

Discrete Math, 17th day, Monday 8/2/04  
REU 2004. Info:  
<http://people.cs.uchicago.edu/~laci/reu04>.

Instructor: László Babai  
Scribe: Charilaos Skiadas and Eric Patterson

## 1 Puzzles

### The Monty Hall Paradox.

On a game show, there are three closed doors. There is a car behind one door. Behind each of the other two doors, there is a goat. You select one of the doors, and the game master opens a different door, behind which is a goat. Then you are offered a choice to stay with your choice or to switch. The chance that we picked the car initially is  $\frac{1}{3}$ , so the chance that we get the car if we stay with our choice is  $\frac{1}{3}$ . Thus, the strategy that switches gives us a  $\frac{2}{3}$  chance of picking the car.

### The 2 Envelope Paradox

You get two envelopes. Each of them contains some money. One of the envelopes has twice as much money as the other. You are allowed to open one of the envelopes, see how much is in it, and then choose which envelope to pick. No matter what is in the envelope that you opened, you would expect that the other envelope would give you a larger expected amount of money. But you do not need to open the envelope to determine that. By this reasoning, we should keep switching back and forth between the two envelopes to increase the amount of money we expect to get.

**Moral:** be careful with the notion of expectation and probability space.

## 2 Statistical Independence vs Linear Independence

**Problem 17.1.** *Can we create  $m$  3-wise independent non-trivial events, such that  $n = |\Omega| = 2m$ ? ( $m = 2^k$ )*

For pairwise independent events, we know that  $n \geq m + 1$ . Following the ideas in the pairwise case, define  $\Omega := \mathbb{F}_q^\ell$  and suppose  $v_1, \dots, v_t$  are vectors in  $\Omega$ . We have shown that  $v_1^\perp, \dots, v_t^\perp$

are independent events iff the  $v_i$  are linearly independent. If  $v_i \neq 0$ ,  $P(v_i^\perp) = \frac{1}{q}$  because the number of elements of a hyperplane is  $q^{\ell-1}$ . Now  $U := v_1^\perp \cap \dots \cap v_t^\perp = \text{span}(v_1, \dots, v_t)^\perp$  had dimension  $\ell - \dim(U) = \ell - r$  where  $r$  is the rank of  $(v_1, \dots, v_t)$ . So the probability of  $U$  is  $\frac{1}{q^r}$ . If the events are independent, then this has to be equal to  $\frac{1}{q^t}$ . Hence  $r = t$ , and the  $v_i$  are linearly independent. Conversely, if the  $v_i$  are linearly independent, then  $r = t$  and the events are independent.

To construct 3-wise independent events, we would need to construct vectors that are 3-wise linearly independent. In  $\mathbb{F}_2^\ell$ , we need to find  $2^{\ell-1}$  3-wise independent vectors, so  $n = 2^\ell$  and  $m = \frac{n}{2}$ . Take an affine hyperplane not passing through 0 (a shift of a subspace). For example, we could put 1 in the first coordinate and either 0 or 1 in the other coordinates for a total of  $2^{\ell-1}$  elements of  $\Omega$ .

Claim: These vectors are 3-wise linearly independent.

Taking any three of the vectors, we would need to show that any nontrivial linear combination cannot be 0. Since the only elements of  $\mathbb{F}_2$  are 0 and 1, nontrivial combinations are sums of one, two, or all three of the vectors. This means (i) no one of them is 0, (ii) no two of them add up to 0, and (iii) no three of them add up to 0. The vectors we chose begin with coordinate 1, so they are not zero. If  $v + v' = 0$ , then  $v = -v' = v'$ , but we assumed that we did not take two identical vectors. Any three of them add to a vector with first coordinate 1, so the sum is not equal to 0. If  $m$  is not a power of 2, we can still get that  $n < 4m$  by taking the smallest  $\ell$  such that  $m < 2^\ell$ .

### 3 Algorithmic Application

A **Boolean function** is a function  $\{0, 1\}^n \rightarrow \{0, 1\}$ . A **Boolean formula** is a formula composed of literals (variables and their negations) using AND and OR operations. For instance,  $\bar{x}_1 \vee (x_2 \wedge x_1 \wedge \bar{x}_2)$ . A **disjunction** is an OR of literals. A **CNF (conjunctive normal form)** is an AND of clauses, each of which is a disjunction.

