The Probabilistic Method: Proof Through a Probabilistic Argument

- Compute

$$\sum_{i=0}^{n} i \binom{n}{i} \left(\frac{1}{2}\right)^n$$

Proof Through a Probabilistic Argument

- Compute

$$\begin{aligned}
\sum_{i=0}^{n} i\binom{n}{i}\left(\frac{1}{2}\right)^n &= \sum_{i=1}^{n} i\frac{n!}{i!(n-i)!}\left(\frac{1}{2}\right)^n \\
&= \sum_{i=1}^{n} n\frac{(n-1)!}{(i-1)!(n-i)!}\left(\frac{1}{2}\right)^n \\
&= \frac{n}{2}\sum_{i=1}^{n}\frac{(n-1)!}{(i-1)!(n-i)!}\left(\frac{1}{2}\right)^{n-1} \\
&= \frac{n}{2}\sum_{i=0}^{n-1}\frac{(n-1)!}{(i)!(n-i-1)!}\left(\frac{1}{2}\right)^{n-1} \\
&= \frac{n}{2}
\end{aligned}$$

Proof Through a Probabilistic Argument

- Compute

$$\sum_{i=0}^{n} i \binom{n}{i} \left(\frac{1}{2}\right)^n$$

- Let $X \sim B(n, 1/2)$,
- X_i independent r.v. with $Pr(X_i = 1) = Or(X_i = 0) = 1/2$.

$$\sum_{i=0}^{n} i \binom{n}{i} \left(\frac{1}{2}\right)^n = E[X] = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i] = \frac{n}{2}$$

- We prove a deterministic statement using a probabilistic argument!

Theorem

Given any graph $G = (V, E)$ with n vertices and m edges, there is a partition of V into two disjoint sets A and B such that at least $m/2$ edges connect vertex in A to a vertex in B.

Proof.

Construct sets A and B by randomly assign each vertex to one of the two sets.
The probability that a given edge connect A to B is $1/2$, thus the expected number of such edges is $m/2$.
Thus, there exists such a partition. $\square$

Sample and Modify

An *independent set* in a graph G is a set of vertices with no edges between them.
Finding the largest independent set in a graph is an NP-hard problem.

Theorem

Let $G = (V, E)$ be a graph on n vertices with $dn/2$ edges. Then G has an independent set with at least $n/2d$ vertices.

Algorithm:

1. Delete each vertex of G (together with its incident edges) independently with probability $1 - 1/d$.
2. For each remaining edge, remove it and one of its adjacent vertices.

$X =$ number of vertices that survive the first step of the algorithm.

$$E[X] = \frac{n}{d}.$$

$Y =$ number of edges that survive the first step.
An edge survives if and only if its two adjacent vertices survive.

$$E[Y] = \frac{nd}{2}\left(\frac{1}{d}\right)^2 = \frac{n}{2d}.$$

The second step of the algorithm removes all the remaining edges, and at most Y vertices.
Size of output independent set:

$$E[X - Y] = \frac{n}{d} - \frac{n}{2d} = \frac{n}{2d}.$$

Conditional Expectation

Definition

$$E[Y \mid Z = z] = \sum_y y \Pr(Y = y \mid Z = z),$$

where the summation is over all y in the range of Y.

Lemma

For any random variables X and Y,

$$E[X] = \sum_y \Pr(Y = y) E[X \mid Y = y],$$

where the sum is over all values in the range of Y.

Derandomization using Conditional Expectations

Given a graph $G = (V, E)$ with n vertices and m edges, we showed that there is a partition of V into A and B such that at least $m/2$ edges connect A to B.
How do we find such a partition?

$C(A, B) =$ number of edges connecting A to B.
If A, B is a random partition $E[C(A, B)] = \frac{m}{2}$.
Algorithm:

1. Let $v_1, v_2, \ldots, v_n$ be an arbitrary enumeration of the vertices.
2. Let x_i be the set where v_i is placed ($x_i \in \{A, B\}$).
3. For $i = 1$ to n do:
    1. Place v_i such that

$$E[C(A, B) \mid x_1, x_2, \ldots, x_i]$$
$$\geq E[C(A, B) \mid x_1, x_2, \ldots, x_{i-1}] \geq m/2.$$

Lemma

For all $i = 1, \ldots, n$ *there is an assignment of* v_i *such that*

$$E[C(A,B) \mid x_1, x_2, \ldots, x_i]$$
$$\geq E[C(A,B) \mid x_1, x_2, \ldots, x_{i-1}] \geq m/2.$$

Proof.

By induction on i.

