Discrepancy Theory

Suppose we have a probability distribution and generate a set S of n random points according to the distribution. If we have a region and we integrate the probability distribution over this region, we would like the number of points in the region to be proportional to the size of the integral (i.e. if value of the integral is $\frac{1}{4}$, we would expect $\frac{1}{4}$ of the points to be inside the region). In symbols, we have:

$$\left|\frac{S \cap R}{n} - p(R)\right| < \epsilon$$

where $p(R)$ is the integral of the probability distribution, $\frac{S \cap R}{n}$ is the fraction of points in the rectangle, and ϵ is some constant that depends on the VC dimension of the shape of the region. Therefore, if we use shapes with a low VC dimension, we are in good shape. We can prove this for rectangles, but what about arbitrary regions?

Theorem: For any rectangle, we want

$$\left|\frac{S \cap R}{n} - p(R)\right| < \epsilon$$

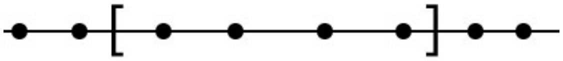
to be true with probability $1 - \epsilon$. Outline of the proof:

1. Prove the statement for one rectangle.

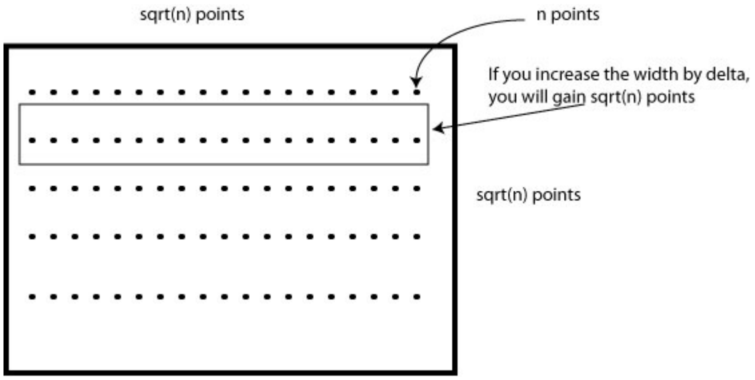
2. Define two rectangles to be in the same equivalence class if they contain the same points.

3. Sum over polynomial number of equivalence classes of rectangles.

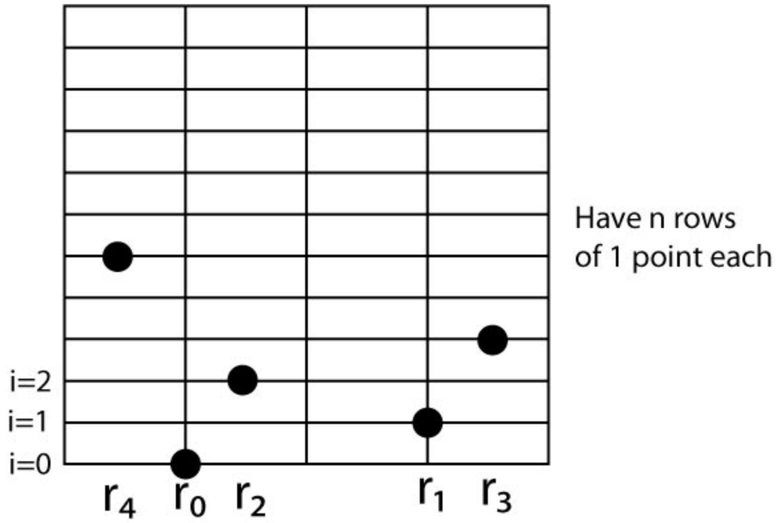
1 dimensional case: We want the proportion of points approximately equal to the area of the integral (if we put n points down, we want an area close to n).



2 dimensional case: We want the area of the rectangle to be proportional to the number of points. We can get a discrepancy of $\pm \log n$.


sqrt(n) points
n points
If you increase the width by delta, you will gain sqrt(n) points
sqrt(n) points

