_i(b)$.

4. Finally, edge from level $i + 1$ to level i has weight 2^i.

Analysis:

We first claim that for $x, y \in X$, $d_T(\phi(x), \phi(y)) \geqslant d(x, y)$.

Proof. Observe that at level i, if x, y in the same node, then $d(x, y) \leqslant 2\beta_i \leqslant 2^{i+1}$. So let i be the highest level the path goes through, this implies that $d(x, y) \leqslant 2^{i+1}$. But also $d_T(\phi(x), \phi(y)) \geqslant 2 \cdot 2^i \geqslant d(x, y)$ as desired. ☺

Proposition 6.1.1

$$E[d_T(\phi(x), \phi(y))] \leqslant O(\log n) d(x, y).$$

Proof.

$$\begin{aligned} d_T(\phi(x), \phi(y)) &\leqslant O(\text{largest edge weight along the path}) \\ &\leqslant \sum_{i=0}^{2 \log n} \mathbb{I}(\text{center}_i(x) \neq \text{center}_i(y)) \cdot 2^{i+1}. \end{aligned}$$

So we have

$$\begin{aligned}E[d_T(\phi(x),\phi(y))] &\leqslant \sum_{i=0}^{2\log n}\int_{r=2^{i-1}}^{2^i} P(B_i = r)\cdot P(\mathrm{center}_i(x)\neq \mathrm{center}_i(y)\mid \beta_i = r)2^{i+1}\mathrm{d}r\\ &= \sum_{i=0}^{2\log n}\int_{r=2^{i-1}}^{2^i} \frac{1}{2^{i-1}}\cdot P(\mathrm{center}_i(x)\neq \mathrm{center}_i(y)\mid \beta_i = r)2^{i+1}\mathrm{d}r\\ &= \sum_{i=0}^{2\log n}\int_{r=2^{i-1}}^{2^i} P(\mathrm{center}_i(x)\neq \mathrm{center}_i(y)\mid \beta_i = r)\mathrm{d}r.\end{aligned}$$

If we look through $\pi_1, \pi_2, \dots, \pi_n$, the first of $\{\mathrm{center}_i(x), \mathrm{center}_i(y)\}$ to appear is w implies that w "cuts" (x, y) (and the centers are non-equal).

Let $w_1, w_2, \dots, w_n$ be elements of x sorted by distance from w_i to $\{x, y\}$.

$$\sum_{i=0}^{2\log n}\int_{r=2^{i-1}}^{2^i} P(\mathrm{center}_i(x)\neq \mathrm{center}_i(y)\mid \beta_i = r)\mathrm{d}r \leqslant \sum_{j=1}^{n}\sum_{i=0}^{2\log n}\int_{r=2^{i-1}}^{2_i} P(w\text{ cuts }(x,y)\mid \beta_i = r)\mathrm{d}r.$$

Call the integrand Ξ. We consider what must occur for w_j to cut (x, y). Suppose WLOG $d(w_j, x) \leqslant d(w_j, y)$. We need

- $d(w_j, x) \leqslant \beta_i \leqslant d(w_j, y)$.
- w_j must appear before $w_1, \dots, w_{j-1}$ in π, which has probability $\leqslant \frac{1}{j}$.

So we get

$$\begin{aligned}\sum_{j=1}^{n}\sum_{i=0}^{2\log n}\int_{r=2^{i-1}}^{2_i} P(w\text{ cuts }(x,y)\mid \beta_i = r)\mathrm{d}r &\leqslant \sum_{j=1}^{n}\int_{z=d(w_j,x)}^{d(w_j,y)}\frac{1}{j}\mathrm{d}z\\ &= \sum_{j=1}^{n}\frac{1}{j}|d(w_j,y) - d_w(w_j,x)|\\ &\leqslant d(x,y)\sum_{j=1}^{n}\frac{1}{j} = d(x,y)O(\log n).\end{aligned}$$

