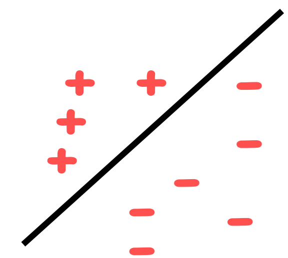
Active and passive learning of linear separators

Maria-Florina Balcan

Joint with Phil Long

2-Minute Version

New results for label efficient, poly time, passive and active learning of linear seps under log-concave distributions



- AL provides exponential improvement over passive learning.

⇩

- A poly time PAC algorithm with optimal sample complexity.

- Solves open question for the uniform distr. [Long'95,'03], [Bshouty'09]
- First tight bound for poly-time PAC algos for an infinite class of fns under a general class of distributions. [Ehrenfeucht et al., 1989; Blumer et al., 1989]

2-Minute Version

New improved bounds for active and passive learning in the case that the data might not be linearly separable.

- agnostic case (disagreement coefficient) and Tsybakov low-noise condition

Nearly log-concave distributions [Applegate&Kannan'91] [Caramanis&Mannor'07], new structural results there as well.

This talk focus on the noise-free setting, log-concave distributions .

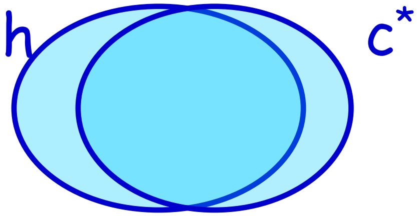
Supervised Learning Formalization

Classic models: PAC (Valiant), SLT (Vapnik)

- X - feature space
- $S=\{(x, l)\}$ - set of labeled examples
  - drawn i.i.d. from distr. D over X and labeled by target concept c^*

- Do optimization over S, find hypothesis $h \in C$.
- Goal: h has small error over D.

$$err(h)=Pr_{x \in D}(h(x) \neq c^*(x))$$



- c^* in C, realizable case; else agnostic
- In PAC, talk about efficient algorithms.

Sample Complexity Results

Infinite C, realizable

Theorem

$$m \geq \frac{d}{\epsilon} \log\left(\frac{1}{\epsilon}\right) + \frac{1}{\epsilon} \log\left(\frac{1}{\delta}\right)$$

labeled examples case are sufficient s.t. with prob. 1-δ all h in C consistent with data satisfy err(h) $\leq \epsilon$.

