CSE 694 – Prob. Analysis and Randomized Algo.

What is it about?

- Probabilistic thinking!

Administrative Stuff

- 5 assignments (to be done individually)
- 1 final presentation and report (I will assign papers and topic)

First few weeks

- Gentle introduction to concepts and techniques from probability theory
- Done via sample problems from many areas (networking, algorithms, combinatorics, coding, learning theory, etc.)

PTCF $=$ *Probability Theory Concepts and Facts*

- Discrete Probability Space, Events, Union Bound

Example 1: Tracing A Single Traitor

- Application: "broadcast" to a group of legitimate users
  - DVD or CD-ROM distribution of movies or softwares
  - Pay-per-view subscriptions
  - Online databases
- Some user might be traitor, giving his key(s) to a pirate
- Pirate sells decryption device on black market
- Problem: obtain device, identify the (single) traitor
- Two extremes, both do not work well
  - Single shared key: can't trace the traitor
  - Each person a key: cipher-text too large!

Traitor Tracing and Sperner Family

- Set of keys T, $|T| = t$
- n users, user i given a subset $F_i \subseteq T$ of keys

Claim

To be able to trace a traitor, $F_i \not\subseteq F_j$, for all $i \neq j$.

- A family $\mathcal{F}$ of sets where no member is contained in another is called a Sperner family

Main Questions

- Given n, find the smallest t for which a Sperner family of n sets on $[t] = \{1, \cdots, t\}$ exists
- Dually, given t find the maximum n for which a Sperner family of n sets on $[t]$ exists

A Classic Theorem by Sperner

Theorem (Sperner, 1928)

The maximum size of a family $\mathcal{F}$ of subsets of $[t]$ whose members do not contain one another is $\binom{t}{\lfloor t/2 \rfloor}$.

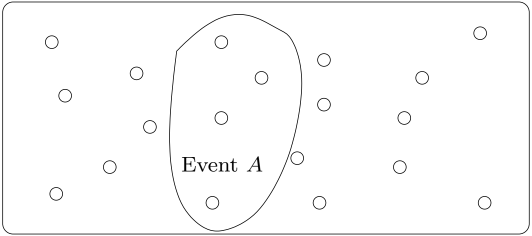
- The collection of $\lfloor t/2 \rfloor$-subsets of $[t]$ is a Sperner family
- Thus, suffices to show that $|\mathcal{F}| \leq \binom{t}{\lfloor t/2 \rfloor}$ for any Sperner family $\mathcal{F}$
- Pick a permutation π of $[t]$ uniformly at random
- For $F \in \mathcal{F}$, let A_F be the event that F is a prefix of π

$$\text{Prob}[A_F] = \frac{k!(t-k)!}{t!} = \frac{1}{\binom{t}{k}} \geq \frac{1}{\binom{t}{\lfloor t/2 \rfloor}}, \quad \text{where } k = |F|$$

- The A_F are mutually exclusive (why?), hence

$$1 \geq \text{Prob}\left[\bigcup_{F \in \mathcal{F}} A_F\right] = \sum_{F \in \mathcal{F}} \text{Prob}[A_F] \geq \frac{|\mathcal{F}|}{\binom{t}{\lfloor t/2 \rfloor}}$$

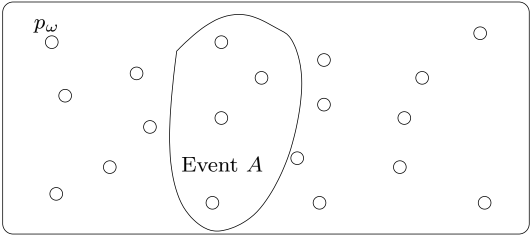
PTCF: Simple Probability Space



Sample Space Ω

- Ω is a finite set of all possible outcomes of some experiment
- Each outcome occurs equally likely
- A subset A of outcomes is an event
  - Think of it as a set of outcomes satisfying a certain property
- $\text{Prob}[A] = \frac{|A|}{|\Omega|}$: the fraction of outcomes in A
- In most cases, **not** a good way to think about probability spaces

PTCF: Discrete Probability Space



Sample Space Ω

- Each $\omega \in \Omega$ is assigned a number $p_\omega \in [0,1]$, such that $\sum_{\omega \in \Omega} p_\omega = 1$.
- For any event A, Prob$[A] = \sum_{\omega \in A} p_\omega$.
- In the simple space, $p_\omega = \frac{1}{|\Omega|}, \forall \omega$
- Note: this is **not** the most general definition, but suffices for now.

PTCF: How do we "assign" the p_ω?

