5. Random Sampling and Arrangement of Lines

A central concept of *statistics* is
A random sample is a good estimator for statisical population

The concept of randomized divide-and-conquer Quick-Sort

- Let N be any set of points in the real line.
- If we pick a random element S from N, then S probably divides the line into interval of roughly equal size. The size mean the number of unchosen points lying in the interval.

Random Sampling without replacement

- Given a set N of objects, a r-element subset R of N is a random sample if every element in N is equally likely to be in R.
  - Choose the first element in R randomly from N
  - Choose the seond element in R from the remaining $n - 1$ elements independently and randomly.
  - Repeat the process untail r elements from N are chosen.

An interesting and important question:
Given a set N of n points in the real line, does a random sample R of N of size r divide the real line into roughly equal size?

- Let $H(R)$ be the partition of the real line formed by R.
- For each interval I in $H(R)$, the conflict size of I is the number of points in $N \setminus R$ lying in I.

Is the conflict size of each interval in $H(R)$ $O(n/r)$ with high probability?
Most researchers conjecture the positive answer, but no one can prove it over several centries.

Main Theorem

For a set N of n points on the real line and a random sample R of N of size r, with probability greather than 1/2, the conflict size of each interval in $H(R)$ is $O([n/r] \log r)$.

More generally, for any fixed $c > 2$ and any $s \geq r > 2$, with probability $1 - O(1/s^{c-2})$, the conflict size of each interval in $H(R)$ is less than $c(n \ln s)/(r-2)$. In other words, the probability of some conflict size exceeding $c(n \ln s)/(r-2)$ is small, $O(1/s^{c-2})$ to be precise.

Proof of Main Theorem

Terminology

- $\Pi = \Pi(N)$ is the set of all pairs the form (p, q) where p, as well as q, is a point in N or a point at infinity.
- A point at infinity means either $-\infty$ or $+\infty$
- σ is any such pair in Π, and thus defines an interval on the real line.
- $D(\sigma)$ is $\{p, q\} \cap N$, and consists of the endpoints of σ not at the infinity. The points in $D(\sigma)$ is said to define σ.
- $d(\sigma)$ is the size of $D(\sigma)$ and is called the *degree* of σ. $d(\sigma)$ is 0, 1, or 2.
  - $d((p, q)) = 2$, $d((-\infty, p))$, and $d((-\infty, +\infty))$.
- $L(\sigma)$ is the set of points in N that lies in the interior of σ. The points in $L(\sigma)$ is said to conflict with σ
- $l(\sigma)$ is the size of $L(\sigma)$ and called the *conflict size* of σ.
- Π is a configuration space of N
  - An interval $\sigma \in \Pi$ is *active* over a subset $R \subseteq N$ if σ is an interval of $H(R)$
  - σ is an interval of $H(R)$ if and only R contains all points in $D(\sigma)$ but no poin in $L(\sigma)$.

Conditional Probability

- Let $R \subseteq N$ denote a random sample of N of size r.
- Let $p(\sigma, r)$ denote the *conditional probability* that R contains no point in conflict with σ, given that it contains the points defining σ.

Claim

$$p(\sigma, r) \leq (1 - \frac{l(\sigma)}{n})^{r-d(\sigma)}$$

Intuition

- Since R must contain $D(\sigma)$, the remaining $r - d(\sigma)$ can be thought of as resulting from independent random draws.
- The probability of choosing a conflicting point in any such draw is greater than or equal to $l(\sigma)/n$.

Rigorous justification

- Let R' be $R \setminus D(\sigma)$
- R' is a random sample of the set $N' = N \setminus D(\sigma)$ of size $n - d(\sigma)$
- R' is obtained from N' by $r - d(\sigma)$ successive random drwas without replacement.
- For each $j \geq 1$, the probability that the point chosen in the j^{th} draw does not conflict with σ, given that no point chosen in any previous draw conflicts with σ, is

$$1 - \frac{l(\sigma)}{n - d(\sigma) - j} \leq 1 - \frac{l(\sigma)}{n}.$$

- Then

$$p(\sigma, r) = \prod_{j=1}^{r-d(\sigma)} 1 - \frac{l(\sigma)}{n - d(\sigma) - j} \leq (1 - \frac{l(\sigma)}{n})^{r-d(\sigma)}$$

Proof of Main Theorem(continue)

- Since $1 - l(\sigma)/n \leq \exp(-l(\sigma)/n)$, the claim implies

$$p(\sigma, r) \leq \exp(-\frac{l(\sigma)}{n}(r - d(\sigma)),$$

  where $\exp(x)$ denotes e^x.

