

Fingerprinting ctd.

Alice $\xrightarrow[1000 \text{ bits}]{\text{sensitive message}}$ 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 transferred to the evil agency

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

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

What to do now?

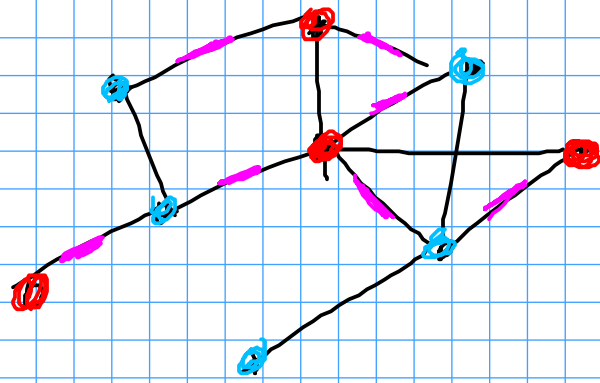
Alice constructs a hash function $h: \{0,1\}^{2000} \rightarrow \{0,1\}^{1000}$ to hash the secret sequence. Alice sends h and her encrypted message (xor with hash value) to Bob, Bob decrypts.

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

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

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

2.4.4. De-randomized MaxLat via Hashing



Max Cut

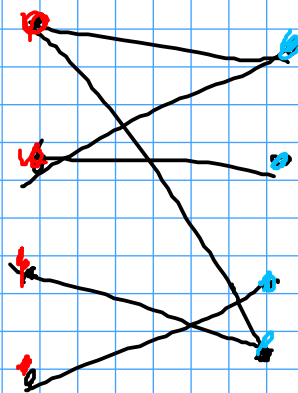
Given a graph $G(V, A)$, partition V into two groups of vertices (red, blue), such that the number of edges with differently colored vertices is maximized.

More formally

$$\phi: V \rightarrow \{\text{red}, \text{blue}\}$$

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

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

$$\begin{aligned} E(X) &= \sum_{\{u,v\} \in A} P(\phi(u) \neq \phi(v)) \\ &= \sum_{\{u,v\} \in A} \frac{1}{2} = \frac{|A|}{2} \end{aligned}$$

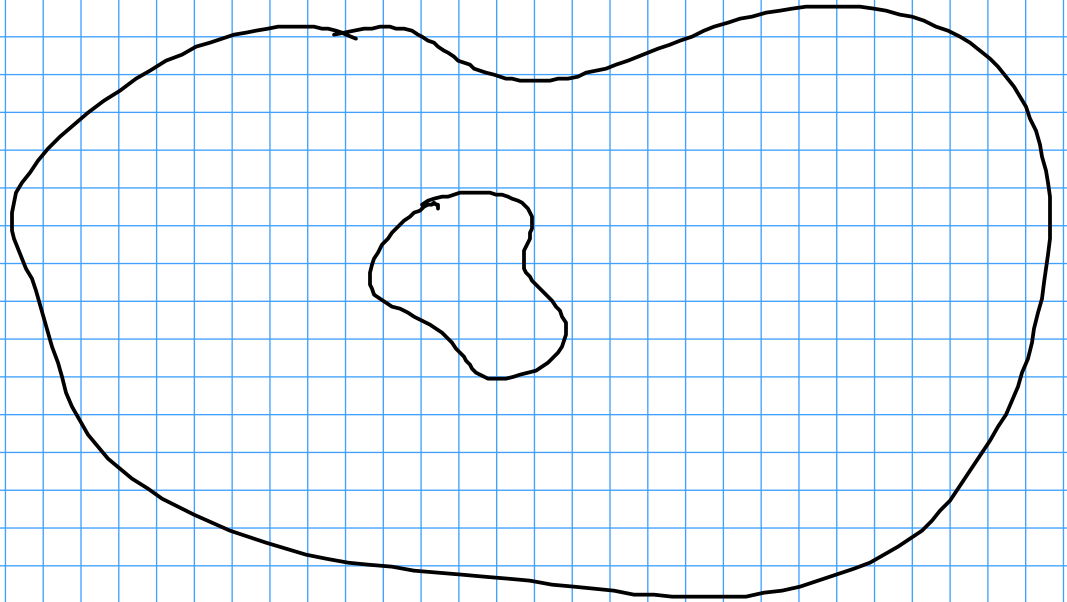
\Rightarrow in expectation we have a factor 2-approximation

