Sample Complexity Bounds for Active Learning

Paper by Sanjoy Dasgupta

Presenter: Peter Sadowski

Passive PAC Learning Complexity

- Based on **VC dimension**

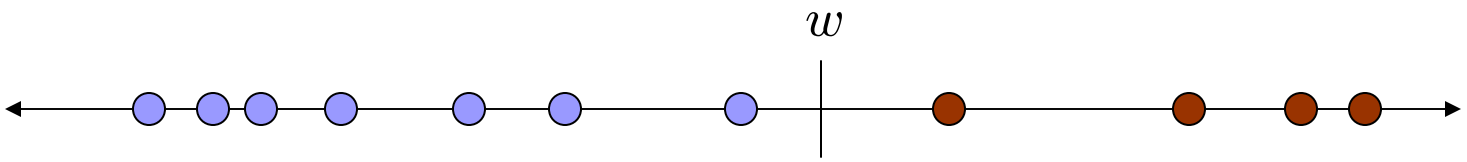
$$\text{To get error} < \epsilon \text{ with probability} \geq 1 - \delta\text{:}$$

$$\text{num samples } \geq \widetilde{O}\left(\frac{1}{\epsilon}\left(VC\left(H\right)\log\left(1/\delta\right)\right)\right)$$

Is there some equivalent for active learning?

Example: Reals in 1-D

P=underlying distribution of points
H=space of possible hypotheses

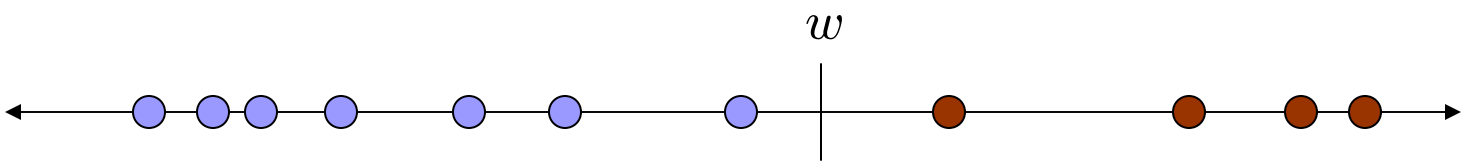


$$\text{H}=\{h_w : w \in \mathbb{R}\} \qquad h_w(x) = \begin{cases} 1 & \text{if } x \geq w \\ 0 & \text{if } x < w \end{cases}$$

$\text{O}(1/\epsilon)$ random labeled examples needed from P to get error rate $< \epsilon$

Example: Reals in 1-D

$$h_w(x) = \begin{cases} 1 & \text{if } x \geq w \\ 0 & \text{if } x < w \end{cases}$$



Passive learning:

$\mathrm{O}(1/\epsilon)$ random labeled examples needed from P to get error rate $< \epsilon$