- Since $d(\sigma) \leq 2$,

$$p(\sigma, r) \leq \exp(-\frac{l(\sigma)}{n}(r - 2)).$$

- If $l(\sigma) \geq c(n \ln s)/(r - 2)$, for some $c > 1$, then

$$p(\sigma, r) \leq \exp(-c \ln s) = \frac{1}{s^c}.$$

Combined probability

- Let $q(\sigma, r)$ denote the probability that R contains all points in $D(\sigma)$.
- The probability that σ is active over R is precisely $p(\sigma, r)q(\sigma, r)$.

The probability that some $\sigma \in \Pi$, with $l(\sigma) > c(n \ln s)/(r-2)$, is active over R is bounded by

$$\sum_{\sigma \in \Pi : l(\sigma) > \frac{cn \ln s}{r-2}} p(\sigma, r)q(\sigma, r) \leq \sum_{\sigma \in \Pi : l(\sigma) > \frac{cn \ln s}{r-2}} q(\sigma, r)/s^c \leq \frac{1}{s^c} \sum_{\sigma \in \Pi} q(\sigma, r).$$

Summary

- Let $\pi(R)$ denote the number of intervals in Π whose defining points are in R.
- $\sum_{\sigma \in \Pi} q(\sigma, r)$ is $\pi(R)$.
- For a random sample R of N, the probability that some $\sigma \in \Pi$, with $l(\sigma) > cn \ln s/(r - 2)$, is active over R is bounded by

$$\frac{1}{s^c} E[\pi(R)].$$

- Since R has r points, $\pi(R) = \binom{r}{2} + 2r + 1 = O(r^2)$.

$$\frac{1}{s^c} E[\pi(R)] = O(\frac{r^2}{s^c}) = O(\frac{1}{s^{c-2}}).$$

Arrangement
Given a set N of hyperplane in $\mathbb{R}^d$, the arrangmenet $G(N)$ formed by N is the natural partition of $\mathbb{R}^d$ by N into faces of varying dimensions together with the adjacencies among them.

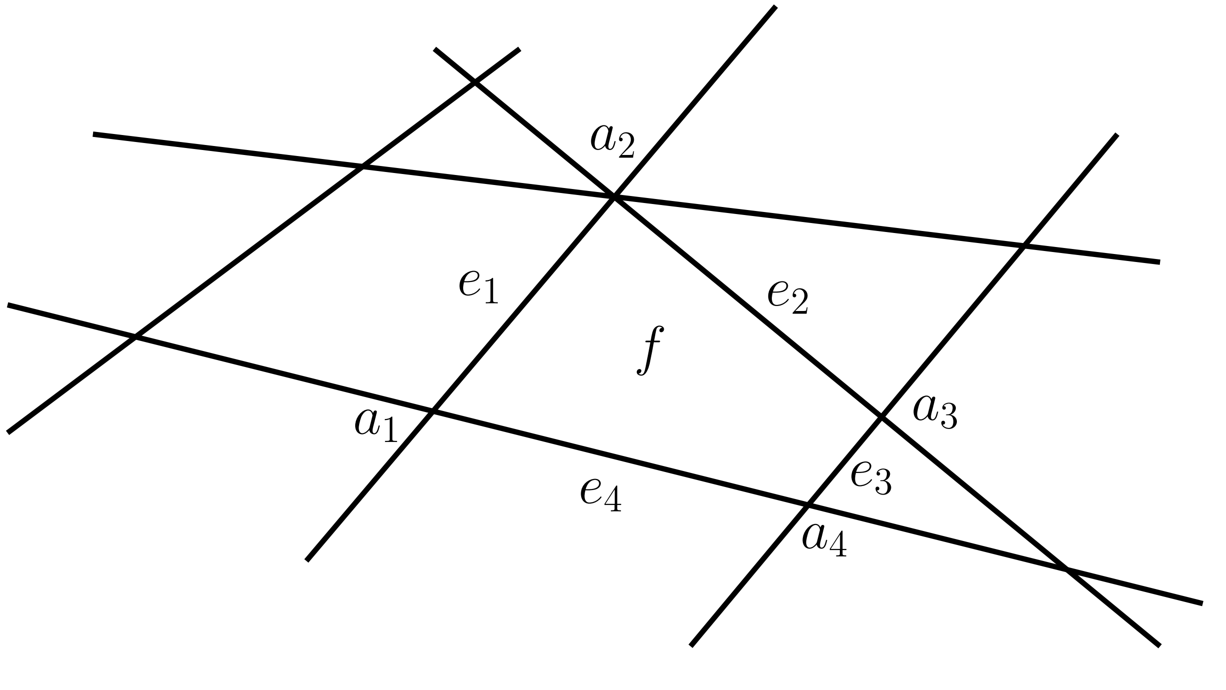
- A face of j dimensions is called a j-face
- A d-face is called a cell
- A $(d-1)$-face is caleld a facet
- A 1-face is called an edge
- A 0-face is called a vertex

