Pigeonhole Principle

Lecturer: Ankur Moitra

This 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 up their partner, gets dressed in another room. They grab some socks from the sock drawer to put on. If there are only two colors of socks, how many socks do they need to guarantee a matching pair?

The answer is three. With three socks and only two colors, there must be two socks of the same color (but with 2 socks, you might not have a matching pair). More generally, if we have more socks than the number of colors (of those socks) then there must always be 2 socks of the same color. This is and example of the pigeonhole principle in action.

1 The Pigeonhole Principle

The basic version of the *Pigeonhole Principle* states:

> *If m objects (or pigeons) are put in n boxes (or pigeonholes) and $m > n$, then at least one box contains more than one object.*

Further, a stronger version of the principle states that at least one box contains at least $\lceil \frac{m}{n} \rceil$ objects[1] This "principle" is so basic, it is natural to suspect one cannot deduce anything interesting from this basic principle. However, this suspicion is incorrect, and we show some examples below.

Two equal degrees

A graph G consists of a set V (the elements of V are called *vertices*) and a set E of pairs of vertices (the elements of E are called *edges*). The *degree* of a vertex v in a graph is the number of edges containing v.

Theorem 1. *In any finite graph, there are two vertices of equal degree.*

As a story, this means that at a party with n persons, there exist two persons who know the same number of people at the party.

Proof. For any graph on n vertices, the degrees are integers between 0 and $n-1$. Therefore, the only way all degrees could be different is that there is exactly one vertex of each possible degree. In particular, there is a vertex v of degree 0 (with no neighbors) and a vertex w of degree $n-1$ adjacent to all other vertices). However, if there is an edge (v, w), then v cannot have degree 0, and if there is no edge (v, w) then w cannot have degree $n-1$. This is a contradiction. □

[1] $\lceil \cdot \rceil$ (pronounced "ceiling") means to round up to the next integer.

Equal sum subsets

Here is a more profound application 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 $2^{30}-1$ distinct nonempty 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 $2^{30}-1 \approx 10^9$. Thus, since there are more subsets (pigeons) than sums (holes), there must be two distinct subsets A and B that have the same sum. These two subsets may not disjoint, but we can eliminate any element common to both. More formally, we can replace A and B by $A' = A \setminus B$ and $B' = B \setminus A$, and A' and B' are now disjoint subsets with the same sum. If we do this, then we get two disjoint subsets which both have the same sum. And notice that at least one of this set is non-empty (and thus the other one as well) since A and B were distinct.

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.

Monotone subsequences

The next result shows that every very long sequence of distinct real numbers contains a long subsequence which is monotone, which means the subsequence is increasing or decreasing. Consider for example the following sequence of length 12 (i.e., has 12 terms):

$$4, 3, 2, 1, 8, 7, 6, 5, 12, 11, 10, 9.$$

The subsequence $4, 8, 11$ (consisting of the first, fifth, and tenth terms) is increasing of length 3, and the subsequence $4, 3, 2, 1$ (consisting of the first four terms) is decreasing of length 4. One can check that, in the sequence above, there is no longer increasing subsequence and no longer decreasing subsequence. The following theorem implies that any longer sequence of distinct real numbers has to have a longer increasing subsequence or a longer decreasing subsequence.

Theorem 2. *Any sequence of $mn+1$ distinct real numbers $a_1, a_2, \ldots, a_{mn+1}$ has an increasing subsequence of length $m+1$ or a decreasing subsequence of length $n+1$.*

Proof. Let $t(i)$ denote the length of the longest increasing subsequence ending with a_i. It $t(i) > m$ for some i, then we have an increasing subsequence of length at least $m+1$ and we are done. So we may assume $t(i) \in \{1, 2, \ldots, m\}$ for all i. As we have $mn+1$ numbers with m possible values, by the pigeonhole principle, there must be some $s \in \{1, \ldots, m\}$ such that $t(i_j) = s$ for at least $n+1$ indices $i_1 < \ldots < i_{n+1}$. We claim that $a_{i_1} > a_{i_2} > \cdots > a_{i_{n+1}}$, so that we have a decreasing subsequence of length $n+1$ and we are done. Indeed, if this was not the case, then there is a pair such that $a_{i_j} < a_{i_{j+1}}$. We could extend the increasing subsequence of length s ending at a_{i_j} by adding the term $a_{i_{j+1}}$ at the end to get an increasing subsequence of length $s+1$ ending at $a_{i_{j+1}}$. However, this contradicts $t(i_{j+1}) = s$, which says that longest increasing subsequence ending at $a_{i_{j+1}}$ has length s. □

Exercise. *For any m and n, find a sequence of mn distinct numbers with no increasing subsequence of length $m + 1$ or a decreasing subsequence of length $n + 1$. This shows that our theorem is the strongest possible result.*

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[2].

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. 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 pigeonhole.

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 (We'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

[2]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.

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.