Pigeonhole Principle

Prof. Peter Shor

1 Pigeonhole Principle

The first lecture is about the pigeonhole principle. This is a very simple, and surprisingly powerful, proof technique.

Getting dressed. Let's start with an example. There is an old puzzle which goes as follows. A mathematician gets up in the dark and, to avoid waking his wife, gets dressed in another room. He grabs some socks from his sock drawer to put on. If he only has two colors of socks, how many socks does he need to guarantee a matching pair?

The answer is three. Suppose the colors are brown and black. Let us assume that the first sock is brown. For this sock not to be matched, the second and third socks must both be black. But then they match each other, showing that three socks (of two colors) always contain a match.

There is another way of thinking about it, which generalizes to the pigeonhole principle. We have three socks and two colors. The important thing her is that there are more socks than colors. Suppose we put each sock into a pigeonhole that depends only on its color. Since we have more socks than pigeonholes, there must be one pigeonhole that contains at least two socks.

This is the simplest way to state the pigeonhole principle. Suppose you have n pigeons and m pigeonholes, with $n > m$. Then, if every pigeon is in a hole, some hole must contain at least two pigeons.

Equal sum subsets. Let's look at some more applications of the pigeonhole principle. Suppose we have 30 7-digit numbers. We will show that there are two disjoint subsets of these numbers which have the same sum. How do we do this? There are 2^{30} distinct subsets of these numbers; these will be the pigeons. For each of these subsets, the sum of the numbers in the subset will be the pigeonhole. Since our numbers are all between 0 and 10^7, the sum of thirty of them is at most $3 \cdot 10^8$, which is less than $2^{30} \approx 10^9$. Thus, since there are more subsets (pigeons) than sums (holes), there must be two subsets that have the same sum. Thus, we have an equation

$$x_{i_1} + x_{i_2} + \ldots + x_{i_k} = x_{j_1} + x_{j_2} + \ldots + x_{j_l}$$

These two subsets may not disjoint, but we can eliminate any variable that appears on both sides of this equation. If we do this, then we get two disjoint subsets which both have the same sum. There is just one last thing to check, which is that we don't get the equation $0 = 0$ after eliminating the common variables. This cannot happen since the original two subsets were different.

One more comment on this question. While it is quite easy to prove that these numbers exist, they are quite hard to find. With 30 7-digit numbers, the problem is in the range of modern computers, but if you increase the number of numbers to 100, and make them correspondingly larger, with all known methods, the computation time to find two subsets of with the same sum is enormous.

20 questions game. Now, to another application. Consider the game 20 questions. Here, one player thinks of an object, and the second player asks yes/no questions until she guesses it. We'll ask for the largest number of distinct objects that can be identified in such a game. One thing to recognize is that the strategy is adaptive: the k'th question may depend on the previous $k-1$ questions. Consider a strategy of the guessing player. On her first question, she may ask the traditional question "is it bigger than a breadbox?" If the answer is "yes," it would be reasonable to continue: "is it smaller than an elephant?" while if the first answer is "no," she will want to ask a different question. However, we restrict her to follow a deterministic strategy: if she gets a "yes" answer to the first question, we assume she always asks the same second question[1].

Even though the second question may depend on the answer to the first, if our player is to identify an object from our set of objects definitively, there cannot be two objects in our set which give the same answers to all the questions. If there were, then the questions would have to be the same for each of these two objects, since the questions may depend only on the previous answers. Thus, when the guesser got to the end of her twenty questions, she would not be able to determine which of these objects was correct, and she would have to guess wrong for at least one of them.

Thus, for every object in our set (these objects will be the pigeons), there must be a unique sequence of yes/no answers (the pigeonholes). There are twenty questions, so there are twenty yes/no answers, each of which can have two values. The total number of possible objects is thus at most 2^{20}, or a little over a million. And this bound can be achieved as there is a strategy for the guesser to distinguish 2^{20} objects with 20 questions.

20 questions with one lie. Now, let's look at a variation of this problem, which was apparently first considered by Stanislaw Ulam (the real inventor of the H-bomb). Suppose the player who has thought of the object is allowed to answer with one lie. What is the maximum number of objects the guesser can distinguish? Again, each sequence of answers will correspond to a pigeonholes.

Consider a given object. How many different sequences of answers can be associated with the object? The answerer might not lie, or he might lie on any one of the twenty questions. Once he has lied, however, his following answers are determined since he is required to tell the truth on the remaining questions. This gives 21 distinct sequences of answers identified with any given object. We will say that each of these sequence of answers is a pigeon. Thus, if we have t objects, we have $21t$ pigeons.

