Active Learning

9.520 Class 22, 03 May 2006

Claire Monteleoni

MIT CSAIL

Outline

Motivation

Historical framework: query learning

Current framework: selective sampling

Some recent results

Open problems

Active learning motivation

Machine learning applications, e.g.

- Medical diagnosis
- Document/webpage classification
- Speech recognition

Unlabeled data is abundant, but labels are expensive.

Active learning is a useful model here.

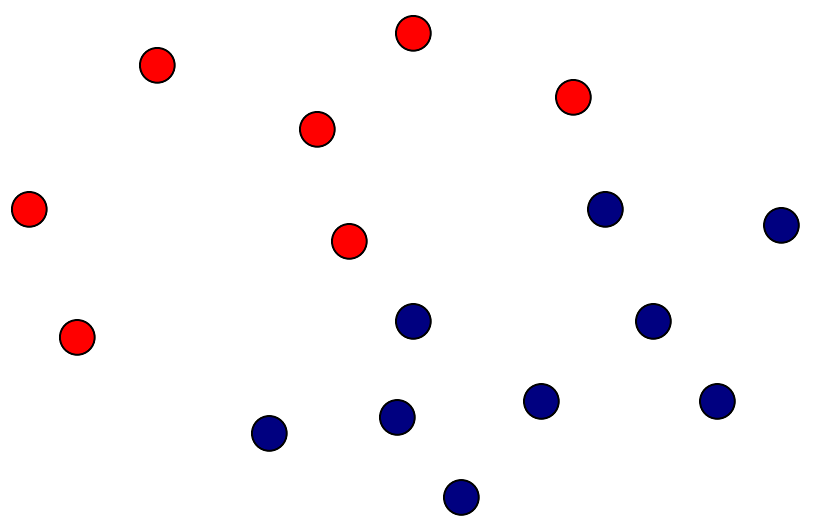
Allows for intelligent choices of which examples to label.

Label-complexity: the number of labeled examples required to learn via active learning

→ can be much lower than the PAC sample complexity!

Supervised learning

Given access to labeled data (drawn iid from an unknown underlying distribution P), want to learn a classifier chosen from hypothesis class H, with misclassification rate $<\varepsilon$.

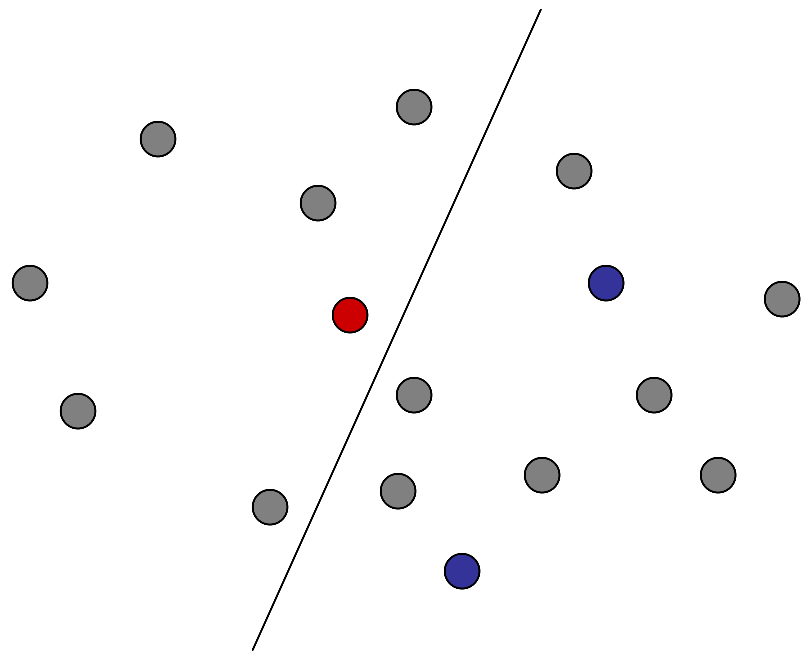


Sample complexity characterized by d = VC dimension of H.
If data is *separable*, need roughly d/ε labeled samples.

Slide credit: Sanjoy Dasgupta

Active learning

In many situations unlabeled data is easy to come by, but there is a charge for each label.



What is the minimum number of labels needed to achieve the target error rate?

Slide credit: S. Dasgupta

Active learning variants

There are several models of active learning:

- Query learning (a.k.a. Membership queries)
- Selective sampling
- Active model selection
- Experiment design

Various evaluation frameworks:

- Regret minimization
- Minimize label-complexity to reach fixed error rate
- Label-efficiency (fixed label budget)

We focus on classification, though regression AL exists too.

Membership queries

Earliest model of active learning in theory work [Angluin 1992]

X = space of possible inputs, like $\{0,1\}^n$
H = class of hypotheses

Target concept $h^* \in H$ to be identified *exactly*.
You can ask for the label of any point in X: *no unlabeled data*.

$H_0 = H$
For $t = 1,2,\ldots$
    pick a point $x \in X$ and query its label $h^*(x)$
    let H_t = all hypotheses in H_{t-1} consistent with $(x, h^*(x))$

What is the minimum number of "membership queries" needed to reduce H to just $\{h^*\}$?

Slide credit: S. Dasgupta

Membership queries: example

$X = \{0,1\}^n$

H = AND-of-positive-literals, like $x_1 \wedge x_3 \wedge x_{10}$

$S = \{\ \}$ (set of AND positions)

For $i = 1$ to n:

ask for the label of $(1,\ldots,1,0,1,\ldots,1)$ [0 at position i]

if negative: $S = S \cup \{i\}$

Total: n queries

General idea: synthesize highly informative points.

Each query cuts the *version space* -- the set of consistent hypotheses -- in half.

Slide credit: S. Dasgupta

Problem

Many results in this framework, even for complicated hypothesis classes.

[Baum and Lang, 1991] tried fitting a neural net to handwritten characters.
Synthetic instances created were incomprehensible to humans!

[Lewis and Gale, 1992] tried training text classifiers.
"an artificial text created by a learning algorithm is unlikely to be a legitimate natural language expression, and probably would be uninterpretable by a human teacher."

Slide credit: S. Dasgupta

Selective sampling [Cohn, Atlas & Ladner, 1992]

Selective sampling:

Given: pool (or stream) of unlabeled examples, *x*, drawn i.i.d. from input distribution.

Learner may request labels on examples in the pool/stream.

(Noiseless) oracle access to correct labels, *y*.

Constant cost per label

The error of any classifier h is measured on distribution *P*:

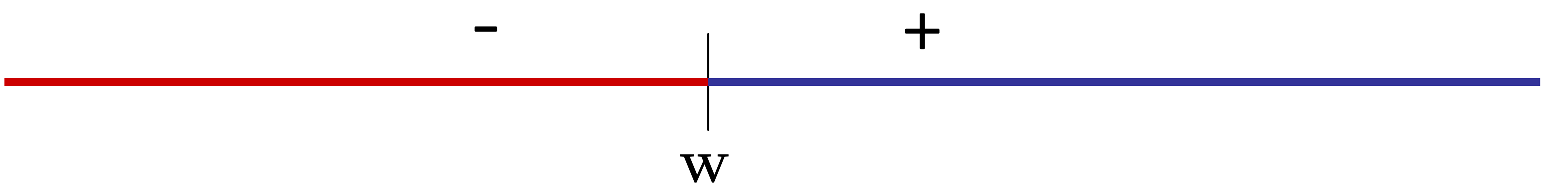
$err(h) = P(h(x) \neq y)$

Goal: minimize label-complexity to learn the concept to a fixed accuracy.

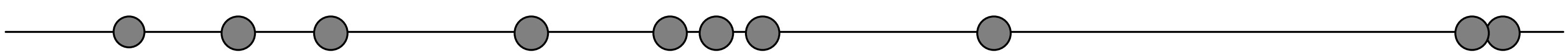
Can adaptive querying really help?

[CAL92, D04]: Threshold functions on the real line

$$h_w(x) = \mathbf{1}(x \geq w), \qquad H = \{h_w : w \in \mathbf{R}\}$$



Start with $1/\varepsilon$ *unlabeled* points



Binary search – need just $\log 1/\varepsilon$ labels, from which the rest can be inferred! Exponential improvement in sample complexity.

Slide credit: S. Dasgupta

More general hypothesis classes

For a general hypothesis class with VC dimension d, is a "generalized binary search" possible?

Random choice of queries $\quad d/\varepsilon$ labels
Perfect binary search $\quad d \log 1/\varepsilon$ labels

Where in this large range does the label complexity of active learning lie?