😄

😄

6.2 Day 3

> **Theorem 6.2.1** Johnson-Lindenstrauss Lemma (1994)
>
> Let $\epsilon \in (0,1)$. Let $S \subseteq \mathbb{R}^d$ be a set of n points. There exists a map $\phi : S \longrightarrow \mathbb{R}^{O(\log n)\cdot\epsilon^{-2}}$ which satisfies:
>
> $$\forall a, b \in S, (1 \pm \epsilon)\|a - b\|_2 = \|\phi(a) - \phi(b)\|_2.$$

Our goal is to construct a linear ϕ such that for any $a \in \mathbb{R}^d$, with high probability in n,

$$\|\phi(a)\|_2 = (1 \pm \epsilon)\|a\|_2.$$

If we can do this, then $\forall a, b \in S$, we have with high probability that

$$\begin{aligned} \|\phi(a) - \phi(b)\|_2 &= \|\phi(a - b)\|_2 \\ &= (1 \pm \epsilon)\|a - b\|_2. \end{aligned}$$

Then by applying the union bound over all $\binom{n}{2}$ pairs, can complete the proof.
We start with a warmup task. We want to construct a linear function $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ such that $E[f(a)^2] = \|a\|_2^2$. If we can find a function like this that also suffices:

$$P(f(a)^2 > K\|a\|_2^2) \leqslant e^{-\Omega(K)},$$

then we can finish the theorem off.
Review of normal random variables: Let v = variance, $\sigma = \sqrt{v}$ (standard-deviation). Then $N(0, v)$ represents a normal random variable with mean 0 and standard deviation σ.
Fact 1: Let $X \sim \mathrm{N}(0, \sigma^2)$. Then

$$P(|X| \geqslant K\sigma) < 2e^{-K^2/3}.$$

Fact 2: If $X \sim (0, v_1)$ and $Y \sim (0, v_2)$, and $X \perp Y$, then $X + Y \sim (0, v_1 + v_2)$.
Fact 3: Let $X \sim \mathrm{N}(0, v)$, $c \in \mathbb{R}$. Then

$$cX \sim \mathrm{N}(0, c^2 v).$$

For vector $\bar{a} = (a_1, \ldots, a_d)$, define

$$f(a) = \sum_{i=1}^{d} X_i \cdot a_i$$

where X_i is an independent random variable such that $X_i \sim \mathrm{N}(0, 1)$. That means $f(a) \sim \mathrm{N}(0, \|a\|_2^2)$.

> **Corollary 6.2.1** 1
>
> $E[f(a)^2] = \|a\|_2^2$.

> **Corollary 6.2.2** 2
>
> $P(f(a)^2 > K\|a\|_2^2) \leqslant e^{-\Omega(K)}$.

Proof. By the previous Chernoff bound,

$$P(|f(a)| > J\|a\|_2) \leqslant e^{-J^2/3}$$
$$P(f(a)^2 > J^2\|a\|_2^2) \leqslant 2e^{-J^2/3}.$$

By setting $K = J^2$, we get the corollary. 😄

We now turn to the proof of the Johnson-Lindenstrauss lemma.

Proof. Define $\phi(\vec{a})$ as

$$\phi(\vec{a}) = \langle \phi_1(\vec{a}), \ldots, \phi_r(\vec{a}) \rangle$$

where $r = c\epsilon^{-2} \log n$ where c is a large constant. Now we define ϕ_i. We take the same argument as in the previous proof:

$$\phi_i(a) = \frac{1}{\sqrt{r}} \sum_{j=1}^{d} X_{i,j} \cdot a_j$$

where all the $X_{i,j} \sim \mathrm{N}(0,1)$ and are independent. We want to show:

$$\begin{aligned} \|\phi(\vec{a})\|_2^2 &= (1 \pm \epsilon)\|a\|_2^2 \\ \iff \sum \phi_i(\vec{a})^2 &= (1 \pm \epsilon)\|a\|_2^2. \end{aligned}$$