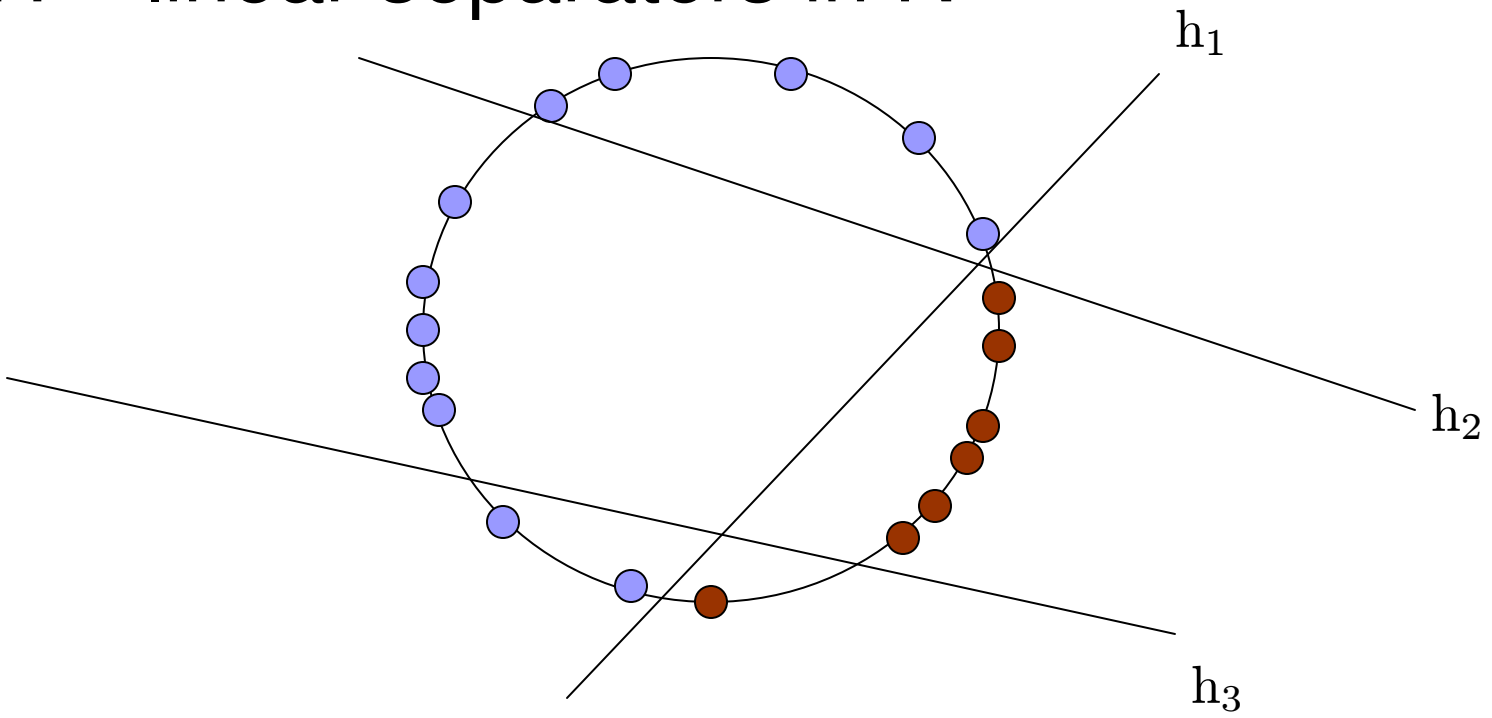
Active learning (Binary Search):

$\mathrm{O}(\log 1/\epsilon)$ examples needed to get error $< \epsilon$

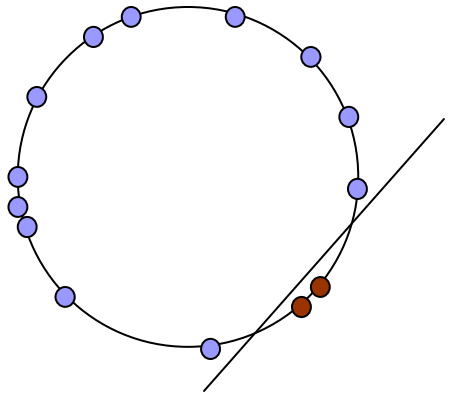
Active learning gives us an exponential improvement!

Example 2: Points on a Circle

- P = some density on circle perimeter
- H = linear separators in R^2



Example 2: Points on a Circle



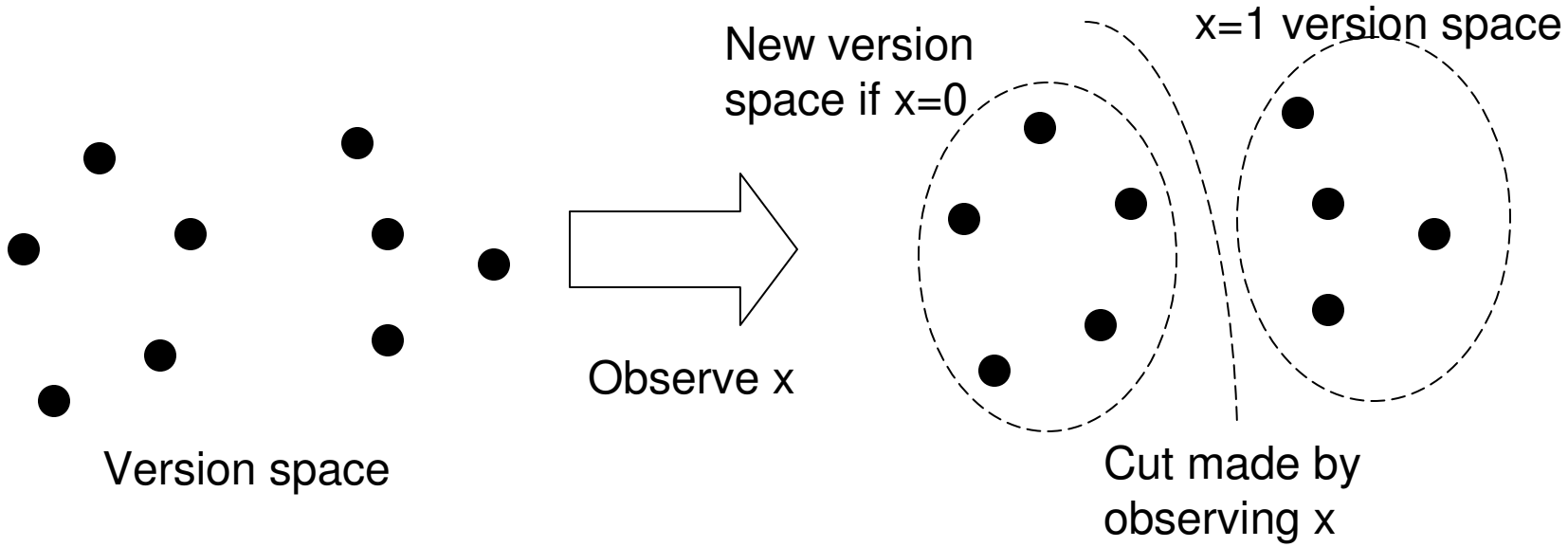
Worst case: small ϵ slice of the circle is different