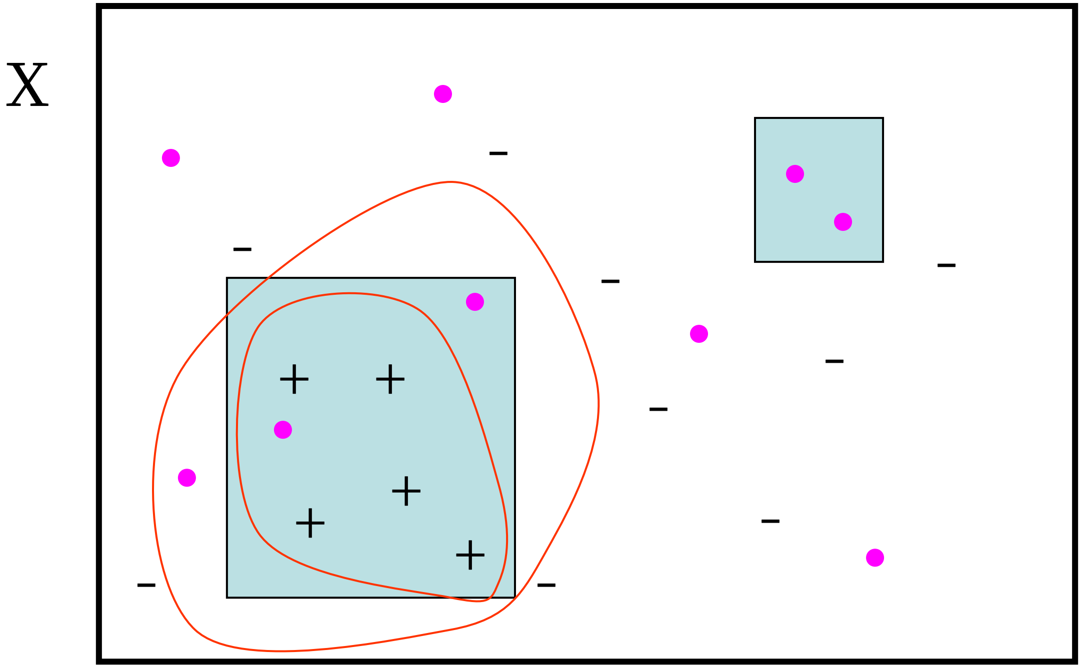
We've already handled linear separators in 1-d…

Slide credit: S. Dasgupta

[1] Uncertainty sampling

Maintain a single hypothesis, based on labels seen so far.
Query the point about which this hypothesis is most "uncertain".

Problem: confidence of a single hypothesis may not accurately represent the true diversity of opinion in the hypothesis class.



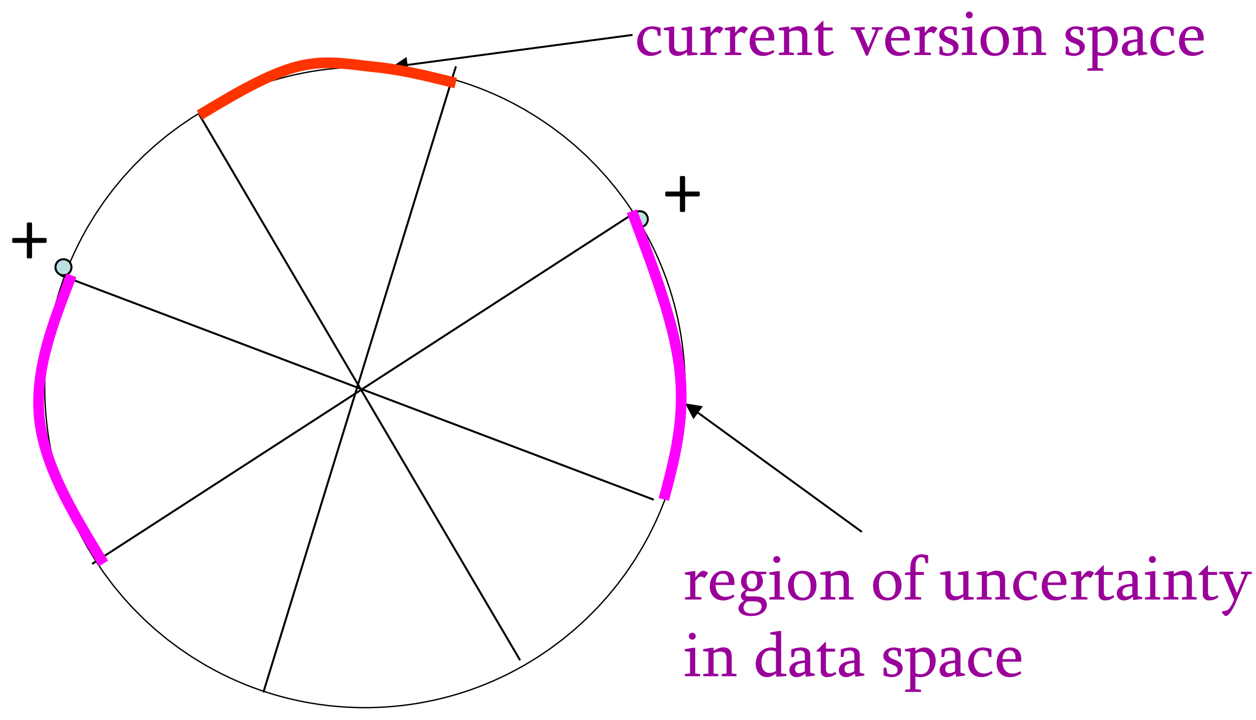
Slide credit: S. Dasgupta

[2] Region of uncertainty

Current version space: portion of H consistent with labels so far.
"Region of uncertainty" = part of data space about which there is still some uncertainty (ie. disagreement within version space)

Suppose data lies on circle in R^2; hypotheses are linear separators.

(spaces X, H superimposed)



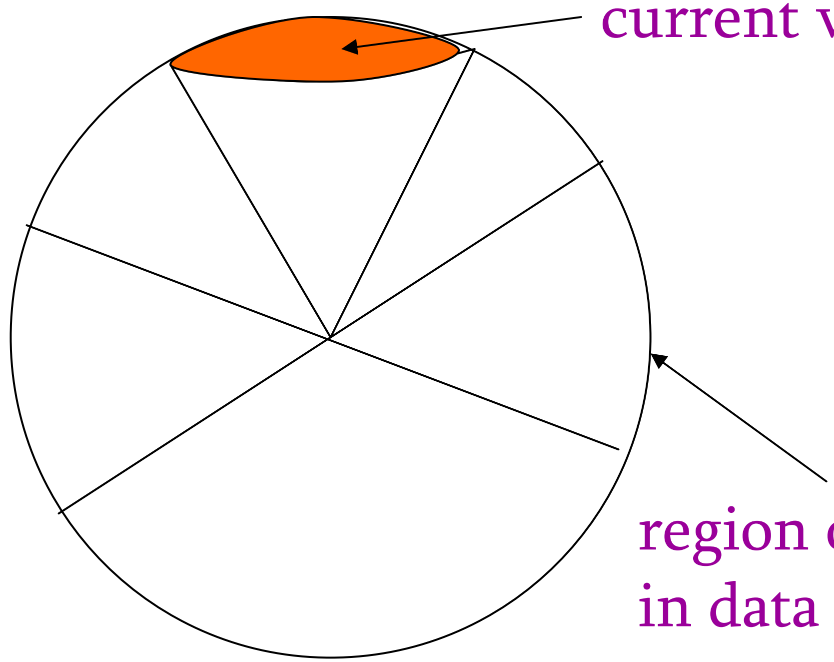
Slide credit: S. Dasgupta

[2] Region of uncertainty

Algorithm [CAL92]:
of the unlabeled points which lie in the region of uncertainty,
pick one at random to query.

Data and hypothesis spaces, superimposed:

(both are the surface of the unit sphere in R^d)



Slide credit: S. Dasgupta

[2] Region of uncertainty

Number of labels needed depends on H and also on P.

Special case: H = {linear separators in R^d}, P = uniform distribution over unit sphere.

Theorem [Balcan, Beygelzimer & Langford ICML '06]:
$\tilde{O}(d^2 \log 1/\varepsilon)$ labels are needed to reach a hypothesis with error rate $< \varepsilon$.

Supervised learning: $\Theta(d/\varepsilon)$ labels.

Slide credit: S. Dasgupta

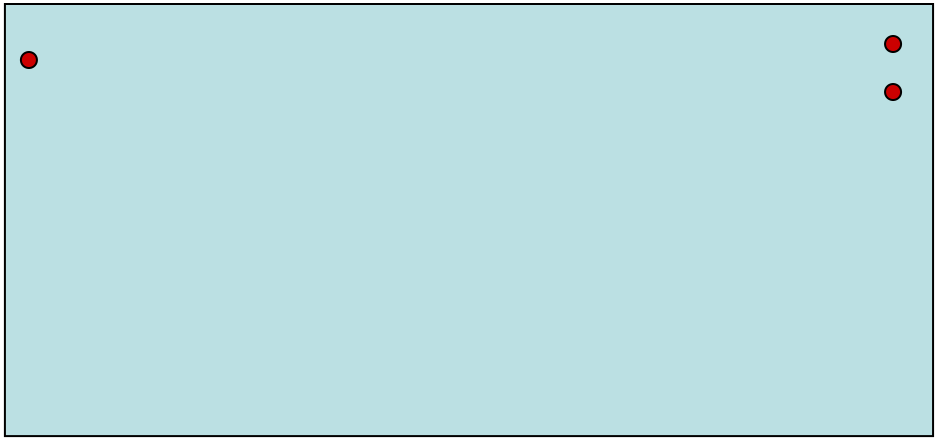
[3] Query-by-committee

[Seung, Opper, Sompolinsky, 1992; Freund, Seung, Shamir, Tishby 1997]

First idea: Try to rapidly reduce volume of version space?