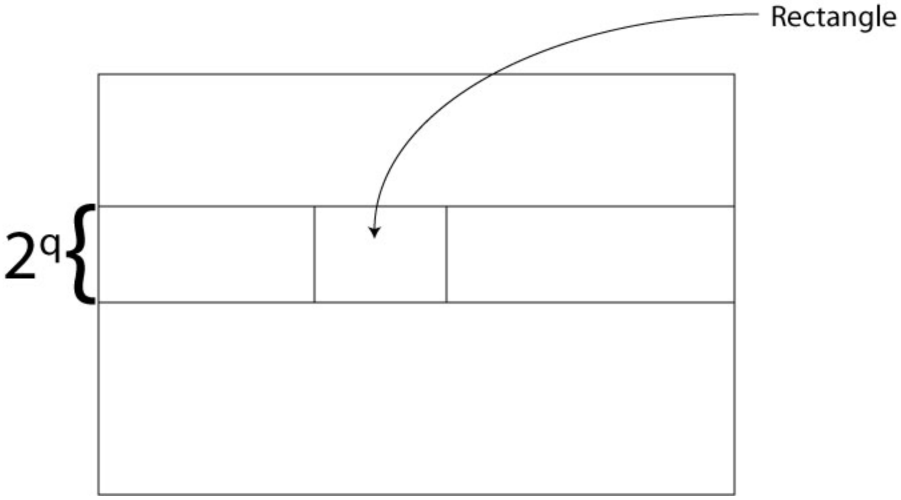
A better way of doing this is the point the points on a grid. We will have n rows of 1 point each. We set up the grid by continuously drawing lines halfway through and place points alternating from side to side. If we write i in binary notation, we have $i = a_0 + 2a_1 + 4a_2 + \ldots$ and $r_i = \frac{a_0}{2} + \frac{a_1}{4} + \frac{a_2}{8} + \ldots$. We place points at (i, r_i). We want to prove that this has very low discrepancy.

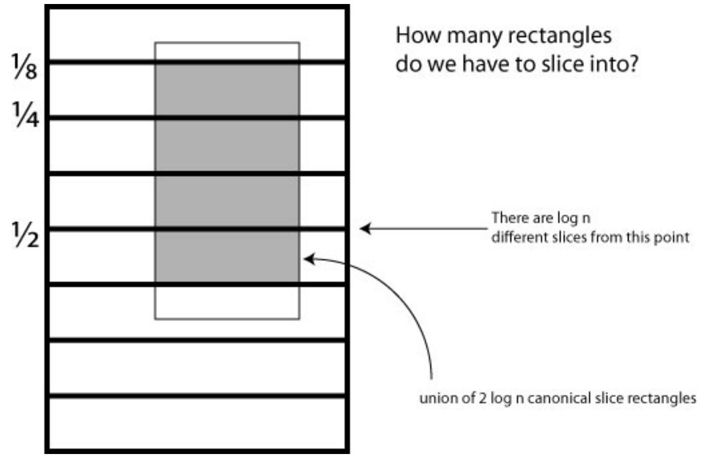


Theorem: There exists an n-point set p where the discrepancy for axis-parallel rectangles is $O(\log n)$.

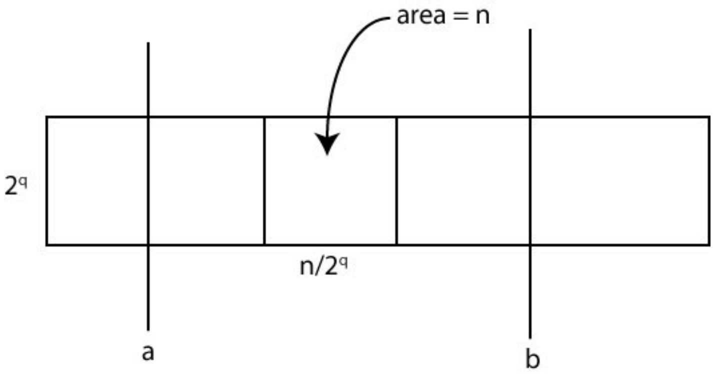
Notation: To simplify notation, consider the unit square to be an $n \times n$ lattice where points are indexed by pairs of integers. We start with $i = 0$ instead of $i = 1$. Let $p = \{(i, r_i) | 0 \leq i < n\}$.

Proof: Consider an interval $I = [k2^q, (k+1)2^q]$. For the canonical rectangle, the discrepancy is at most 1.




1/8
1/4
1/2
How many rectangles do we have to slice into?
There are log n different slices from this point
union of 2 log n canonical slice rectangles

Each canonical slice contains 2^q points and each piece of the slice is of width $\frac{n}{2^q}$. This gives each piece an area of n. We want to argue that there is one point in each piece. We can see that this is the case since we have 2^q rectangles and 2^q points.



Data Streams

Consider a situation where we have a large stream of data that is impossible to store completely (e.g. a stream of all grocery store purchases in the USA). What if we would like to know how many distinct numbers we have seen in the stream? If we have m distinct numbers, we cannot save them in less than m bits of space.