\hookrightarrow Now let H be a universal hash family that maps V to $\{0,1\}$.
You read
if we choose $h \in H$ u.a.v. we expect
a cut of size $\frac{|A|}{2}$ (analysis from
above applies here, too, because of
universality of H). Now remember that
we can 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.

→ 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 $|S| \geq \epsilon \cdot |U|$ $\epsilon \in]0, 1[$. Then a random sample R of size $\frac{1}{\epsilon} \ln \frac{1}{1-d}$ from U intersects S with a prob. of $1-d$.

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

$$\text{So } P(R \cap S = \emptyset) = (1-\epsilon)^{\frac{1}{\epsilon} \ln \frac{1}{1-d}}$$

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

$$a^{nm} = (a^n)^m$$

We can rewrite our term as

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

$$= \frac{1}{e^{\ln \frac{1}{1-d}}} = \frac{1}{\frac{1}{1-d}} = 1-d$$

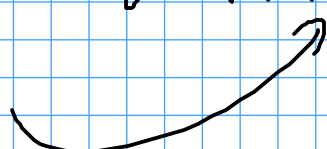
Therefore $P(R \cap S \neq \emptyset) \geq 1-d$

Example (Traffic control) In Germany there are 32 million drivers. Every year about 0.5 million people are banned from driving. Counting minor offences 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 90% to catch a speeder?

$$\epsilon\text{-value: } \frac{|U|}{|S|} = \frac{32 \text{ million}}{4 \text{ million}} = \frac{1}{8} = \epsilon$$

$$1-d = 0.9 \quad \epsilon \cdot \ln n_0 \approx 1.9$$

$$\frac{1}{d} = n_0$$


Exercise How many randomly chosen votes 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 \mathcal{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 n is large or subjects overlap to great extent, 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 modeled 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

$\forall S \in \mathcal{S}: R \cap S \neq \emptyset$ and $|R|$ 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 sampler 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 $|S^*| = \epsilon|U|$ then we only have a $\epsilon|U|$ -APX.