Problem: doesn't take data distribution into account.

H:



Which pair of hypotheses is closest? Depends on data distribution P.

Distance measure on H: $d(h,h') = P(h(x) \neq h'(x))$

Slide credit: S. Dasgupta

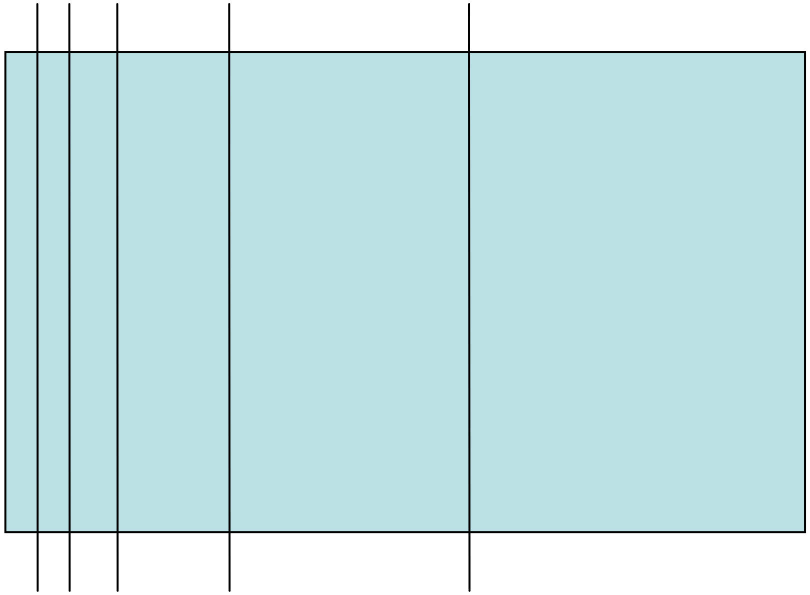
[3] Query-by-committee

First idea: Try to rapidly reduce volume of version space?

Problem: doesn't take data distribution into account.

To keep things simple, say d(h,h') $\propto$ Euclidean distance in this picture.

H:



Error is likely to remain large!

Slide credit: S. Dasgupta

[3] Query-by-committee

Elegant scheme which decreases volume in a manner which is sensitive to the data distribution.

Bayesian setting: given a prior π on H

$H_1 = H$
For t = 1, 2,
    receive an unlabeled point x_t drawn from P
    [informally: is there a lot of disagreement about x_t in H_t?]
    choose two hypotheses h,h' randomly from (π, H_t)
    if $h(x_t) \neq h'(x_t)$: ask for x_t's label
    set H_{t+1}

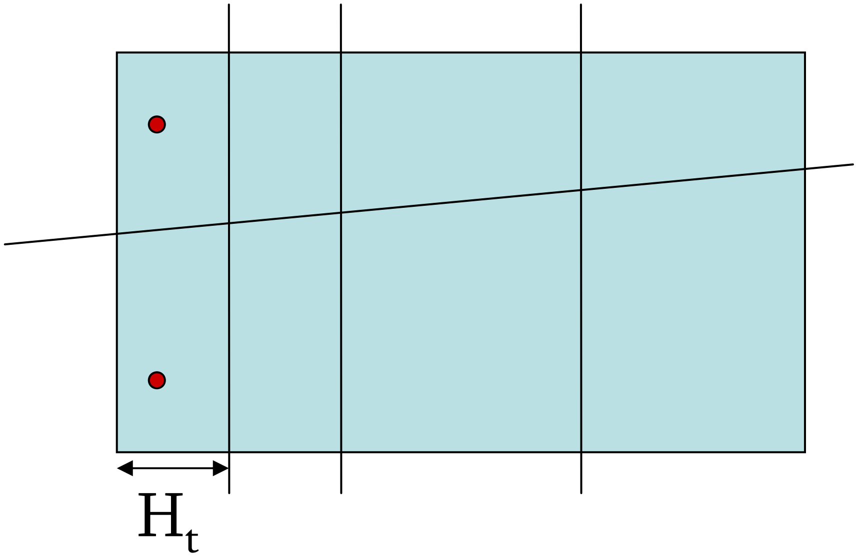
Slide credit: S. Dasgupta

[3] Query-by-committee

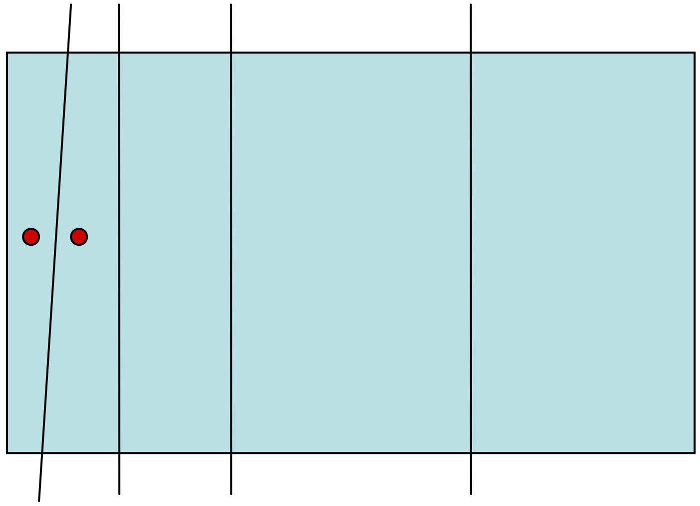
For $t = 1, 2, \ldots$

receive an unlabeled point x_t drawn from P
choose two hypotheses h,h' randomly from (π, H_t)
if $h(x_t) \neq h'(x_t)$: ask for x_t's label
set H_{t+1}

Observation: the probability of getting pair (h,h') in the inner loop (when a query is made) is proportional to $\pi(h)\ \pi(h')\ d(h,h')$.



vs.



Slide credit: S. Dasgupta

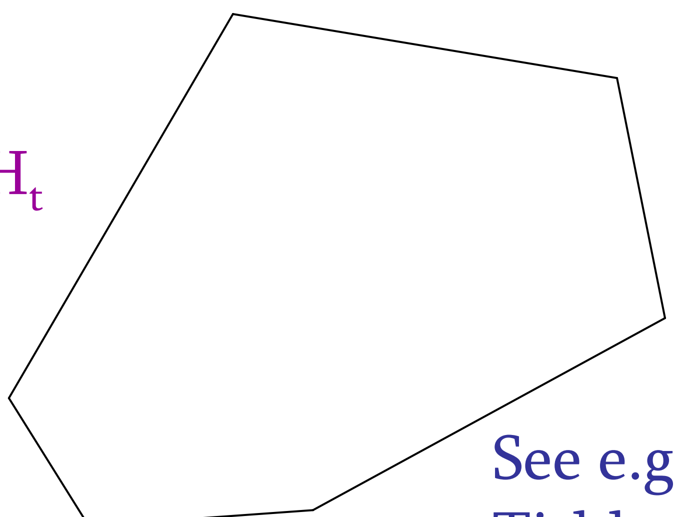
[3] Query-by-committee

Label bound, Theorem [FSST97] :
For H = {linear separators in R^d}, P = uniform distribution, then $\tilde{O}(d \log 1/\varepsilon)$ labels to reach a hypothesis with error $< \varepsilon$.

Implementation: need to randomly pick h according to (π, H_t).

e.g. H = {linear separators in R^d}, π = uniform distribution:

H_t



How do you pick a random point from a convex body?

See e.g. [Gilad-Bachrach, Navot & Tishby NIPS '05]

Slide credit: S. Dasgupta

Online active learning

Under Bayesian assumptions, QBC can learn a half-space through the origin to generalization error ε, using $\tilde{O}(d \log 1/\varepsilon)$ labels.

→ But not online: space required, and time complexity of the update both scale with number of seen mistakes!

Online algorithms:

- See unlabeled data streaming by, one point at a time
- Can query current point's label, at a cost
- Can only maintain current hypothesis (memory bound)

Online learning: related work

Standard (supervised) Perceptron: a simple online algorithm:

If $y_t \neq \mathrm{SGN}(v_t \cdot x_t)$, then: Filtering rule

$v_{t+1} = v_t + y_t\, x_t$ Update step

Distribution-free mistake bound $O(1/\gamma^2)$, if exists margin γ.