- Passive learning: $\mathrm{O}(1/\epsilon)$
- Active learning: $\mathrm{O}(1/\epsilon)$

No improvement!

Active Learning Abstracted

- Goal: Narrow down the **version space**, (hypotheses that fit with known labels
- Idea: Think of hypotheses as points



Shrinking the Version Space

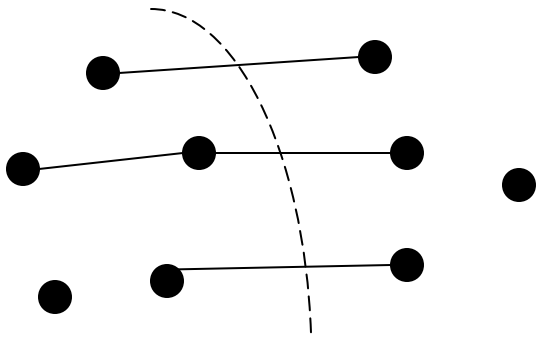
- Define distance between hypotheses:

$$\mathrm{d(h,h')=P\{x:h(x)\neq h'(x)\}}$$

- Ignore distances less than ϵ

$$\mathrm{Q=H\times H}$$

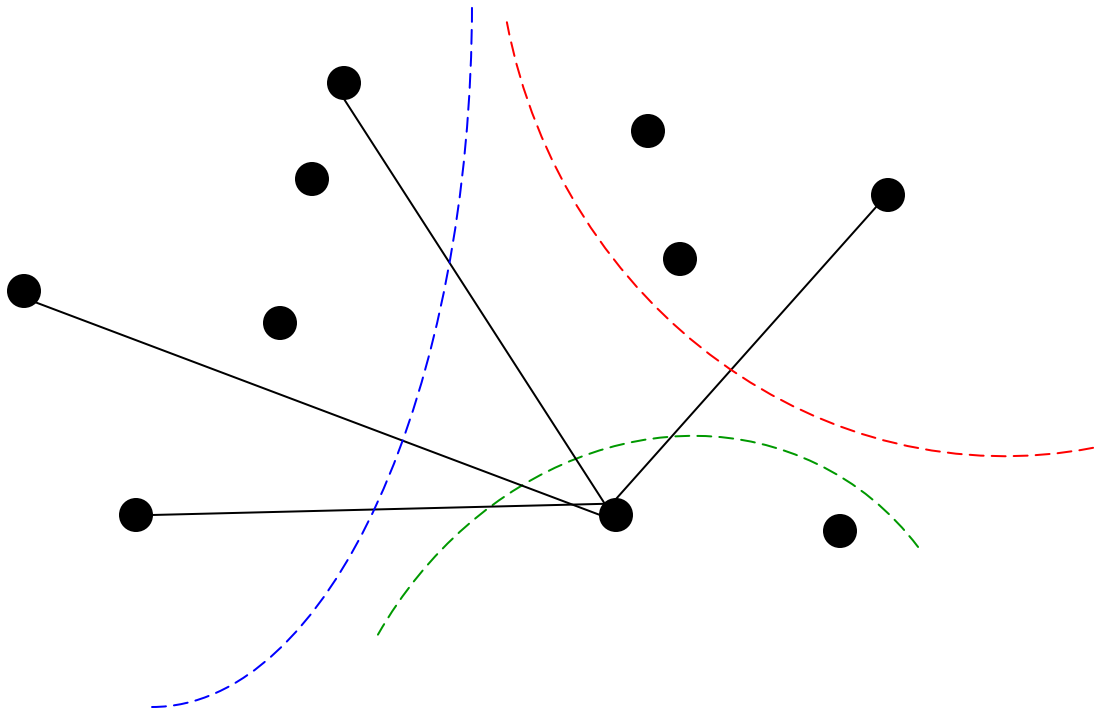
$$\mathrm{Q}_\epsilon = \{(h, h') \in Q : d(h, h') > \epsilon\}$$