General Position Assumption

- No two hyperplane are parallel to each other
- For $2 \leq j \leq d+1$, the intersection among j hyperplane is exactly a $(d+1-j)$-face

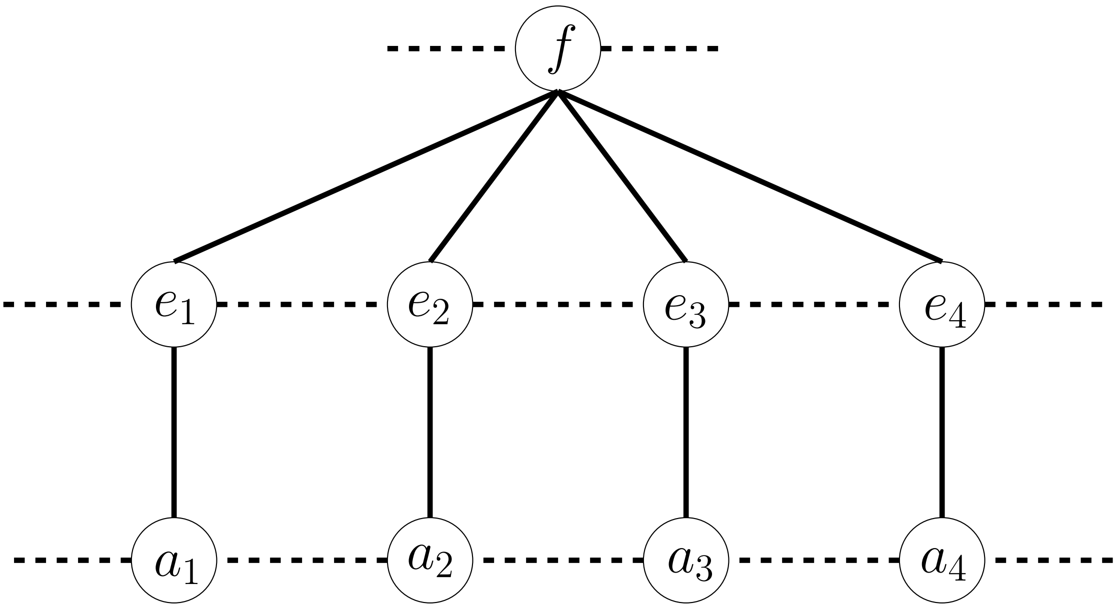
Arrangement in the plane
An arrangement of n lines is one of the simplest geometric structure

- $O(n^2)$ faces in total



Facial lattice of an arrangement

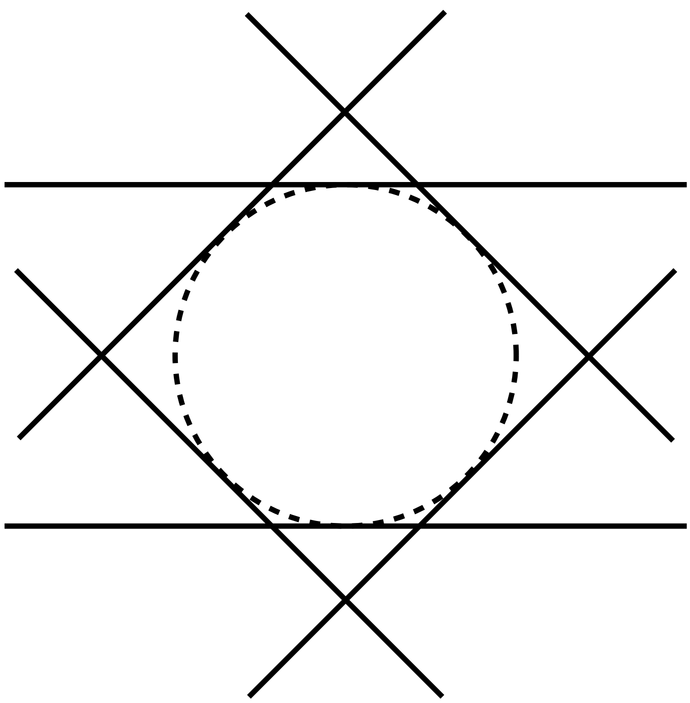
- The lattice contains a node for each face of $G(N)$
- Each node contains auxiliary information, such as pointers to the hyperplanes containing the corresponding face
- A node for a j-face f is linked to a node for a $(j-1)$-face g if f and g are adjacent