Theorem [Baum'89]: Perceptron, given sequential labeled examples from the uniform distribution, can converge to generalization error ε after $\tilde{O}(d/\varepsilon^2)$ mistakes.

Fast online active learning [Dasgupta, Kalai & M, COLT '05]

A lower bound for Perceptron in active learning context of $\Omega(1/\varepsilon^2)$ labels.

A modified Perceptron update with a $\tilde{O}(d \log 1/\varepsilon)$ mistake bound.

An active learning rule and a label bound of $\tilde{O}(d \log 1/\varepsilon)$.

A bound of $\tilde{O}(d \log 1/\varepsilon)$ on total errors (labeled or not).

Selective sampling, online constraints

Sequential selective sampling framework:

Unlabeled examples, x_t, are received one at a time,
sampled i.i.d. from the input distribution.

Learner makes a prediction at each time-step.

A noiseless oracle to label y_t, can be queried at a cost.

Goal: minimize number of *labels* to reach error ε.

ε is the error rate (w.r.t. the target) on the input distribution.

Online constraints:

Space: Learner cannot store all previously seen examples (and then perform batch learning).

Time: Running time of learner's belief update step should not scale with number of seen examples/mistakes.

AC Milan vs. Inter Milan







Problem framework

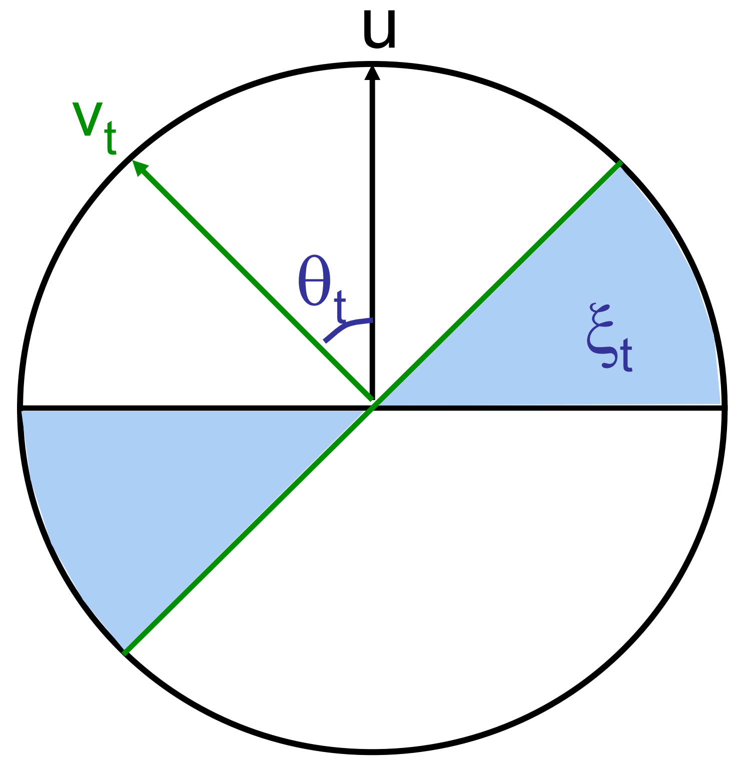
$$S = \{x \in \mathbb{R}^d \mid \|x\| = 1\}, \quad x_t \in S, \quad y_t \in \{-1, +1\}$$

Target: $u : y_t(u \cdot x_t) > 0 \ \ \forall t, \quad \|u\| = 1$

Current hypothesis: v_t

$$\theta_t = \arccos(u \cdot \hat{v}_t) \ : \ \hat{v}_t = \frac{v_t}{\|v_t\|}$$

Error region: ξ_t



Assumptions:

Separability

u is through origin

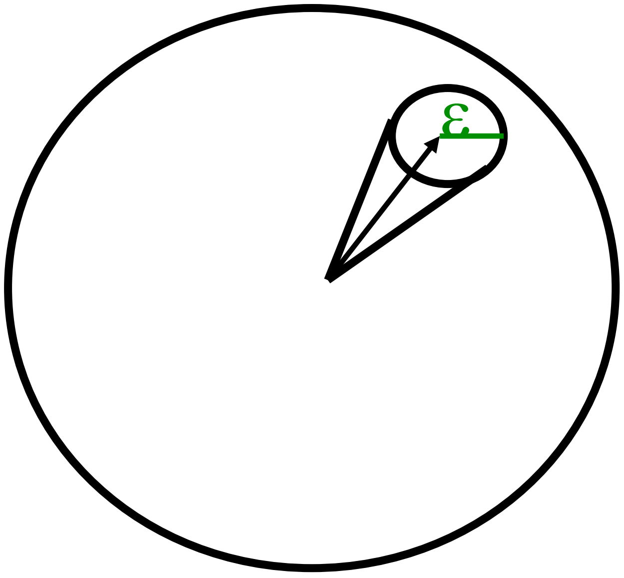
x~Uniform on S

error rate: $\epsilon_t = P_{x \in S}[x \in \xi_t] = \frac{\theta_t}{\pi}$

OPT

Fact: Under this framework, any algorithm requires $\Omega(d \log 1/\varepsilon)$ labels to output a hypothesis within generalization error at most ε.

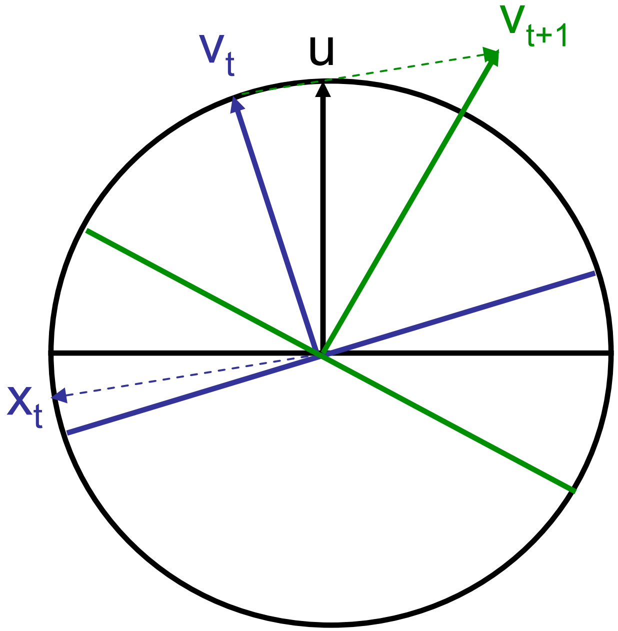
Proof idea: Can pack $(1/\varepsilon)^d$ spherical caps of radius ε on surface of unit ball in $\mathbb{R}^d$. The bound is just the number of bits to write the answer.



Perceptron

Perceptron update: $v_{t+1} = v_t + y_t\, x_t$

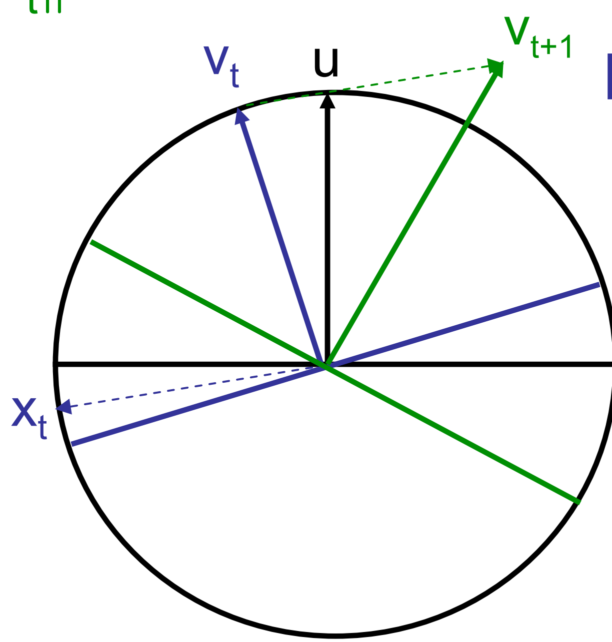
$\rightarrow$ error does not decrease monotonically.



Lower bound on labels for Perceptron

Theorem [DKM05]: The Perceptron algorithm, using any active learning rule, requires $\Omega(1/\varepsilon^2)$ labels to reach generalization error ε w.r.t. the uniform distribution.

Proof idea: Lemma: For small θ_t, the Perceptron update will increase θ_t unless $\|v_t\|$ is large: $\Omega(1/\sin\theta_t)$.



But, $\|v_t\|$ growth rate: $O(\sqrt{t})$

So need $t \geq 1/\sin^2\theta_t$.

Under uniform,

$\varepsilon_t \propto \theta_t \geq \sin\theta_t$.