A good cut!

Quick Example

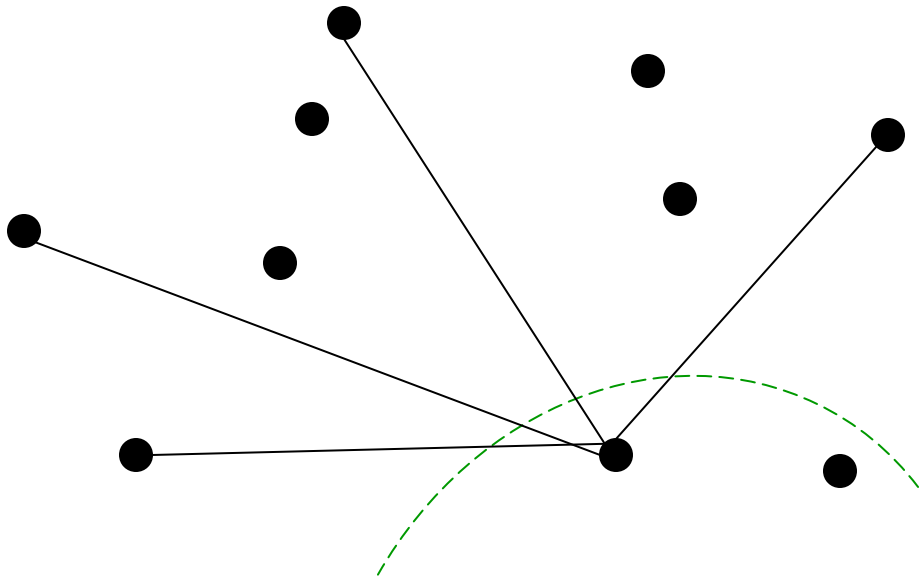
- What is the best cut?



$$Q_\epsilon = \{(h, h') \in Q : d(h, h') > \epsilon\}$$

Quick Example

- Cut edges => shrink version space

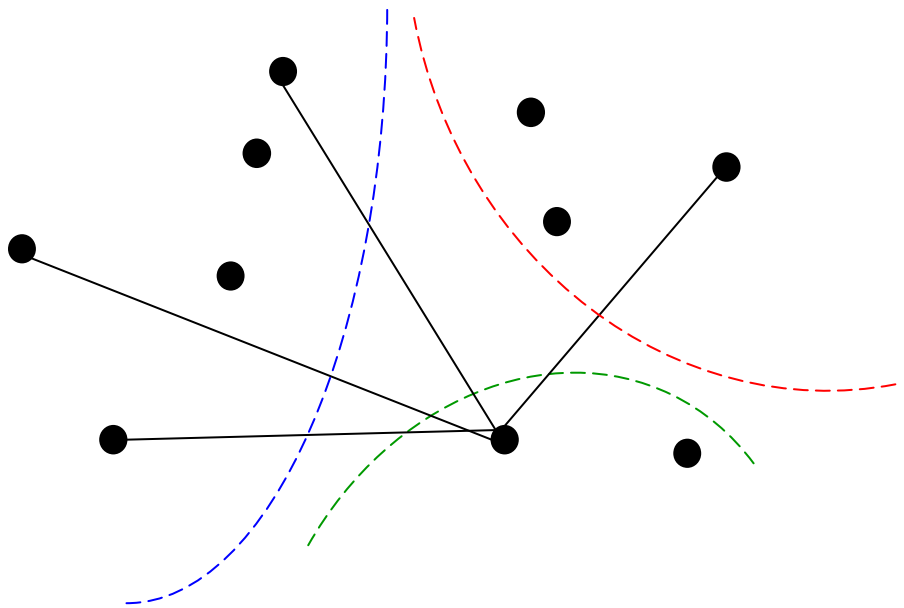


After this cut, we have a solution!