For $i = 1$, $E[E[C(A,B) \mid X_1]] = E[C(A,B)] = m/2$

For $i > 1$, if we place v_i randomly in one of the two sets,

$$\begin{aligned} & E[C(A,B) \mid x_1, x_2, \ldots, x_{i-1}] \\ = \; & \frac{1}{2} E[C(A,B) \mid x_1, x_2, \ldots, x_i = A] \\ & + \frac{1}{2} E[C(A,B) \mid x_1, x_2, \ldots, x_i = B]. \end{aligned}$$

$$\begin{aligned} & \max(E[C(A,B) \mid x_1, x_2, \ldots, x_i = A], \\ & E[C(A,B) \mid x_1, x_2, \ldots, x_i = B]) \\ \geq \; & E[C(A,B) \mid x_1, x_2, \ldots, x_{i-1}] \\ \geq \; & m/2 \end{aligned}$$

How do we compute

$$\max(E[C(A,B) \mid x_1, x_2, \ldots, x_i = A], E[C(A,B) \mid x_1, x_2, \ldots, x_i = B])$$
$$\geq E[C(A,B) \mid x_1, x_2, \ldots, x_{i-1}] \geq m/2$$

We just need to consider edges between v_i and $v_1, \ldots, v_{i-1}$.

Simple Algorithm:

1. Place v_1 arbitrarily.
2. For $i = 2$ to n do
    1. Place v_i in the set with smaller number of neighbors.

Perfect Hashing

Goal: Store a **static disctionary** of n items in a table of $O(n)$ space such that any search takes $O(1)$ time.

Universal hash functions

Definition

Let U be a universe with $|U| \geq n$ and $V = \{0, 1, \ldots, n-1\}$. A family of hash functions $\mathcal{H}$ from U to V is said to be *k-universal* if, for any elements $x_1, x_2, \ldots, x_k$, when a hash function h is chosen uniformly at random from $\mathcal{H}$,

$$\Pr(h(x_1) = h(x_2) = \ldots = h(x_k)) \leq \frac{1}{n^{k-1}}.$$

Example of 2-Universal Hash Functions

Universe $U = \{0, 1, 2, \ldots, m-1\}$
Table keys $V = \{0, 1, 2, \ldots, n-1\}$, with $m \geq n$.
A family of hash functions obtained by choosing a prime $p \geq m$,

$$h_{a,b}(x) = ((ax + b) \bmod p) \bmod n,$$

and taking the family

$$\mathcal{H} = \{h_{a,b} \mid 1 \leq a \leq p-1, 0 \leq b \leq p\}.$$

Lemma

$\mathcal{H}$ is 2-universal.

Lemma

$\mathcal{H}$ is 2-universal.

Proof.

We first observe that for $x_1, x_2 \in \{0, \ldots, p-1\}$, $x_1 \neq x_2$,

$$ax_1 + b \neq ax_2 + b \mod p.$$

Thus, if $h_{a,b}(x_1) = h_{a,b}(x_2)$ there is a pair (s, r) such that $s \neq r$, $s = r \mod n$, and

$$\begin{aligned} (ax_1 + b) \mod p &= r \\ (ax_2 + b) \mod p &= s \end{aligned}$$

There are p choices of r, and for each pair (r, s) there is only one pair (a, b) that satisfies the relation.
For each r there are $\leq \lceil \frac{p}{n} \rceil - 1$ values $s \neq r$ such that $s = r \mod n$.
Thus, the probability of a collision is $\leq \frac{p(\lceil \frac{p}{n} \rceil - 1)}{p(p-1)} \leq \frac{1}{n}$. □

Lemma

If $h \in \mathcal{H}$ is chosen uniformly at random from a 2-universal family of hash functions mapping the universe U to $[0, n-1]$, then for any set $S \subset U$ of size m, with probability $\geq 1/2$ the number of collisions is bounded by m^2/n.

Proof.

Let $s_1, s_2, \ldots, s_m$ be the m items of S. Let X_{ij} be 1 if the $h(s_i) = h(s_j)$ and 0 otherwise. Let $X = \sum_{1 \leq i < j \leq n} X_{ij}$.

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{1 \leq i < j \leq n} X_{ij}\right] = \sum_{1 \leq i < j \leq m} \mathbf{E}[X_{ij}] \leq \binom{m}{2}\frac{1}{n} < \frac{m^2}{2n},$$

Markov's inequality yields
$\Pr(X \geq m^2/n) \leq \Pr(X \geq 2\mathbf{E}[X]) \leq \frac{1}{2}$. □

Definition

A hash function is perfect for a set S if it maps S with no collisions.

Lemma

If $h \in \mathcal{H}$ is chosen uniformly at random from a 2-universal family of hash functions mapping the universe U to $[0, n-1]$, then for any set $S \subset U$ of size m, such that $m^2 \leq n$ with probability $\geq 1/2$ the hash function is perfect

Theorem

The two-level approach gives a perfect hashing scheme for m items using $O(m)$ bins.