We now have 2^{20} holes, and $21t$ pigeons. If $21t > 2^20$, then there are two pigeons in the same hole. These two cannot come from the same object (I'll let you figure out why), so this means that for that some sequence of yes/no answers (the pigeonhole) there would are two possible objects compatible with it. Thus the maximum number of pigeons we can have is $\lfloor 2^{20}/21 \rfloor$, or 49932. ($\lfloor x \rfloor$ denotes the greatest integer less than or equal to x, and is pronounced "floor")

Can we actually find a strategy for the guesser that can distinguish this many objects? It turns out that for some numbers of questions we can, and for the rest we can still do reasonably well. This will be discussed later in the class, when we get to the section on error-correcting codes.

Increasing/decreasing sequences. The last example I want to present has to do with permutations. A permutation of n is an ordering of the numbers from 1 to n. For example, a permuation

[1]The largest number of distinct objects which can be guessed doesn't change if we let her use a probabilistic strategy, but the argument is more complicated, and we haven't even talked about probability yet.

of 7 is

$$1\ 4\ 3\ 7\ 2\ 6\ 5\,.$$

we will prove:

Theorem: Every permutation of n either has an increasing subsequence or a decreasing subsequence of length $\lceil\sqrt{n}\rceil$.

Here $\lceil\cdot\rceil$ ("pronounced ceiling") means to round up to the next integer. For a permutation of 7, the theorem guarantees an increasing or a decreasing sequence of length at least $\lceil\sqrt{7}\rceil = 3$. The permutation above has has an increasing subsequence 1, 4, 7 and a decreasing subsequence 4, 3, 2.

Proof: We will associate to every number in the permutation an ordered pair of two integers. The first integer associated with a number k will be the length of the longest increasing subsequence ending with k, and the second will be the length of the longest decreasing subsequence ending with k. Thus, for the subsequence above, the ordered pairs will be

$$\begin{array}{ccccccc} 1 & 4 & 3 & 7 & 2 & 6 & 5 \\ \left(\begin{array}{c}1\\1\end{array}\right) & \left(\begin{array}{c}2\\1\end{array}\right) & \left(\begin{array}{c}2\\2\end{array}\right) & \left(\begin{array}{c}3\\1\end{array}\right) & \left(\begin{array}{c}2\\3\end{array}\right) & \left(\begin{array}{c}3\\2\end{array}\right) & \left(\begin{array}{c}3\\3\end{array}\right) \end{array}\,.$$

Here, the ordered pair associated with 2 is $(2,3)$ because the longers increasing subsequence ending with 2 is 1, 2, which has length 2, and the longest decreasing subsequence ending is 4, 3, 2, which has length 3.

Claim: All these ordered pairs are distinct.

First, we use the pigeonhole principle to show how this claim implies our theorem, and then we will prove this claim.

The pigeons will be the numbers 1, 2, ..., n, of the permutation, and the holes will be the ordered pairs. Let ℓ_i be the length of the longest increasing subsequence of our permutation, and ℓ_d be the length of the longest decreasing subsequence. It is clear that all the ordered pairs are less than or equal to (ℓ_i, ℓ_d), so the number of pairs is at most $\ell_i\ell_d$.

If we assume the claim that each pigeon fits into a different hole, then we have that the number of pigeons is at most the number of holes. There are n pigeons and $\ell_i\ell_d$ holes. This gives us

$$n \le \ell_i\ell_d,$$

showing that either $\ell_i \ge \sqrt{n}$ or $\ell_d \ge \sqrt{n}$. Since ℓ_i and ℓ_d are integers, we can strengthen this to $\ell_i \ge \lceil\sqrt{n}\rceil$ or $\ell_d \ge \lceil\sqrt{n}\rceil$.

Proof of Claim: Suppose that we have two numbers in our permutation s and t, with s coming before t, and let (a,b) and (c,d) be the ordered pairs associated with them.

$$\begin{array}{ccccccc} \ldots & & s & \ldots & & t & \ldots \\ \ldots & \left(\begin{array}{c}a\\b\end{array}\right) & & \ldots & \left(\begin{array}{c}c\\d\end{array}\right) & & \ldots \end{array}$$

There is an increasing sequence of length a that ends with s. If $s < t$, we can add t to this increasing sequence to obtain an increasing sequence of length $a+1$ that ends with t, showing that $c \ge a+1$. A similar argument shows that if $s > t$, then $d \ge b+1$. Thus, the two ordered pairs associated with s and t are distinct. This proves the claim.

It is a fairly easy exercise to find permutations of n that have no increasing or decreasing sequences longer than $\lceil\sqrt{n}\rceil$ for any n, showing that our theorem is the strongest possible result.