- Could think of it as a mathematical function, like saying "give each outcome ω a number p_ω equal to $1/|\Omega|$"
- That's **not** the probabilistic way of thinking!
- Probabilistic way of thinking:
  - An experiment is an *algorithm* whose outcome is not deterministic
  - For example, algorithms making use of a random source (like a bunch of "fair" coins)
  - Ω is the set of all possible outputs of the algorithm
  - p_ω is the "likelihood" that ω is output

Example 2: Tracing a Group of Traitors

- Suppose we know there are $\leq d < n$ traitors out of n users
- User j gets key set F_j, set system $\mathcal{F} = \{F_j\}_{j=1}^n$

Claim (Property $\mathcal{F}$ must satisfy)

For arbitrary $j_0, j_1, \cdots, j_d \in [n]$,

$$F_{j_0} \not\subseteq F_{j_1} \cup \cdots \cup F_{j_d}.$$

- Such a family $\mathcal{F}$ is called a d-cover-free family
- d-cover-free family of size n on $[t]$ is equivalent to d-disjunct matrix

Non-Adaptive Group Testing

- A $t \times n$ binary matrix $\mathbf{A}$ is called d-disjunct iff the union of any d columns does not contain another column
- Columns are codewords of superimposed codes
- Rate of the code is $R(\mathbf{A}) = \frac{\log n}{t}$
- Want codes with high rates. But, as $n \to \infty$ and $d \to \infty$

$$\frac{1}{d^2 \log e}(1 + o(1)) \leq \limsup_{\mathbf{A}} R(\mathbf{A}) \leq \frac{2 \log d}{d^2}(1 + o(1))$$

(From Dyachkov, Rykov (1982), and Dyachkov, Rykov and Rashad (1989))

- We'll prove the lower bound

The Probabilistic Method

Want to prove that $t \times n$ d-disjunct matrix exists with small t

Strategy:

- Fix t (which we'll choose later)
- Choose a $t \times n$ matrix $\mathbf{A}$ at random, somehow
- Prove that, with $t = t(d, n)$,

$$\text{Prob}[\mathbf{A} \text{ is } d\text{-disjunct}] > 0.$$

- Or, equivalently

$$\text{Prob}[\mathbf{A} \text{ is not } d\text{-disjunct}] < 1.$$

Existence of Good d-disjunct Matrix

- Set a_{ij} to 1 with probability p
- Fix j_0 and a set $C = \{j_1, \cdots, j_d\} \subseteq [n]$, $j_0 \notin C$
- (j_0, C) is bad for $\mathbf{A}$ if column j_0 is contained in the union of columns in C
- Let $B_{j_0,C}$ be the event that (j_0, C) is bad
- $\mathbf{A}$ is not d-disjunct implies $\bigcup_{(j_0,C)} B_{j_0,C}$, thus

$$\text{Prob}[A \text{ is not } d\text{-disjunct}] \leq \text{Prob}\left[\bigcup_{j_0,C} B_{j_0,C}\right] \leq \underbrace{\cdots}_{how?} < 1$$

PTCF: The Union Bound

Lemma

Let $B_1, B_2, \dots$ be any finite or countably infinite sequence of events. Then,

$$\text{Prob}\left[\bigcup_{i\geq 1} B_i\right] \leq \sum_{i\geq 1} \text{Prob}[B_i]$$

Note:

- this bound holds for **any** probability space (not just simple spaces).
- the bound is simple but extremely useful!

Existence of Good d-disjunct Matrix

$$\text{Prob}\left[\bigcup_{j_0,C} B_{j_0,C}\right] \leq \sum_{j_0,C} \text{Prob}[B_{j_0,C}] = \sum_{j_0,C}\left[1-p(1-p)^d\right]^t$$

- Set $p = 1/(d+1)$, $\mathbf{A}$ is **not** d-disjunct with probability at most

$$(d+1)\binom{n}{d+1}\left[1-\frac{1}{d+1}(1-\frac{1}{d+1})^d\right]^t$$

- $f(x) = (1-1/(x+1))^x$ is decreasing when $x \geq 1$, and $\lim_{x\to\infty} f(x) = 1/e$, hence $f(x) \geq 1/e$
- RHS is upper-bounded by

$$(d+1)\binom{n}{d+1}\left[1-\frac{1}{e(d+1)}\right]^t \leq (d+1)\left(\frac{ne}{d+1}\right)^{d+1} e^{-1/e(d+1)}$$

- This is < 1 as long as $t \geq 2e(d+1)^2 \ln\left(en/(d+1)\right)$.

PTCF: Two Very Useful Inequalities

$$1 + x \le e^x, \qquad \forall x \in \mathbb{R} \tag{1}$$

$$\sum_{i=0}^{d} \binom{n}{i} \le \left(\frac{ne}{d}\right)^d, \qquad \forall d \le n \tag{2}$$

Some Key Ideas We've Learned

- To show the existence of some combinatorial object, set up some probability space and show that it exists with probability > 0
- The above is essentially a pigeonhole principle kind of proof, casted in probabilistic language
- We will see throughout the course that the probabilistic language is crucial!