1 Element Distinctness

In the Element Distinctness problem ED_n, we are given a function $f : [n] \mapsto [m]$, $m \geq n$, as an oracle. f can also be represented as a list of numbers: $f = (f_1, f_2, \ldots, f_n)$. We would like to know if f is a one-one (i.e., an injective function) or not. If f is not one-one, we say there is a *collision*. The goal of the algorithm is to answer this question with as few queries to the oracle as possible.

Ambainis recently designed an algorithm based on quantum walks that achieves the optimal bound of $O(n^{2/3})$ on the query complexity of ED_n. Ambainis' quantum algorithm is fashioned out of a random walk based algorithm, which we will see shortly. We will study a general method due to Mario Szegedy which, inspired by the Ambainis algorithm, gives a recipe for translating classical walk based algorithms into quantum ones, often leading to a computational speed-up.

The straight forward algorithm to solve Element Distinctness(ED_n) is to go through the list f one number at a time, and check for duplications. In fact, this is optimal.

Fact 1 *Classical query complexity of* $ED_n \in \Theta(n)$.

Interestingly, the random walk based algorithm we will see has worse performance (for certain parameters) and yet leads to a speed-up on quantization. On the other hand, there is no obvious way to speed up the naive algorithm.

A classical walk based algorithm for ED_n. The idea behind the algorithm is to perform a random walk on the space of subsets of a fixed size r until a collision is detected:

1. Pick a uniformly random set r-subset I (a subset of size of r elements out of $[n]$), and query f at all the points in I. If there is a collision within I, stop, and output the solution.
2. Repeat for $\leq T$ steps,
   (a) Pick $i \in I$ and $j \in [n] - I$ uniformly at random. Swap these elements, i.e. $I \leftarrow (I - \{i\}) \cup \{j\}$.
   (b) Query f_j and if there is a collision, stop and output the elements that collide.
3. Output "no collision" (i.e. f is one-one).

This walk is *irreducible* in the sense that starting from any subset we can reach any other r-subset by a sequence of transitions as described above. In particular, starting with a subset that does not contain collision, we can reach a subset that contains the collision in a finite number of steps. We can easily bound the (expected) number of steps the random walk takes to find a collision, if there is one.

Let T be the first time the walk reaches an r-subset containing a collision (called a *stopping time*), starting at a random r-subset.

Note that

$$Pr(\text{walk stops within two steps starting at an arbitrary state}) \;\geq\; 1 \cdot \frac{2}{n-r} \cdot \frac{r-1}{r} \cdot \frac{1}{n-r}$$

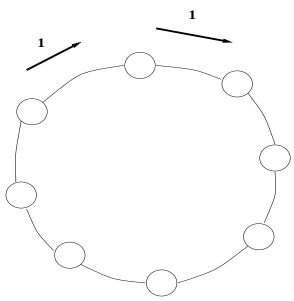

Figure 1: An Example of a periodic Markov Chain

This is because in the worst case, we don't have any colliding element in the r-subset initially, and then we pick one of the two colliding elements from $n - r$ elements not in the set, followed by choosing an element in the r-subset that is not part of the colliding set, swapping it with the other colliding element not in the r-subset.

This implies that in expectation, T is bounded as

$$\mathrm{E}(T) \;\leq\; \frac{r(n-r)^2}{r-1}.$$

(If the probability of success of an experiment is p, the expected number of times we need to repeat the experiment until it succeeds is $1/p$.)

Next we will turn to more sophisticated arguments that give better bounds on $\mathrm{E}(T)$ (called a hitting time).

2 Hitting times in classical walks

Consider a Markov Chain on the state space X, $(|X| = N)$ given by the transition matrix P, where $P = (p_{x,y})$, $x, y \in X$ and $p_{x,y} = \Pr(\text{next state} = y | \text{current state} = x)$. P is a stochastic matrix, i.e., $\sum_y p_{x,y} = 1$ for all x. So all the rows of P sum up to 1.

We will assume that the Markov Chain is

1. Symmetric: $p_{x,y} = p_{y,x}$. In particular, the underlying graph of the walk is undirected.
2. Irreducible: There is a path between every pair of states.
3. Aperiodic: The underlying graph that is not bipartite.

   An equivalent formulation is: there exists $x \in X$ and $t_x \geq 1$ such that the probability that we return to x in t steps starting from x is positive for all $t \geq t_x$. This implies the same property for all $y \in X$ if the Markov chain is irreducible.