A modified Perceptron update

Standard Perceptron update:

$$v_{t+1} = v_t + y_t\, x_t$$

Instead, weight the update by "confidence" w.r.t. current hypothesis v_t:

$$v_{t+1} = v_t + 2\, y_t\, |v_t \cdot x_t|\, x_t \qquad (v_1 = y_0 x_0)$$

(similar to update in [Blum et al.'96] for noise-tolerant learning)

Unlike Perceptron:

Error decreases monotonically:

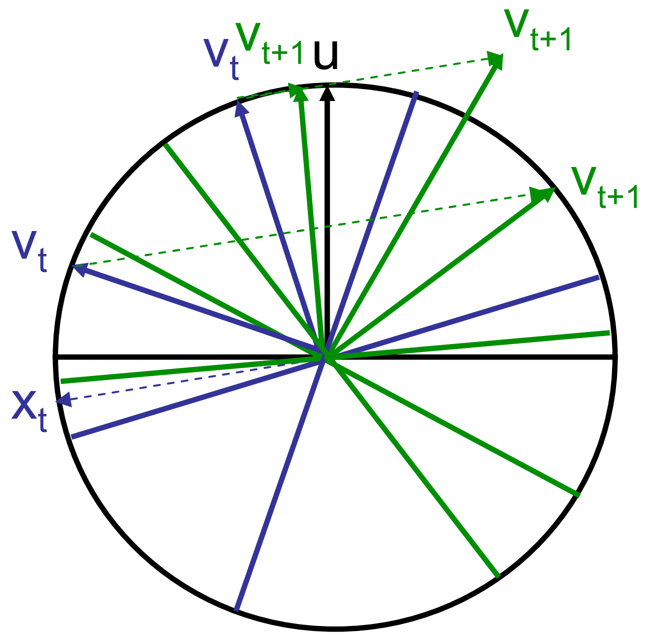
$$\cos(\theta_{t+1}) = u \cdot v_{t+1} = u \cdot v_t + 2\,|v_t \cdot x_t||u \cdot x_t|$$
$$\geq u \cdot v_t = \cos(\theta_t)$$

$\|v_t\| = 1$ (due to factor of 2)

A modified Perceptron update

Perceptron update: $v_{t+1} = v_t + y_t\, x_t$

Modified Perceptron update: $v_{t+1} = v_t + 2\, y_t\, |v_t \cdot\ x_t|\, x_t$



Mistake bound

Theorem [DKM05]: In the supervised setting, the modified Perceptron converges to generalization error ε after $\tilde{O}(d \log 1/\varepsilon)$ mistakes.

Proof idea: The exponential convergence follows from a multiplicative decrease in θ_t:

$$1 - \cos\theta_{t+1} \leq (1 - \frac{c}{d})(1 - \cos\theta_t)$$

On an update,

$$\begin{aligned} \cos\theta_{t+1} &= u \cdot v_{t+1} = u \cdot v_t + 2y_t|v_t \cdot x_t|(u \cdot x_t) \\ &= u \cdot v_t + 2|v_t \cdot x_t||u \cdot x_t| \\ &= \cos\theta_t + 2|v_t \cdot x_t||u \cdot x_t| \end{aligned}$$

$\rightarrow$ Lower bound $2|v_t \cdot x_t||u \cdot x_t|$, with high probability, using distributional assumption.

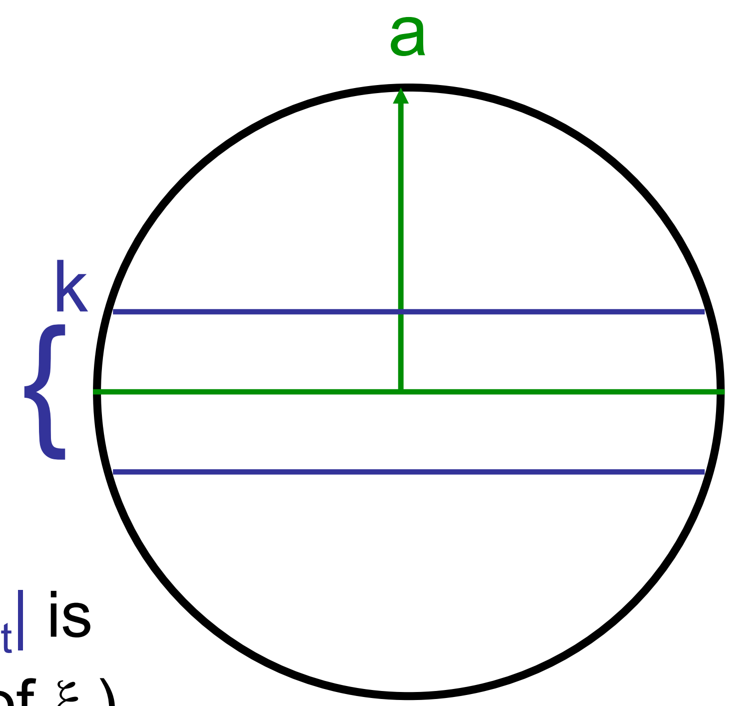
Mistake bound

Theorem 2: In the supervised setting, the modified Perceptron converges to generalization error ε after $\tilde{O}(d \log 1/\varepsilon)$ mistakes.

Lemma (band): For any fixed a: $\|a\|=1$, $\gamma \leq 1$ and for x~U on S:

$$\frac{\gamma}{4} \leq P_{x \in S}\left[|a \cdot x| \leq \frac{\gamma}{\sqrt{d}}\right] \leq \gamma$$

$\{x : |a \cdot x| \leq k\} =$



Apply to $|v_t \cdot x|$ and $|u \cdot x| \Rightarrow 2|v_t \cdot x_t||u \cdot x_t|$ is large enough in expectation (using size of ξ_t).

Active learning rule

Goal: Filter to label just those points in the error region.

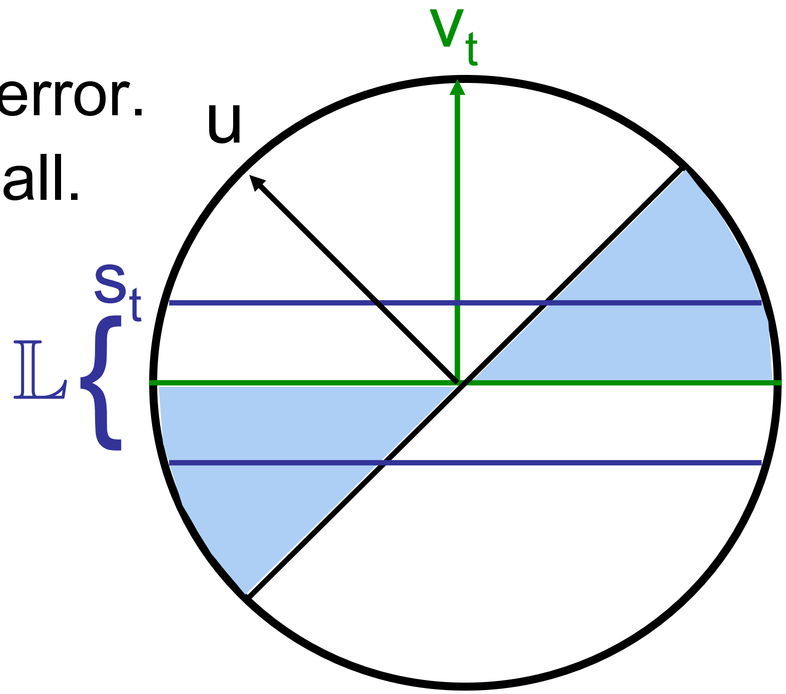
$\rightarrow$ but θ_t, and thus ξ_t unknown!

Define labeling region: $\mathbb{L} = \left\{ x \;\middle|\; |v_t \cdot x| \leq s_t \right\}$

Tradeoff in choosing threshold s_t:

If too high, may wait too long for an error.

If too low, resulting update is too small.



$\mathbb{L} = \left\{ x \;\middle|\; |v_t \cdot x| \leq \frac{\sin \theta_t}{\sqrt{d}} \right\}$ makes

$P_{x \in S} \left[x \in \mathbb{L} \;\middle|\; x \in \xi_t \right]$ *constant.*

$\rightarrow$ But θ_t unknown!

Active learning rule

Choose threshold s_t adaptively:

Start high.