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

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

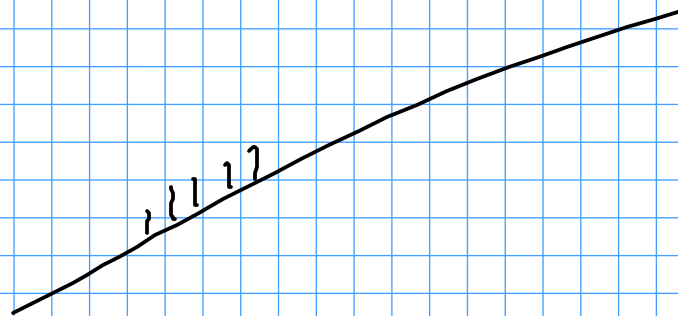
2.5.3. VC-dimension

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

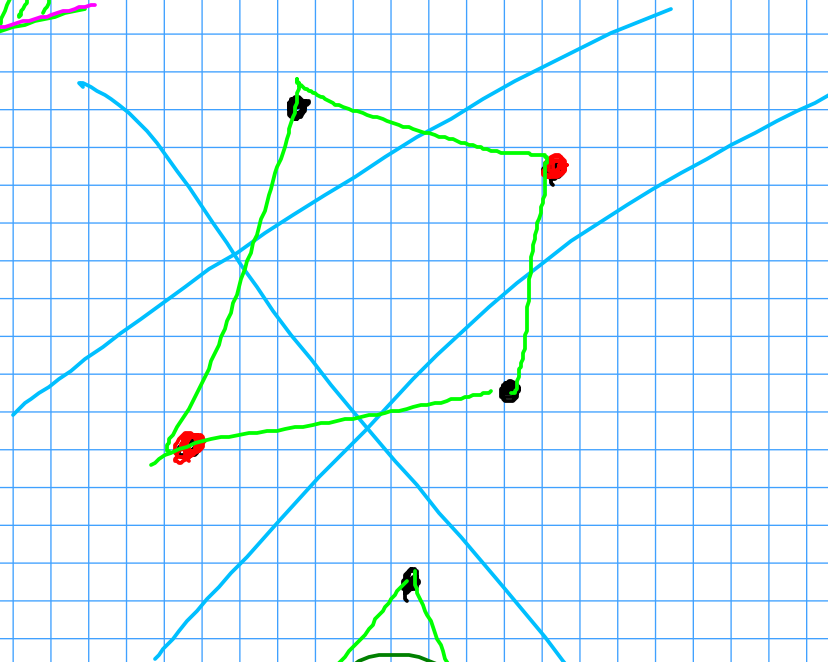
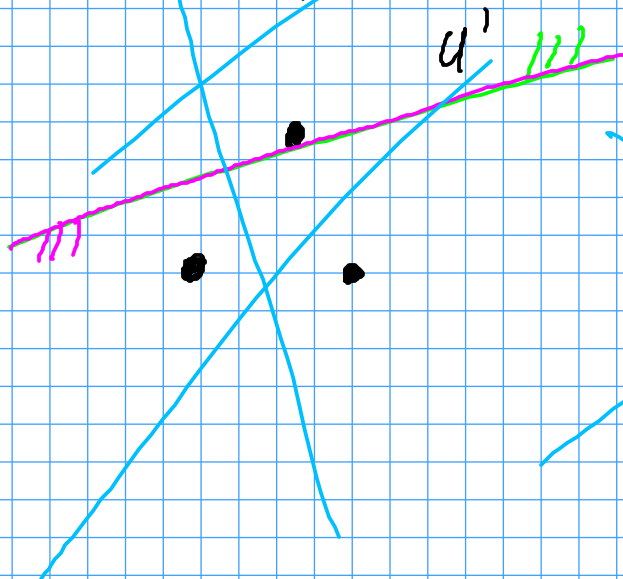
Examples (Points and Half-Spaces)

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

\mathcal{J} - half-spaces



The VC-dim. of this system is 3.

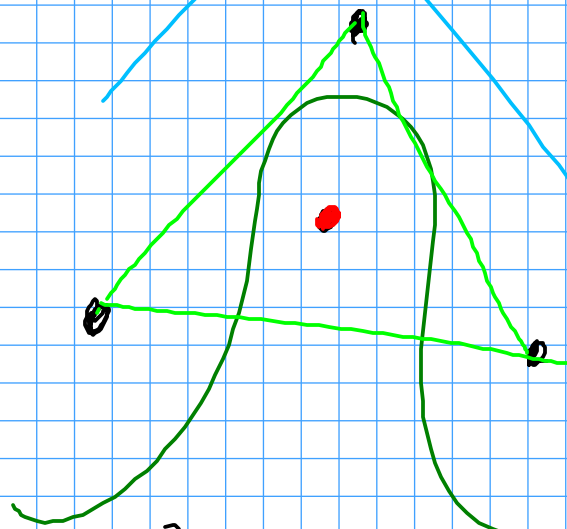


To prove consider
the convex hull:

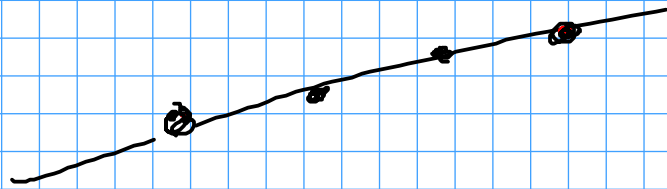
→ if it consists of
 n points

↳ 4 points on
same pos.

↳ as soon as 2 points share a position, the probability
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 separated

↳ if $|CH| = 4$ quadrangle → two diagonal points can not be separated 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 no $n+2$ points can be shattered (wlog Radon's theorem).

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