Fingerprinting ctd.

Alice —sensitive message (1000 bits)→ Bob

- via internet is insecure because evil spies will eavesdrop
- if they have 'unlimited' comp. power, schemes like RSA are not safe enough
- so last Alice and Bob met in person, they threw a coin 2000 times and wrote down the bit sequence

⇒ Alice uses the first 1000 secret bits to encode her message via bit-wise XOR, Alice sends the result to Bob and Bob decrypts it accordingly

BUT: spy at Bob's home
Bob kills him after 200 bits were transfered to the evil agency

If there are 200 consecutive bits of the secret sequence, Alice and Bob could just reuse bits to encode.

⇒ Bob observes, that the 200 bits are about the secret sequence, not a cut-out

What to do now?

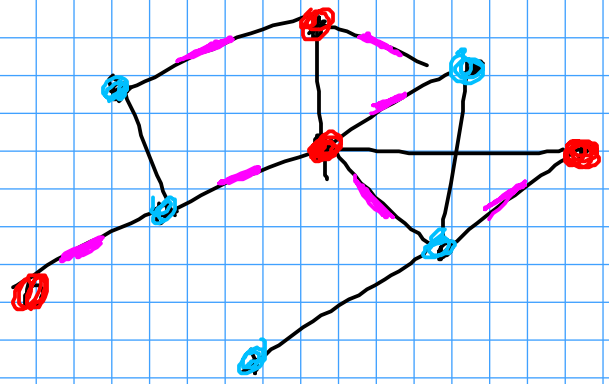
Alice constructs a hash function $h : \{0,1\}^{2000} \to \{0,1\}^{1000}$ to hash the secret sequence. Alice sends h and her encrypted message (xor with hash value) to Bob, Bob decrypts.

⇒ By hashing all information gained by the spy is shredded. We expect the evil agency to learn only

$$2^{1000 - 2000 + 200} = 2^{-800} \text{ bits}$$

if h is drawn u.a.r. from a universal hash family H, so for a practical purposes, this is zero.

2.4.4. De-randomized MaxCut via Hashing



Max Cut

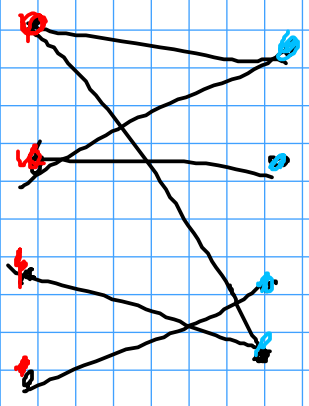
Given a graph $G(V, A)$, partition V into two groups of vertices (red, blue), such that the number of edges with differently colored vertices is maximized.

More formally

$$\phi : V \longrightarrow \{red, blue\}$$

$$c(\phi) = \left| \{\{v, w\} \in A \mid \phi(v) \neq \phi(w)\} \right|$$

cut size



NP-hard for general graphs

Monte Carlo Alg.:

- assign each vertex the color blue or

red each with a prob. of $\frac{1}{2}$

X ... number of edges in the cut

$$E(X) = \sum_{\{u,w\} \in A} P(\phi(u) \neq \phi(v))$$

$$= \sum_{\{v,w\} \in A} \frac{1}{2} = \frac{|A|}{2}$$

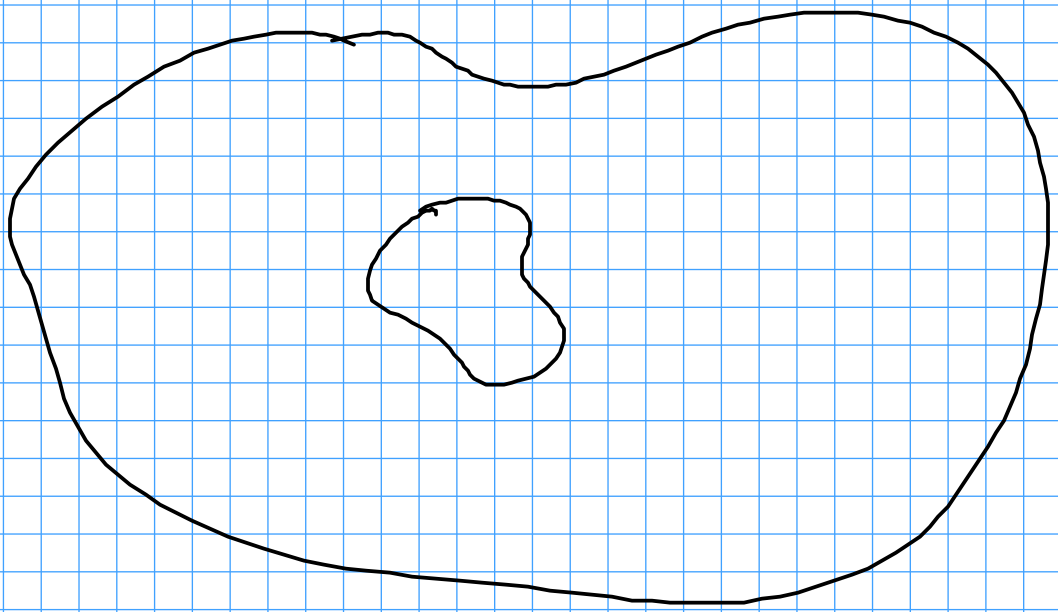
$\Rightarrow$ in expectation we have a factor 2 - approximation

$\Leftrightarrow$ Now let H be a universal hash family that maps V to $\{0,1\}$. Use result if we choose $h \in H$ u.a.r. we expect a cut of size $\frac{|A|}{2}$ (analysis from above applies here, too, because of universality of H). Now remember that we construct H such that $|H|$ is polynomial in the size of the universe.

Therefore we can de-randomize the algorithm by simply trying all hash functions in H and then return the result with maximum C-value.

$\Rightarrow$ polytime alg. (deterministic!) which guarantees a 2-APX

2.5. Random Sampling, VC-dimension and ε-nets



Basic Principle: very very very very $(very)^{1000}$ large universe (too expensive to consider as a whole), want to have small but representative subset to perform computations on and then extrapolate the results

2.5.1 Basic Sampling Theorem

Let U be our universe of elements (people, objects, points) and $S \subseteq U$ a not too small subset. Then a

random sample of U is likely to intersect S.

Thm: $S \subseteq U$ is subset of U with $|S| \geq \varepsilon \cdot |U|$ $\varepsilon \in]0,1[$. Then a random sample R of size $\frac{1}{\varepsilon} \ln \frac{1}{\delta}$ from U intersects S with prob. of $1-\delta$.

Proof. The prob. for a single element u.a.r. chosen from U to be in S is ε, and not to be in S is $1-\varepsilon$.

So $P(R \cap S = \emptyset) = (1-\varepsilon)^{\frac{1}{\varepsilon} \ln \frac{1}{\delta}}$ ($|R|$)

As $\left(1-\frac{1}{n}\right)^n$ converges to $\frac{1}{e}$,

$a^{km} = (a^k)^m$

we can rewrite our term as

$$P(R \cap S = \emptyset) \leq \left(\frac{1}{e}\right)^{\ln \frac{1}{\delta}}$$

$$= e^{\frac{1}{\ln \frac{1}{\delta}}} = \frac{1}{\frac{1}{\delta}} = \delta$$

Therefore $P(R \cap S \neq \emptyset) \geq 1-\delta$

//

Example (Traffic control) In Germany there are 32 million drivers. Every year about 0.5 million people are banned from driving. (ouling minor offenses and unreported cases, it's somehow safe to assume that at least 4 million drivers are speeders.

How many drivers should the police check to be 50% to catch a speeder?