Halve, if no error in R consecutive labels.

$$\mathbb{L} = \left\{ x \;\middle|\; |v_t \cdot x| \le s_t \right\}$$

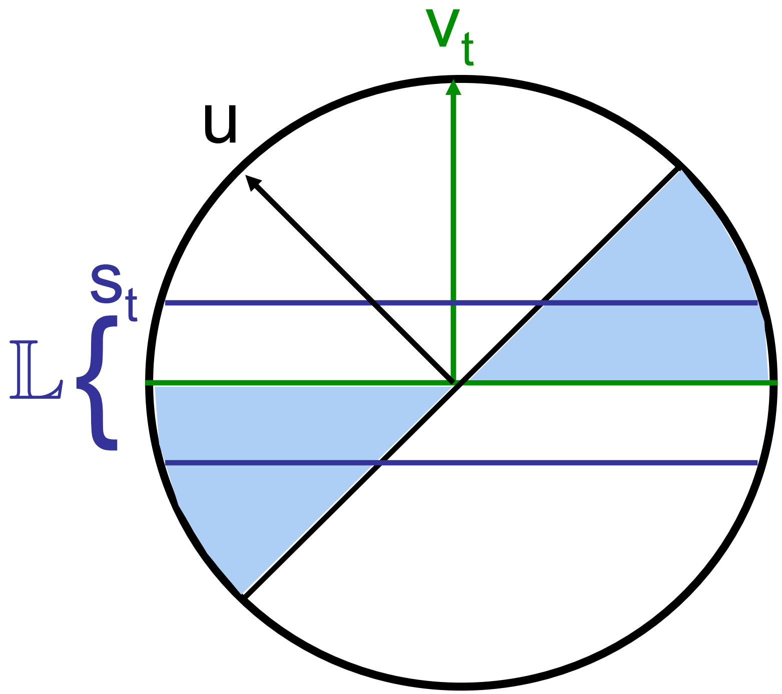
Start with threshold s_t high:

$$s_1 = \frac{\sin \frac{\pi}{2}}{\sqrt{d}} = \frac{1}{\sqrt{d}}$$

After R consecutive labeled points,

if no errors:

$$s_{t+1} = \frac{s_t}{2}$$



Label bound

Theorem [DKM05]: In the active learning setting, the modified Perceptron, using the adaptive filtering rule, will converge to generalization error ε after $\tilde{O}(d \log 1/\varepsilon)$ labels.

Corollary [DKM05] : The total errors (labeled and unlabeled) will be $\tilde{O}(d \log 1/\varepsilon)$.

Proof technique

Proof outline: We show the following lemmas hold with sufficient probability:

Lemma 1. s_t does not decrease too quickly: $s_t \geq \frac{\sin \theta_t}{4\sqrt{d}}$

Lemma 2. We query labels on a constant fraction of ξ_t.

Lemma 3. With constant probability the update is *good*.

By algorithm, ~1/R labels are mistakes. $\exists$ R = Õ(1).

$\Rightarrow$ Can thus bound labels and total errors by mistakes.

[DKM05] in context

	samples	mistakes	labels	total errors	online?
PAC complexity [Long'03] [Long'95]	$\tilde{O}(d/\varepsilon)$ $\Omega(d/\varepsilon)$				
Perceptron [Baum'97]	$\tilde{O}(d/\varepsilon^3)$ $\Omega(1/\varepsilon^2)$	$\tilde{O}(d/\varepsilon^2)$ $\Omega(1/\varepsilon^2)$	$\Omega(1/\varepsilon^2)$		✓
CAL [BBL'06]	$\tilde{O}((d^2/\varepsilon) \log 1/\varepsilon)$	$\tilde{O}(d^2 \log 1/\varepsilon)$	$\tilde{O}(d^2 \log 1/\varepsilon)$		×
QBC [FSST'97]	$\tilde{O}(d/\varepsilon \log 1/\varepsilon)$	$\tilde{O}(d \log 1/\varepsilon)$	$\tilde{O}(d \log 1/\varepsilon)$		×
[DKM'05]	$\tilde{O}(d/\varepsilon \log 1/\varepsilon)$	$\tilde{O}(d \log 1/\varepsilon)$	$\tilde{O}(d \log 1/\varepsilon)$	$\tilde{O}(d \log 1/\varepsilon)$	✓

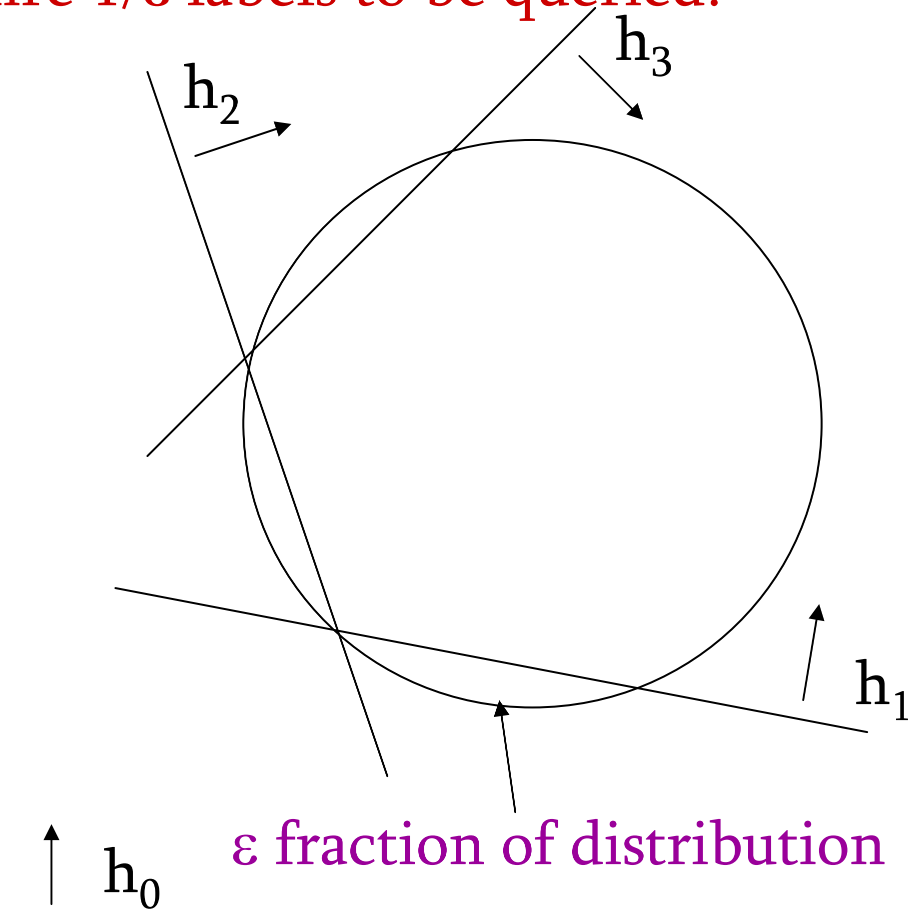
Lower bounds on label complexity

For linear separators in R^1, need just $\log 1/\varepsilon$ labels.
Theorem [D04]: when H = {non-homogeneous linear separators in R^2}: some target hypotheses require $1/\varepsilon$ labels to be queried!

Consider *any* distribution over the circle in R^2.

Need $1/\varepsilon$ labels to distinguish between $h_0, h_1, h_2, \ldots, h_{1/\varepsilon}$!

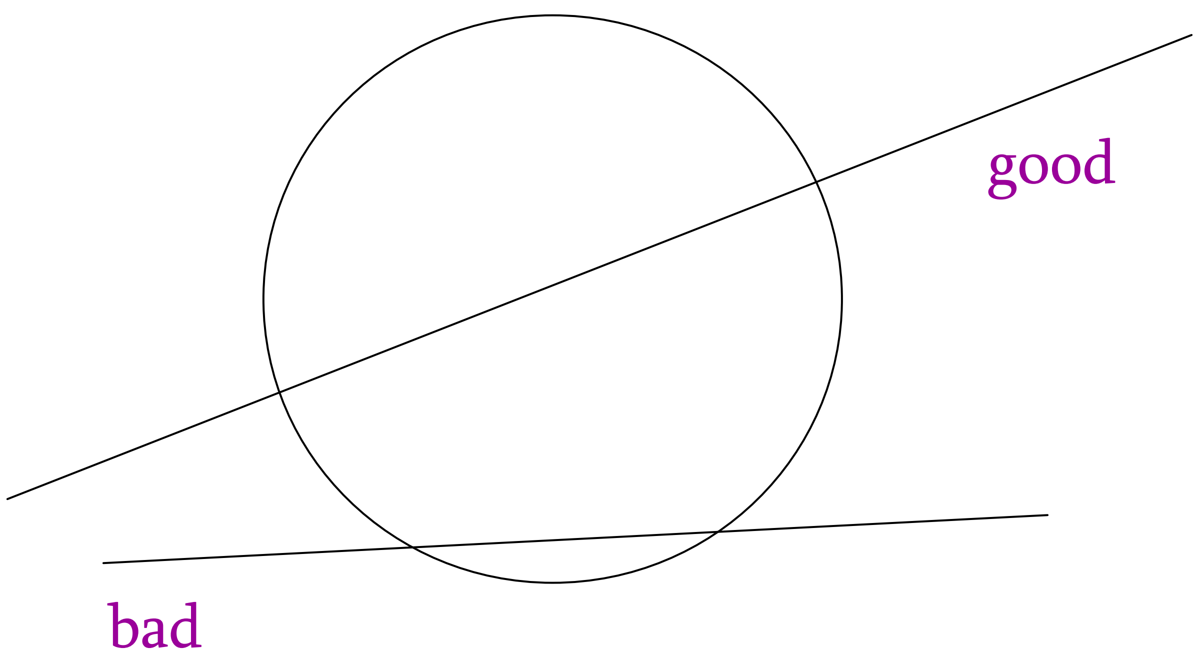
$\rightarrow$ Leads to analagous bound: $\Omega(1/\varepsilon)$ for homogeneous linear separators in R^d.