Fact

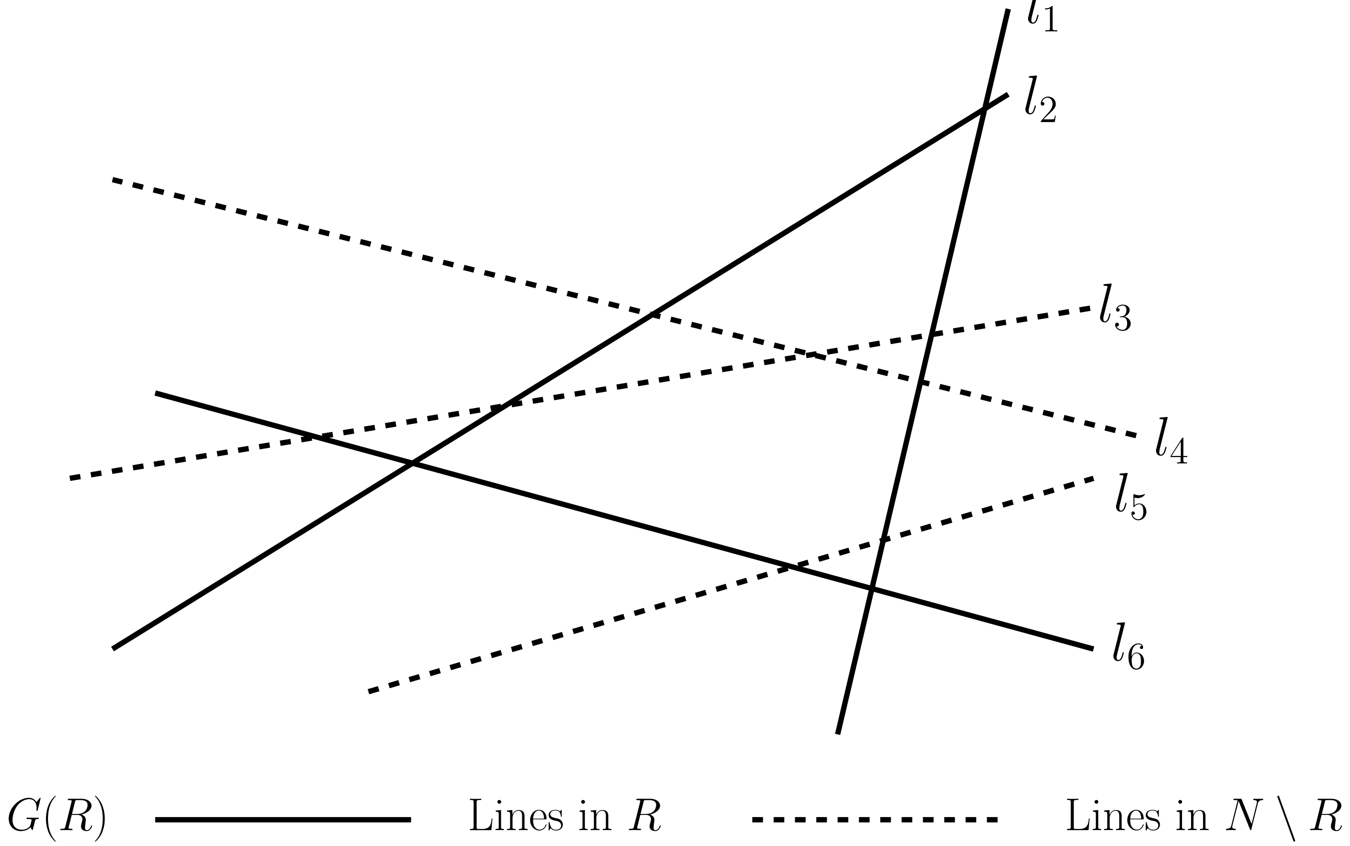
Cells of an arragement of lines in the plane does not allows the random sampling technique