We now go over a Chernoff bound for geometric random variables. Let $X_1, \ldots, X_n$ be independent. Let $\mu = E[\sum X_i]$. Suppose each X_i satisfies

$$P(|X_i| \geqslant K) \leqslant e^{-\Omega(k)}.$$

Let $X = \sum X_i$. In the small-deviation regime ($K \in [1, \sqrt{\mu}]$), we have

$$P(|X - E[X]| \geqslant K\sqrt{\mu}) \leqslant e^{-\Omega(k^2)}.$$

Then for $J \geqslant 2$, we have

$$P(X \geqslant J\mu) \leqslant \left(\frac{1}{2}\right)^{\Omega(J\mu)}.$$

Note that the large-deviation regime bound we're used to does not apply because it is not true in the $n = 1$ case. Now for each ϕ_i, we have

$$r \cdot \phi_i(\vec{a}) \sim \mathrm{N}(0, \|a\|_2^2).$$

If we wanted to normalize this, we would get

$$Y_i := \left(\frac{\sqrt{r} \cdot \phi_i(\vec{a})}{\|a\|_2}\right)^2 \sim \mathrm{N}(0,1)^2.$$

Y_i has mean 1, and $|Y_i|$ is a geometric random variable. This means

$$P\left[\left|\sum_{i=1}^{r} Y_i - r\right| \geqslant K\sqrt{r}\right] \leqslant e^{-\Omega(k^2)}.$$

Let's set $K\sqrt{r} = \epsilon r$, or that $K = \epsilon\sqrt{r}$. So,

$$P\left[\sum_{i=1}^{r} Y_i \neq (1 \pm \epsilon) r\right] = e^{-\Omega(\epsilon^2 r)}.$$

But $-\Omega(\epsilon^2 r) = \frac{1}{\mathrm{poly}(n)}$. What follows is:

$$\begin{aligned} \sum \frac{r\phi_i(\vec{a})^2}{\|a\|_2^2} &= (1 \pm \epsilon) r \\ \implies \sum \phi_i(\vec{a})^2 &= (1 \pm \epsilon)\|a\|_2^2. \end{aligned}$$

This is exactly what we wanted to prove. ☺

We now prove the facts about normal random variables.

> **Corollary 6.2.3** from Central Limit Theorem
>
> Suppoose you define $X^{(n)} := \sum_{i=1}^{n} X_i$ where X_i takes on 1 or -1 with probability $1/2$. Then as $n \to \infty$, $\frac{\sqrt{v}}{\sqrt{n}} X^{(n)}$ converges pointwise to a single distribution. This is the definition of $\mathrm{N}(0, v)$.

As a shorthand,

$$\lim_{n\to\infty} \frac{\sqrt{v}}{n} X^{(n)} \sim N(0, v).$$

We skip the first fact because it's an easy Chernoff bound. Fact 3 is also straight-forward from the definition of a random variable. So we will just prove that if we have $A, B \sim \mathrm{N}(0, v_1), \mathrm{N}(0, v_2)$, then $A + B \sim \mathrm{N}(0, v_1 + v_2)$. As an easy example, take $v_1 = v_2 = 1$. So then the distribution of $A + B$ would be

$$\begin{aligned}
\lim_{n\to\infty} \frac{1}{\sqrt{n}}(X^{(n)} + Y^{(n)}) &= \lim_{n\to\infty} \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n + Y_1 + \cdots + Y_n) \\
&= \lim_{n\to\infty} \frac{1}{\sqrt{n}} X^{(2n)} \\
&= \lim_{n\to\infty} \frac{\sqrt{2}}{\sqrt{m}} X^{m} && \text{(where } m = 2n\text{)} \\
&= \lim_{m\to\infty} \frac{\sqrt{2}}{\sqrt{m}} \cdot X^{(m)} \sim \mathrm{N}(0, 2).
\end{aligned}$$