Slide credit: S. Dasgupta

A fuller picture

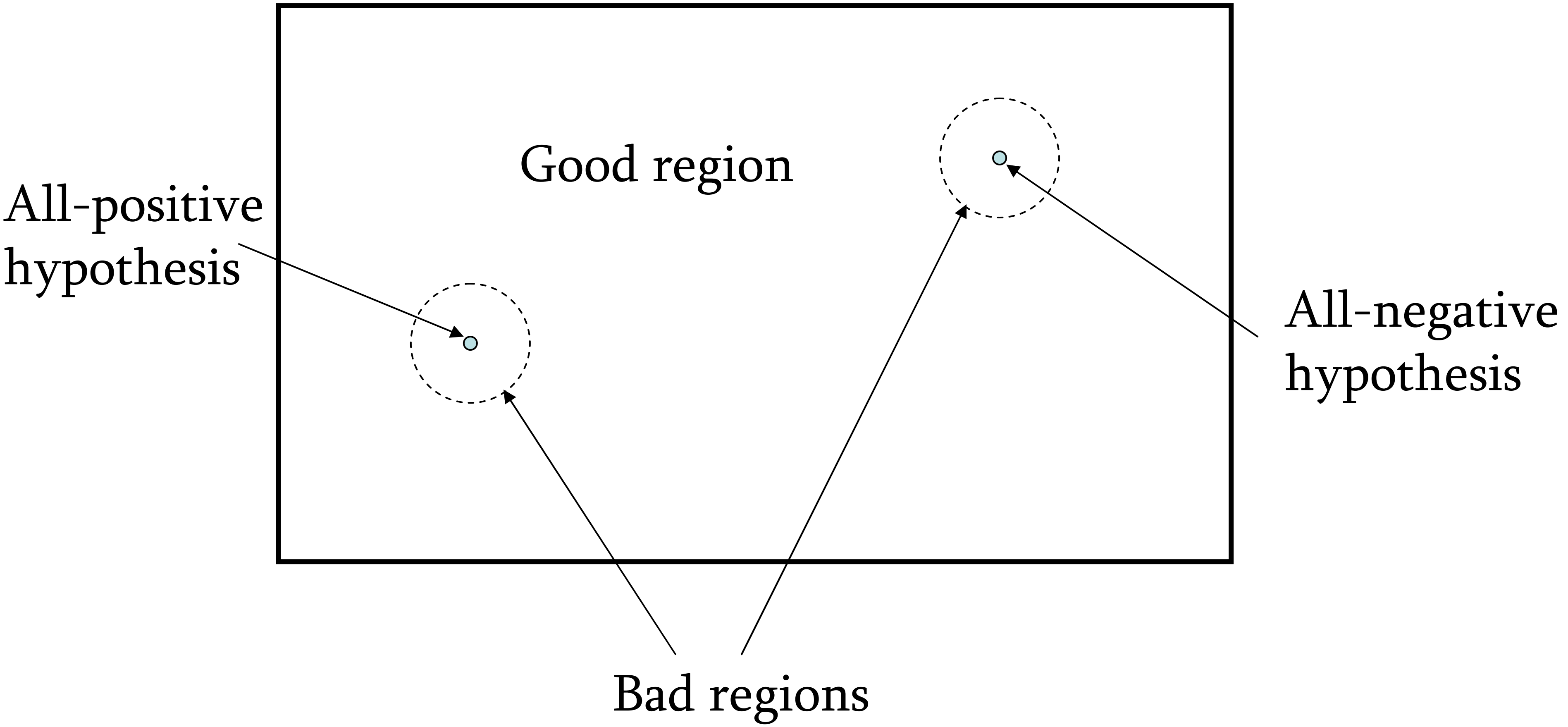
For non-homogenous linear separators in R^2: some bad target hypotheses which require $1/\varepsilon$ labels,
but "most" require just $O(\log 1/\varepsilon)$ labels…



Slide credit: S. Dasgupta

A view of the hypothesis space

H = {non-homogeneous linear separators in R^2}

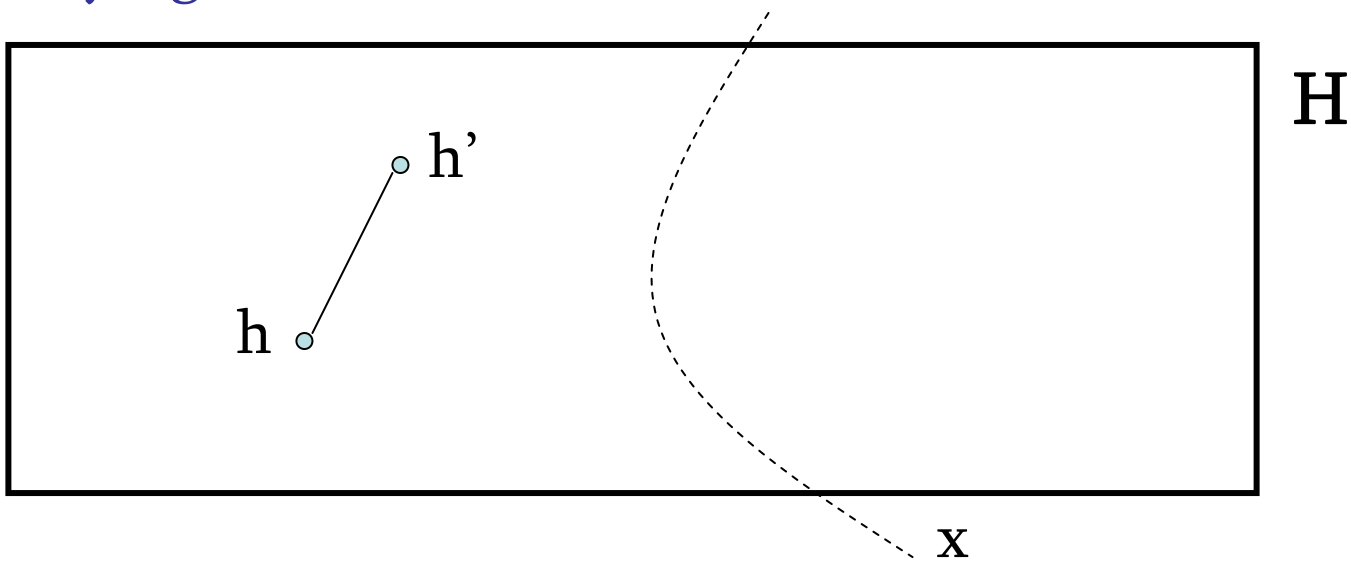


Slide credit: S. Dasgupta

Geometry of hypothesis space

H = any hypothesis class, of VC dimension $d < \infty$.

P = underlying distribution of data.