ε-value: $\frac{|U|}{|S|} = \frac{32 \text{ million}}{4 \text{ million}} = \frac{n}{8} = \varepsilon$

$1 - d = 0.9$
$d = 0.1$
$\frac{n}{d} = 10$

$8 \cdot \ln 10 \approx 19$

Exercise How many randomly chosen voters should you consider today in Germany in a survey to be 80% sure to include an FDP-supporter?

But in general we are not only interested in a single subset S of U but a family of subsets $S_1, \dots, S_m$.

Naively, we could combine random samples for every S_i to have the desired result.

But if m is large or subsets overlap to great extend, this obviously is not a good idea.

2.5.2 Hitting Sets and Universal Sampling

We call a sample universal if its size is independent of $|U|$. Now given a family of different sized subsets $S_1, \dots, S_m \subseteq U$. Can we find a universal sample that hits every S_i?

This problem is classically modelled via the Hitting Set formulation.

Def. (HS): Let $(U, \mathcal{S})$ be a set system, i.e. U is a universe and $\mathcal{S}$ a collection of subsets of U. Find a subset $R \subseteq U$, such that

$$\forall S \in \mathcal{S}: R \cap S \neq \emptyset \text{ and } |R| \text{ is minimized.}$$

This problem is NP-complete. It is moreover hard to approximate better than $\log n$ with $n = |U|$.

So in general, no universal sample can be found. If we restrict the sets to size $\leq k$, there exists a k-approximation. But again, if S^* is the largest of the subsets and $|S^*| = \varepsilon|U|$ then we only have a $\varepsilon|U|$-APX.

BUT there are certain features of set systems which allow to get beyond the inapproximability bound of HS.

In fact the notion of VC-dim. and ε-nets describe when this is possible.

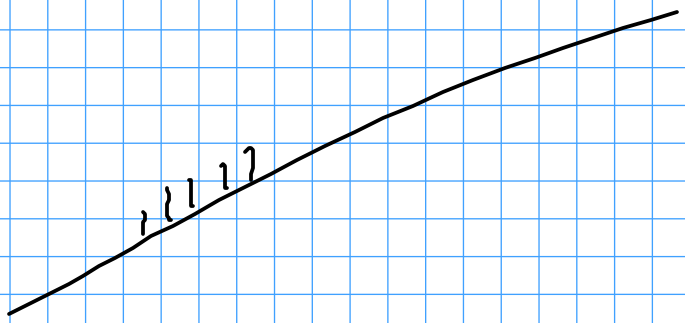
2.5.3. VC-dimension

Def. (VC-dim). The VC-dimd of a set system $(U, \mathcal{S})$ is defined as the size of the largest subset of U that can be shattered. Thereby, a subset $U' \subseteq U$ is called shattered if for any subset $A \subseteq U'$ there exists $B \in \mathcal{S}$ with $U' \cap B = A$.

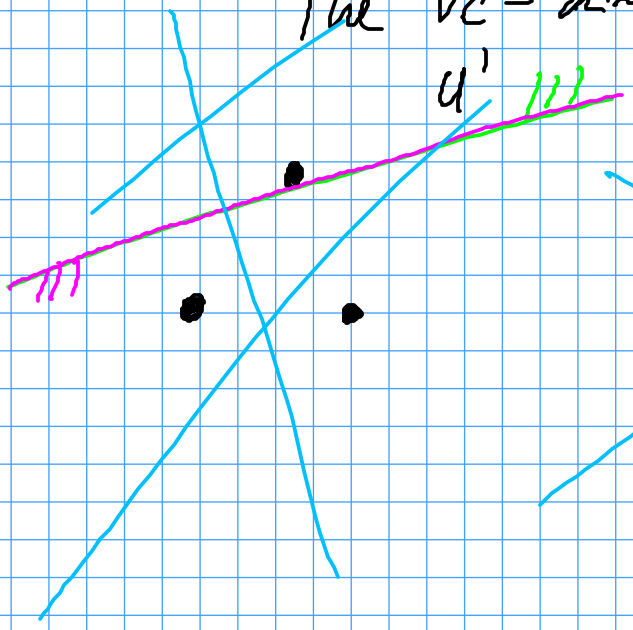
Examples (Points and Half-spaces)