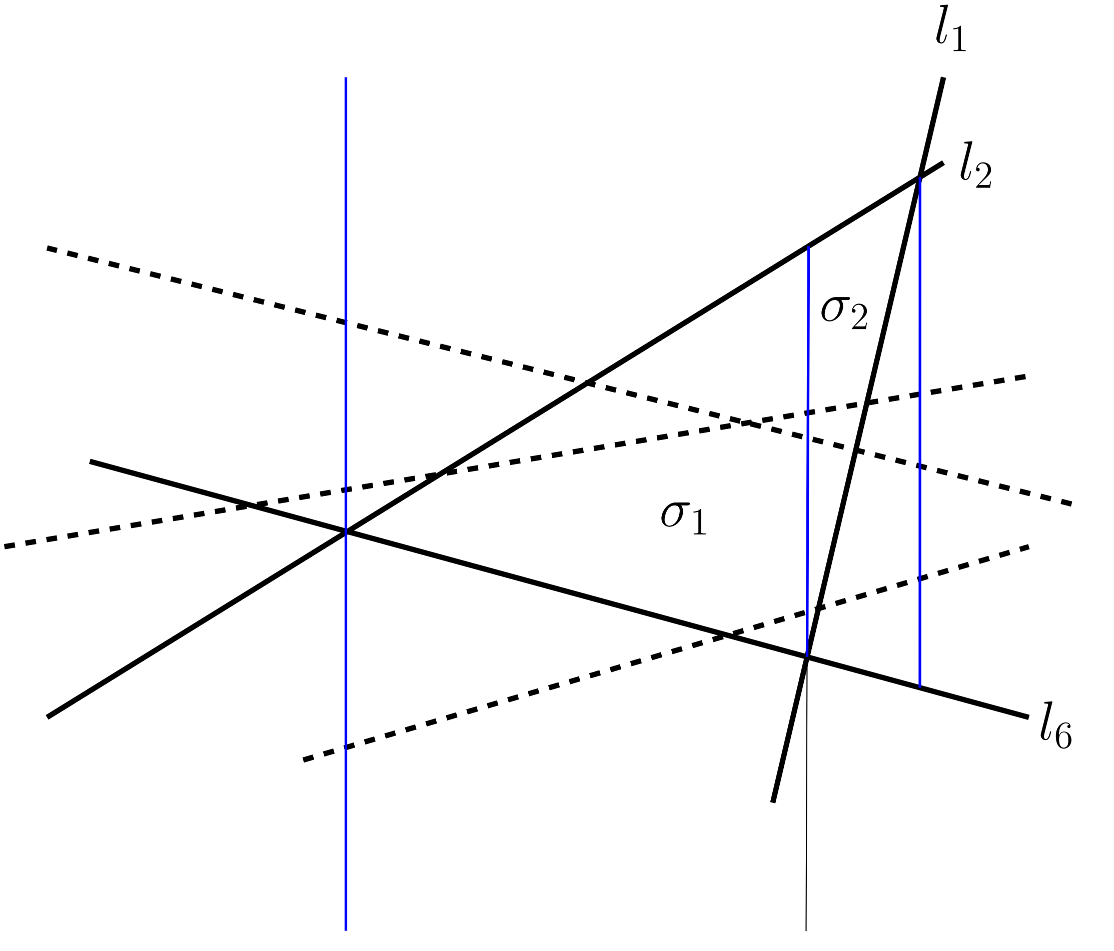
- When all lines in N are tangent to the same circle, for any subset R of N, the central cell of the arragnement of R is intersected by all lines in $N \setminus R$.



A cell of an arrangment $G(R)$ does not satisfied the *bounded degree property.*

That is, the collection of cells is not a configuration space.





$H(R)$: the vertical trapezoidal decomposition of $G(R)$

Bounded Valence

A configuration space $\Pi(N)$ is said to have *bounded valence* if the number of configurations in $\Pi(N)$ sharing the same trigger sets is bounded by a constant

General Form for Main Theorem

Given a set N of n objects, a configuration space $\Pi(N)$ of N with bounded valance, and the maximum degree d of a configuration in $\Pi(N)$, for any random sample R of N of size r, with probability greater than $1/2$, the conflict size for each active configurations over R is at most $c(n/r)\log r$, where c is a large enough constant.

More generally, fixed any $c > d$, for any $s \geq r > d$, with probability 1-$O(1/s^{c-d})$, the conflict size of each active configuration over R is less than $c(n \log s)/(r-d)$

Sketch of Proof:

For the same reasoning, we have the following fact.

Fact

The probability that some $\sigma \in \Pi(N)$, with $l(\sigma) \geq c(\ln s)/(r-d)$, is active over a random sample R is bounded by $E[\pi(R)]/s^c$, where $\pi(R)$ is the number of configurations in $\Pi(N)$ whose defining objects are in R.

$\pi(R) = O(r^d)$

- For each $b \leq d$, there are at most $\binom{r}{b} \leq r^b$ trigger sets contained in R
- Since $\Pi(N)$ has bounded valence, only a constant number of configurations in $\Pi(N)$ share the same trigger set.

$$\frac{E[\pi(R)]}{s^c} = O(\frac{r^d}{s^c}) = O(\frac{1}{s^{c-d}}).$$