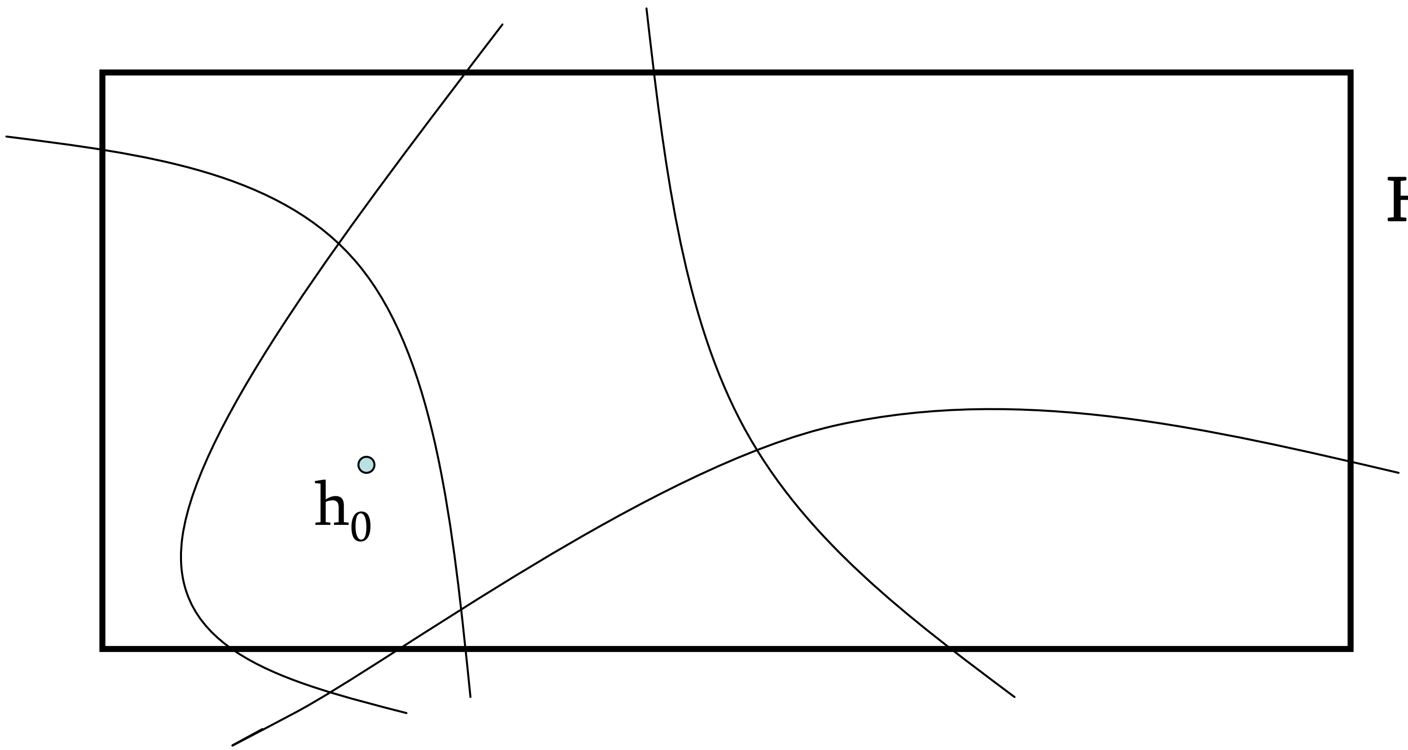
(i) Non-Bayesian setting: no probability measure on H

(ii) But there is a natural (pseudo) metric: $d(h,h') = \mathbf{P}(h(x) \neq h'(x))$

(iii) Each point x defines a cut through H

Slide credit: S. Dasgupta

Label upper bounding technique [Dasgupta NIPS‘05]



(h_0 = target hypothesis)

Proof technique: analyze how many labels until the diameter of the remaining version space is at most ε.

Slide credit: S. Dasgupta

Searchability index [D05]

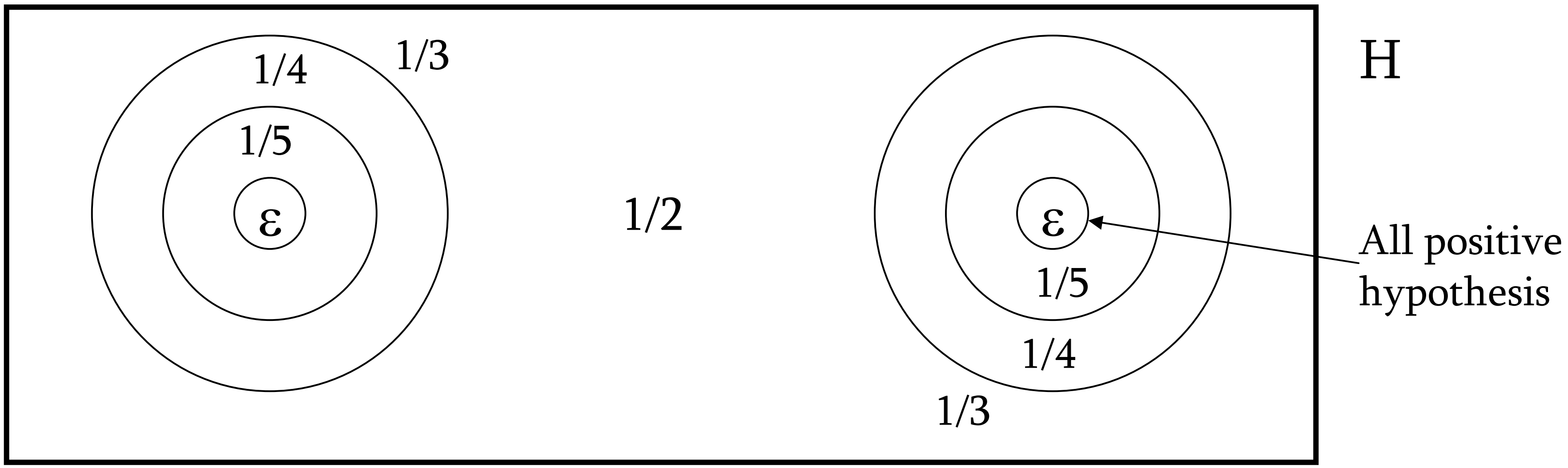
Accuracy ε
Data distribution P
Amount of unlabeled data

⇒

Each hypothesis $h \in H$ has a "searchability index" $\rho(h)$

$\varepsilon \leq \rho(h) \leq 1$, bigger is better

Example: linear separators in R^2, data on a circle:



$\rho(h) \propto \min(\text{pos mass of } h, \text{neg mass of } h)$, but never $< \varepsilon$

Slide credit: S. Dasgupta

Searchability index [D05]

Accuracy ε
Data distribution P
Amount of unlabeled data

⇒

Each hypothesis h ∈ H has a "searchability index" ρ(h)

Searchability index lies in the range: ε ≤ ρ(h) ≤ 1

Upper bound. For any H of VC-dim d<∞, there is an active learning scheme* which identifies (within accuracy ≤ ε) any h ∈ H, with a label complexity of at most: $\frac{1}{\rho(h)} \cdot \tilde{O}\left(d \log \frac{1}{\epsilon}\right)$

Lower bound. For any h ∈ H, any active learning scheme for the neighborhood B(h, ρ(h)) has a label complexity of at least: $\frac{1}{\rho(h)}$

[When ρ(h) ≫ ε: active learning helps a lot.]

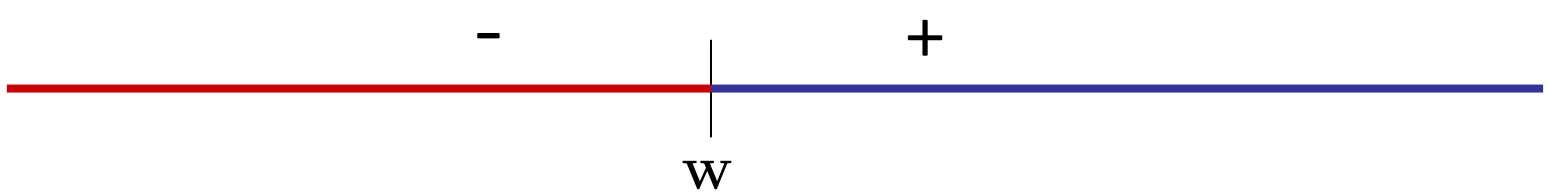
Slide credit: S. Dasgupta

Example: the 1-d line

Searchability index lies in range: $\varepsilon \leq \rho(h) \leq 1$

Theorem [D05]: $\frac{1}{\rho(h)} \leq$ # labels needed $\leq \frac{1}{\rho(h)} \cdot \tilde{O}\left(d \log \frac{1}{\epsilon}\right)$

Example: Threshold functions on the line



Result: $\rho = 1/2$ for any target hypothesis and any input distribution

Slide credit: S. Dasgupta

Open problem: efficient, general AL

[M, COLT Open Problem '06]: Efficient algorithms for active learning under general input distributions, *D*.

$\rightarrow$ Current UB's for general distributions are based on *intractable* schemes!

Provide an algorithm such that w.h.p.:

1. After *L* label queries, algorithm's hypothesis *v* obeys:

   $P_{x \sim D}[v(x) \neq u(x)] < \varepsilon$.

2. *L* is at most the PAC sample complexity, and for a general class of input distributions, *L* is significantly lower.
3. Total running time is at most *poly*(d, 1/ε).

Specific variant: homogeneous linear separators, realizable case, D known to learner.

Open problem: efficient, general AL

[M, COLT Open Problem ‘06]: Efficient algorithms for active learning under general input distributions, *D*.

Other open variants:

- Input distribution, *D,* is unknown to learner.
- Agnostic case, certain scenarios ([Kääriäinen, NIPS Foundations of Active Learning workshop ‘05]: negative result for general agnostic setting).
- Add the online constraint: memory and time complexity (of the online update) must not scale with number of seen labels or mistakes.
- Same goal, other concept classes, or a general concept learner.

Other open problems

Extensions to DKM05:

- Relax distributional assumptions.
  - Uniform is sufficient but not necessary for proof.
- Relax realizable assumption.
- Analyze margin version
  - for exponential convergence, without d dependence.

Testing issue: Testing the final hypothesis takes $1/\varepsilon$ labels!

$\rightarrow$ Is testing an inherent part of active learning?

Cost-sensitive labels

Bridging theory and practice.

How to benchmark AL algorithms?

Thank you!