**Exercise 17.2.** Every Boolean formula can be represented as a CNF.

A 3-CNF formula is a formula in which every clause has 3 literals. To evaluate a Boolean formula on some substitution of values of the variables, recall that an OR of two variables is zero if and only if both variables are zero, and an AND of two variables is one if and only if both variables are one. An assignment of values to the variables in a Boolean formula **satisfies** the formula if the substitution of the values returns 1.

**Theorem 17.3.** *Satisfiability of 3-CNF formulae is NP-complete.*

Claim: For a 3-CNF formula, there always exists a substitution that satisfies at least 7/8 of the clauses.

Let  $n$  be the number of variables and  $m$  be the number of clauses.

Hint 1: Flip coins for the value of each variable; that is, make a random substitution. If each set of values is an element of  $\Omega$ , then  $|\Omega| = 2^n$ .

Hint 2: Find the expected number of satisfied clauses.

Let  $X$  be the number of satisfied clauses, so  $X$  is a random variable from the uniform probability space  $\Omega$ . We can write  $X = \sum_{i=1}^m \vartheta_i$ , where  $\vartheta_i$  is the indicator of the event that the  $i$ th clause is satisfied. Then the expected value is the sum of the expected values of the  $\vartheta_i$ . Since the  $\vartheta_i$  are indicator variables,  $E(\vartheta_i)$  equals the probability that the clause  $C_i$  is satisfied. By the rules for evaluating Boolean formulae, the probability is  $\frac{7}{8}$  that a disjunction of 3 literals with random variables is satisfied. Therefore, the expected number of satisfied clauses is  $\sum_{i=1}^m \frac{7}{8} = \frac{7m}{8}$ . Hence there exists an outcome  $x$  such that  $X(x) \geq \frac{7}{8}m$ . So there exists a substitution that satisfies at least  $\frac{7}{8}$  of the clauses.

If we want to find such a substitution deterministically in polynomial time, we should notice that we only used the fact that the variables occurring in a clause are 3-wise independent. Hence we can replace our space of outcomes with a space of size less than  $4n$  by the construction for finding 3-wise independent vectors above. This gives us a tiny fraction of all the substitutions, but the expected number of satisfied clauses is the same. Now we can simply try everything in this space, which will finish in quadratic time.

**Moral:** It is worth fighting for small sample space.

Recall the following exercise: if  $X_1, \dots, X_m$  are 4-wise independent nonconstant random variables, then  $n = |\Omega| \geq \binom{m}{2}$ .

**Proof:** Without loss of generality, we can assume that the expected value of each random variable is 0. The space of random variables has dimension  $\dim \mathbb{R}^\Omega = n$ . We want to construct  $\binom{m}{2}$  random variables that will be linearly independent in this space. Look at the pairwise products of the  $X_i$ .  $\square$

**Exercise 17.4.** Prove that the  $\binom{m}{2}$  products  $X_i X_j$  for  $i \leq j$  are linearly independent.

## 4 Algebraic Coding Theory

Question: what is the maximum number of  $k$ -wise linearly independent vectors in  $\mathbb{F}_q^\ell$ ?

A **codeword** is a sequence  $(\alpha_1, \dots, \alpha_n) \in \mathbb{F}_q^\ell$ . When you transmit the codeword, there is some noise, that is, some of the entries in the codeword might change value. If we have a given set of codewords from  $\mathbb{F}_q^\ell$ , we want to be able to distinguish them from one another even after some reasonable amount of noise interference. The **Hamming distance** between two codewords is the number of substitutions that must be made to change one into the other. If the Hamming distance is large among the codewords, the codewords will still be distinguishable after some interference. The **code space** is a subspace  $U \leq \mathbb{F}_q^\ell$ . We want  $\dim U$  to be large (i.e., lots of codewords) and the Hamming distance between codewords to be at least something.

**Exercise 17.5.** Establish what this something is and relate it to the  $k$  above, i.e., establish a connection between the two data above.