Level I: use a hash table with $n = m$. Let X be the number of collisions,

$$\Pr(X \geq m^2/n) \leq \Pr(X \geq 2\mathbf{E}[X]) \leq \frac{1}{2}.$$

When $n = m$, there exists a choice of hash function from the 2-universal family that gives at most m collisions.

Level II: Let c_i be the number of items in the i-th bin. There are $\binom{c_i}{2}$ collisions between items in the i-th bin, thus

$$\sum_{i=1}^{m} \binom{c_i}{2} \leq m.$$

For each bin with $c_i > 1$ items, we find a second hash function that gives no collisions using space c_i^2. The total number of bins used is bounded above by

$$m + \sum_{i=1}^{m} c_i^2 \leq m + 2\sum_{i=1}^{m} \binom{c_i}{2} + \sum_{i=1}^{m} c_i \leq m + 2m + m = 4m.$$

Hence the total number of bins used is only $O(m)$.

The First and Second Moment

Theorem

For an integer random variable X,

- $Pr(X > 0) = Pr(X \geq 1) \leq E[X]$
- $Pr(X = 0) \leq Pr(|X - E[X]| \geq E[X]) \leq \frac{Var[X]}{(E[X])^2}$

Application: Number of Isolated Nodes

Let $G_{n,p} = (V, E)$ be a random graph generated as follows:

- The graph has n nodes.
- Each of the $\binom{n}{2}$ pairs of vertices are connected by an edge with probability p independently of any other edge in the graph.

A node is isolated if it is adjacent to no edges.

If $p = 0$ all vertices are isolated (have no edges). If $p = 1$ no vertex is isolated. What can we say for $0 < p < 1$?

Application: Number of Isolated Nodes

Let $G_{n,p} = (V, E)$ be a random graph generated as follows:

- The graph has n nodes.
- Each of the $\binom{n}{2}$ pairs of vertices are connected by an edge with probability p independently of any other edge in the graph.

A node is isolated if it has no edges.

Theorem

For any function $w(n) \to \infty$

- *If $p = \frac{\log n - w(n)}{n}$, then whp the graph has isolated nodes.*
- *If $p = \frac{\log n + w(n)}{n}$, then whp the graph has no isolated nodes.*

Proof

For $i = 1, \ldots, n$, let $X_i = 1$ if node i is isolated, otherwise $X_i = 0$. Let $X = \sum_{i=1}^{n} X_i$.

$$E[X] = n(1-p)^{n-1}$$

For $p = \frac{\log n + w(n)}{n}$

$$E[X] = n(1-p)^{n-1} \leq e^{\log n - (n-1)p} \leq e^{-w(n)} \to 0$$

Thus, for $p = \frac{\log n + w(n)}{n}$,

$$Pr(X > 0) \leq E[X] \to 0$$

To use the second moment method we need to bound $Var[x]$.

$$Var[X_i] \leq E[X_i^2] = E[X_i] = (1-p)^{n-1}$$

$$Cov(X_i, X_j) = (1-p)^{2n-3} - (1-p)^{2n-2}$$

$$\begin{aligned} Var[X] &\leq \sum_{i=1}^{n} Var[X_i] + \sum_{i \neq j} Cov(X_i, X_i) \\ &= n(1-p)^{n-1} + n(n-1)(1-p)^{2n-3} - n(n-1)(1-p)^{2n-2} \\ &= n(1-p)^{n-1} + n(n-1)p(1-p)^{2n-3} \end{aligned}$$

$$
\begin{aligned}
Var[X] &= \sum_{i=1}^{n} Var[X_i] + \sum_{i \neq j} Cov(X_i, X_i) \\
&= n(1-p)^{n-1} + n(n-1)p(1-p)^{2n-3}
\end{aligned}
$$

$$
\begin{aligned}
Pr(X=0) &= Pr(|X - E[X]| \geq E[X]) \leq \frac{Var[X]}{(E[X])^2} \\
&= \frac{n(1-p)^{n-1} + n(n-1)p(1-p)^{2n-3}}{n^2(1-p)^{2n-2}} \\
&= \left(1 - \frac{1}{n}\right) \frac{p}{1-p} + \frac{1}{n(1-p)^{n-1}}
\end{aligned}
$$

For $p = \frac{\log n - w(n)}{n}$,

$$\begin{aligned} Pr(X = 0) &\leq \frac{Var[X]}{(E[X])^2} \\ &= \left(1 - \frac{1}{n}\right) \frac{p}{1-p} + \frac{1}{n(1-p)^{n-1}} \to 0 \end{aligned}$$

Since

$$n(1-p)^{n-1} \geq ne^{-p(n-1)}(1 - \frac{p^2}{n}) \geq \frac{1}{2}e^{w(n)}$$

We use: for $|X| \leq 1$

$$e^x \left(1 - \frac{x^2}{n}\right) \leq \left(1 + \frac{x}{n}\right)^n \leq e^x$$