U - set of points in $\mathbb{R}^2$

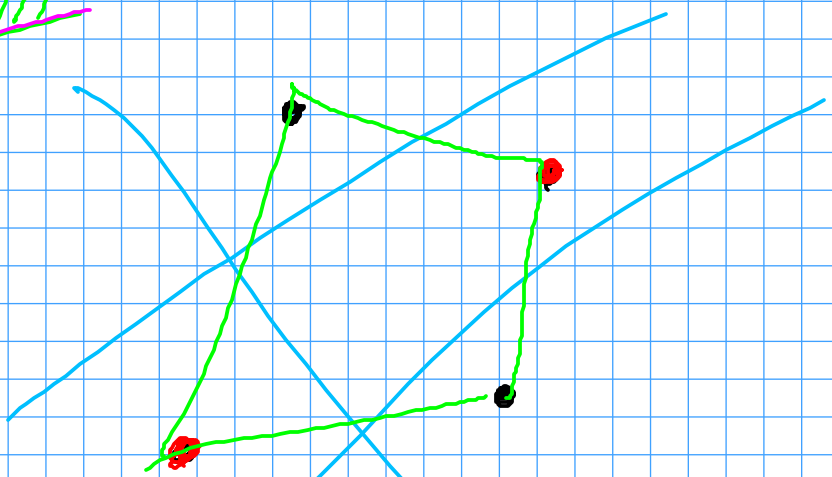
$\mathcal{F}$ - half-spaces

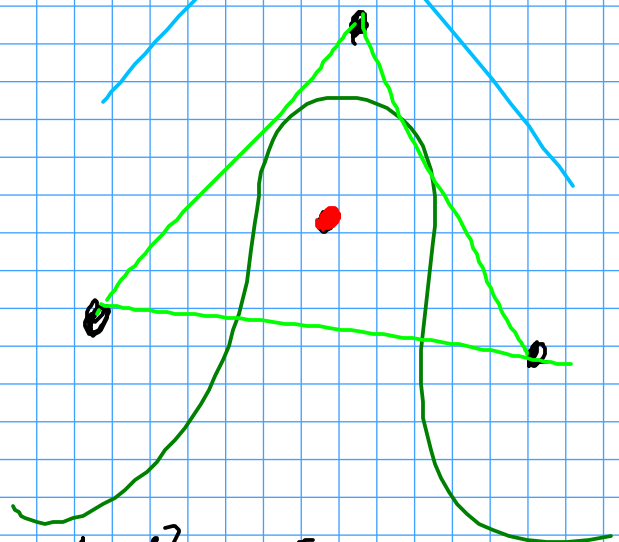


The VC-dim. of this system is 3.



U'

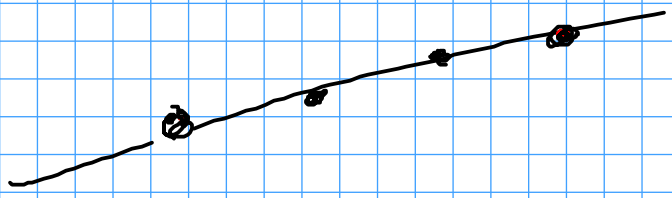




To prove consider the convex hull:

→ if it consists of
- 4 points
  - ↳ 4 points on same pos.
- ↳ as soon as 2 point share a position, the shattering possible

↳ if CH consists of 2 points



point in middle can not be sep. from end points

↳ if $|CH| = 3$ point inside can not be seperated

↳ if $|CH| = 4$ quadrangel → two diagonal points can not be seperated from the other 2

↳ $d < 4$

Exercise: Show that the VC-dim. of half-spaces in $\mathbb{R}^n$ is at most $n+1$. To show that two $n+2$ points can be shattered use Radon's theorem.

If S is a set of $n+2$ points, then S can be partitioned into S_1 and S_2 (disjoint) where convex hulls overlap.