The hypotheses left are insignificantly different.

Quantifying “Usefulness” of Points

A point x$\in X$ is said to $\rho - split\ Q_\epsilon$
IF its label reduces the number of edges by a fraction $\rho > 0$



¼-split

1-split

¾-split

Quantifying the Difficulty of Problems

Definition:
Subset S of hypotheses is (ρ, ϵ, τ)-splittable if

$$\mathrm{P}\{x : x \ \rho\text{-splits } Q_\epsilon\} \geq \tau$$

"At least a fraction of τ samples are ρ-useful in splitting S."

ρ small $\Rightarrow$ smaller splits

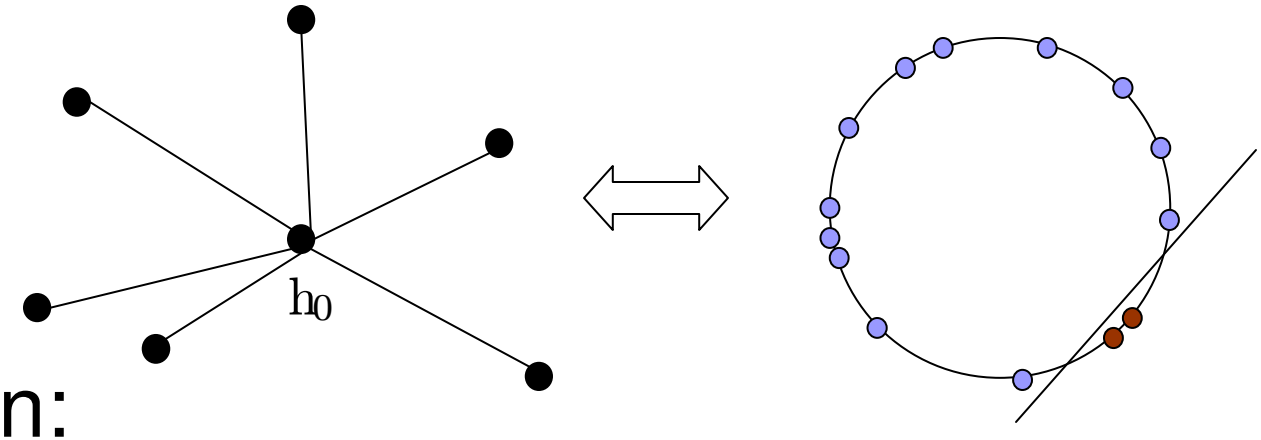
ϵ small $\Rightarrow$ small error

τ small $\Rightarrow$ lots of samples needed to get a good split

Lower Bound Result

Suppose for some hypothesis space H:

- $\mathrm{d}(h_0, h_i) > \epsilon$ for some hypotheses $\mathrm{h}_1, h_2, ..., h_N$
- "disagree sets" $\{x : h_0(x) \neq h_i(x)\}$ are disjoint



Then:

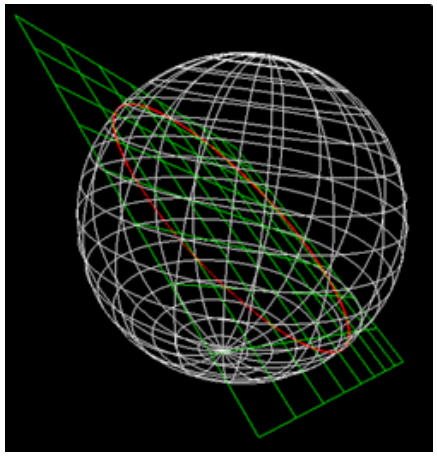
For any τ and $\rho > 1/N$, Q is not (ρ, ϵ, τ)-splittable.

An Interesting Result

There is constant $c > 0$ such that for any dimension $d \geq 2$, if

1. H is the class of homogeneous lenear separators in R^d, and
2. P is the uniform distribution over the surface of the unit sphere,

then H is $(1/4, \epsilon, c\epsilon)$-splittable for all $\epsilon > 0$.



$\Rightarrow$ For any $h \in H$, any $\epsilon \leq 1/(32\pi^2\sqrt{d})$,
$B(h, 4\epsilon)$ is $\left(\frac{1}{8}, \epsilon, \Omega\left(\epsilon/\sqrt{d}\right)\right)$-splittable.

Conclusions

- Active learning not always much better than passive.
- “Splittability” is the VC dimension for active learning.
- We can use this framework to fit bounds for specific problems.