In the next homework, we will analyse such Markov chains and show that they have the following properties.

Since P is symmetric, P is doubly stochastic: all entries of the matrix are non-negative, and each row and each column sums up to 1. Thus, the uniform distribution $\pi^{\mathrm{T}} = (\frac{1}{N}, \frac{1}{N}, \ldots, \frac{1}{N})$ is a stationary distribution: $\pi^{\mathrm{T}} P = \pi^{\mathrm{T}}$. In other words, π is an eigenvector with eigenvalue 1. Since P is symmetric, it orthogonally diagonalizable, and all its eigenvalues are real. Since it is a doubly stochastic matrix, these eigenvalues all lie in the interval $[-1, 1]$. Irreducibility implies that π is the unique 1 eigenvector, and aperiodicity implies that all eigenvalues are strictly larger than -1. Let s be any start distribution on X. The above properties together imply that $s^{\mathrm{T}} P^t \longmapsto \pi^{\mathrm{T}}$ as $t \mapsto \infty$ (in the ℓ_1 metric). So regardless of the initial state, the Markov chain converges to a unique stationary distribution.

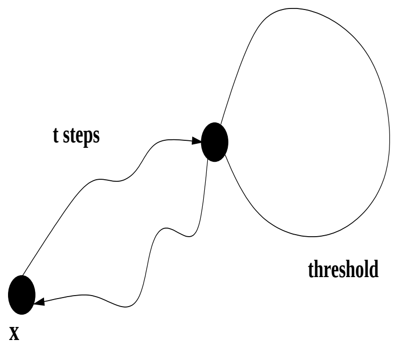

Figure 2: Reaching x in t steps starting from x

Now consider a search algorithm over the space X, where we attempt to locate a "marked" state by simulating the Markov chain P until the marked state is reached. This is precisely the form of the random walk based algorithm for element distinctness (there, a marked state is a subset that contains a collision).

Since we stop at a marked state, the transition matrix for this search process is different from the original Markov chain. Suppose $M \subseteq X$ is a given set of marked states. The transition matrix now looks like

$$\tilde{P}_M = \begin{pmatrix} P_M & P' \\ 0 & I \end{pmatrix},$$

where $(P_M|P')$ are the rows of P corresponding to states $X - M$, P_M is the matrix obtained by deleting rows and columns corresponding to M from P.

What is the hitting time of M?

Lemma 2 *Let T be the stopping time for finding a marked state starting in distribution s on X. Then, $E(T) = s_M^{\mathrm{T}} \cdot (I - P_M)^{-1} \cdot \mathbf{1}$, where s_M is the restriction of s onto $X - M$, and $\mathbf{1} = (1, 1, \ldots, 1)^{\mathrm{T}}$.*

Proof : For any non-negative integer-valued random variable T, $E(T) = \sum_{t=0}^{\infty} Pr(T > t)$.

In our case, $\Pr(T > t)$ is the probability that we have not reached the marked state after t steps, which is also the probability that the walk is in a state in $X - M$. The state distribution after t steps is $s^{\mathrm{T}} \tilde{P}_M^t$. Note that

$$\tilde{P}_M^t = \begin{pmatrix} P_M^t & P'(t) \\ 0 & I \end{pmatrix},$$

where $P'(t) = (P_M^{t-1} + P_M^{t-2} + \cdots + I)P'$. If $\chi(X - M)$ is the characteristic vector of the subset $X - M$, we have

$$\Pr(T > t) \;=\; s^{\mathrm{T}} \tilde{P}_M^t \chi(X - M) \;=\; s_M^T P_M^t \mathbf{1},$$

Then

$$\begin{aligned} \mathrm{E}(T) &= \sum_{t=0}^{\infty} s_M^{\mathrm{T}} P_M^t \mathbf{1} \\ &= s_M^{\mathrm{T}} \left(\sum_{t=0}^{\infty} P_M^t \right) \mathbf{1} \\ &= s_M^{\mathrm{T}} (I - P_M)^{-1} \mathbf{1}. \end{aligned}$$

Note that if P is symmetric and irreducible, and $M \neq \emptyset$, then all the eigenvalues of P_M are strictly smaller than 1. Therefore $I - P_M$ is invertible, and the hitting time is finite.

□