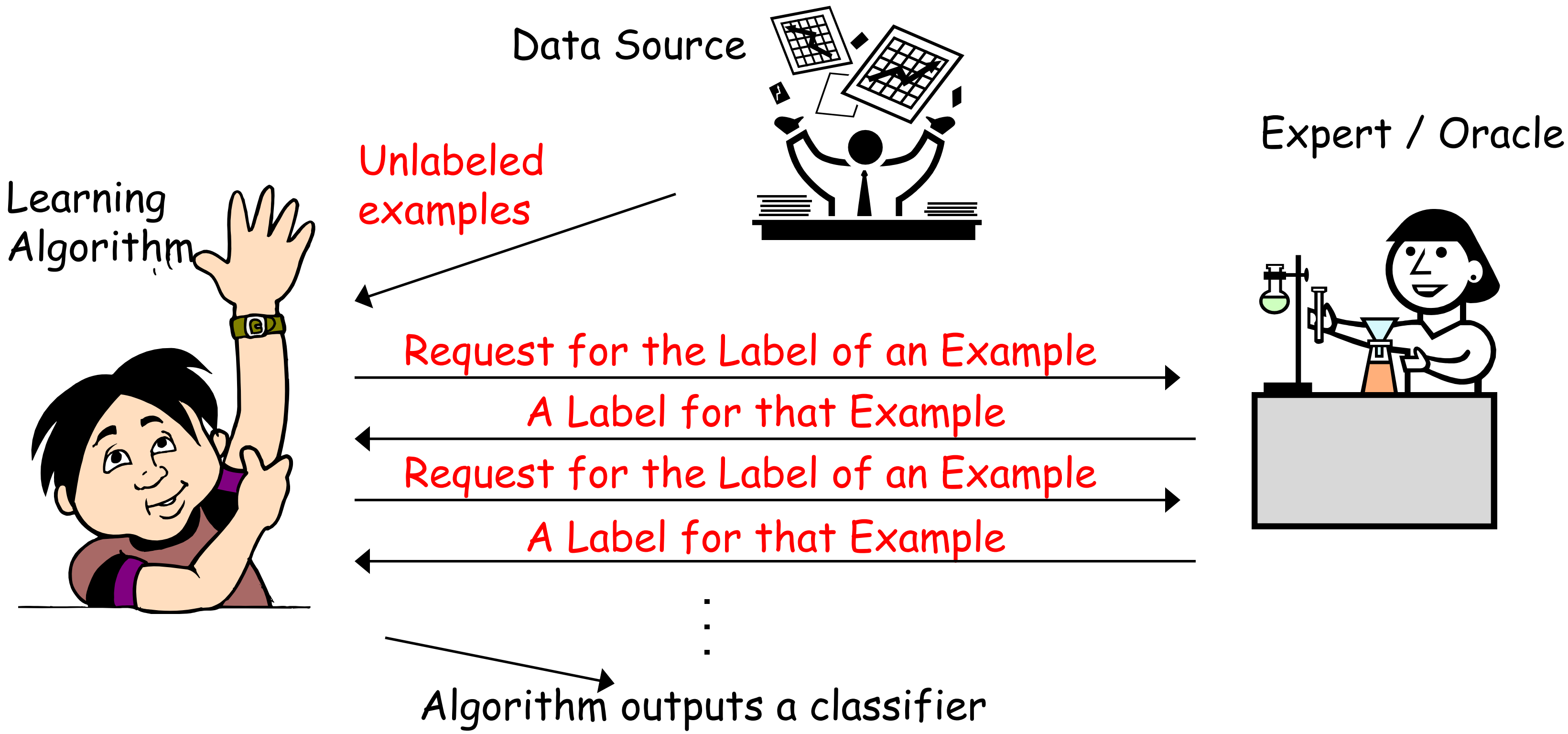
Theorem

Lower bound: $m \geq \frac{d}{\epsilon} + \frac{1}{\epsilon} \log\left(\frac{1}{\delta}\right)$, even for linear separators under uniform distribution.

- Lots of work on tighter bounds [e.g., Haussler,Littlestone,Wartmuth'94'; Gine and Koltchinski'06]

Still pesky gaps between upper and lower bounds, even for lin. separators under uniform distr.

Active Learning

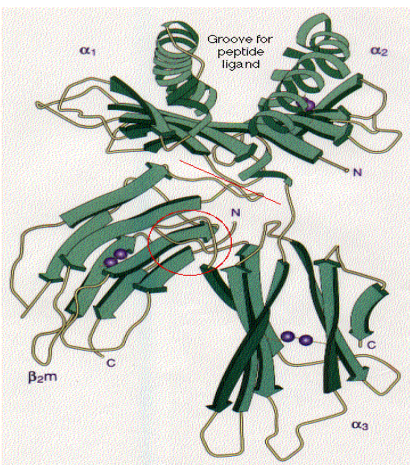


- The learner can choose specific examples to be labeled.
- He works harder, to use fewer labeled examples.

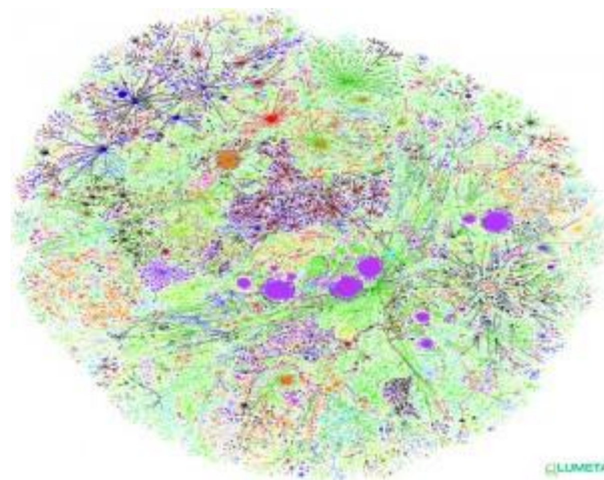
Classic Paradigm Insufficient Nowadays

Modern applications: massive amounts of raw data.

Only a tiny fraction can be annotated by human experts.



Protein sequences



Billions of webpages

Images

When Active Learning Helps

Lots of exciting activity in recent years.

- Several specific analyses for re[...]le case. E.g., linear separators, uniform distr[...]
  - QBC [Freun[...]
  - Active [...]eptron [Dasgupta, Kalai, Monteleoni '05]

Very Specific.

- Generic algos that work even in the agn[...] under various noise conditions
  - A^2 [Balcan, Beygelzimer, [...]] [Hanneke'10] [Wang'09]
  - DKM algo [D[...]] [Hanneke '10]
  - Kolt[...]nksi'10]

Typically suboptimal in query complexity.

Margin Based Active Learning

This talk: C - homogeneous linear seps in R^d, D – logconcave

- Realizable: exponential improvement, only $O(d \log 1/\varepsilon)$ labels to find a hypothesis with error ε. [Bounded noise].
- Tsybakov noise: polynomial improvement.

Log-concave distributions: log of density fnc concave

- wide class: includes uniform distr. over any convex set, Gaussian distr., Logistic, etc
- played a major role in sampling, optimization, integration, learning [LV'07, KKMS'05,KLT'09]

Margin Based Active Learning

This talk: C - homogeneous linear seps in R^d, D – logconcave

- Realizable: exponential improvement, only $O(d \log 1/\varepsilon)$ labels to find a hypothesis with error ε. [Bounded noise].
- Tsybakov noise: polynomial improvement.

Broadens the class of pbs for which we have concrete and optimal bounds for AL.

- Bounds show improvement in the $1/\epsilon$ factor without increase in the d factor.

Margin Based Active-Learning, Realizable Case

Algorithm

Draw m_1 unlabeled examples, label them, add them to $W(1)$.

iterate $k=2, \ldots, s$

- find a hypothesis w_{k-1} consistent with $W(k-1)$.
  - $W(k)=W(k-1)$.
  - sample m_k unlabeled samples x satisfying $|w_{k-1} \cdot x| \leq \gamma_{k-1}$;
  - label them and add them to $W(k)$.

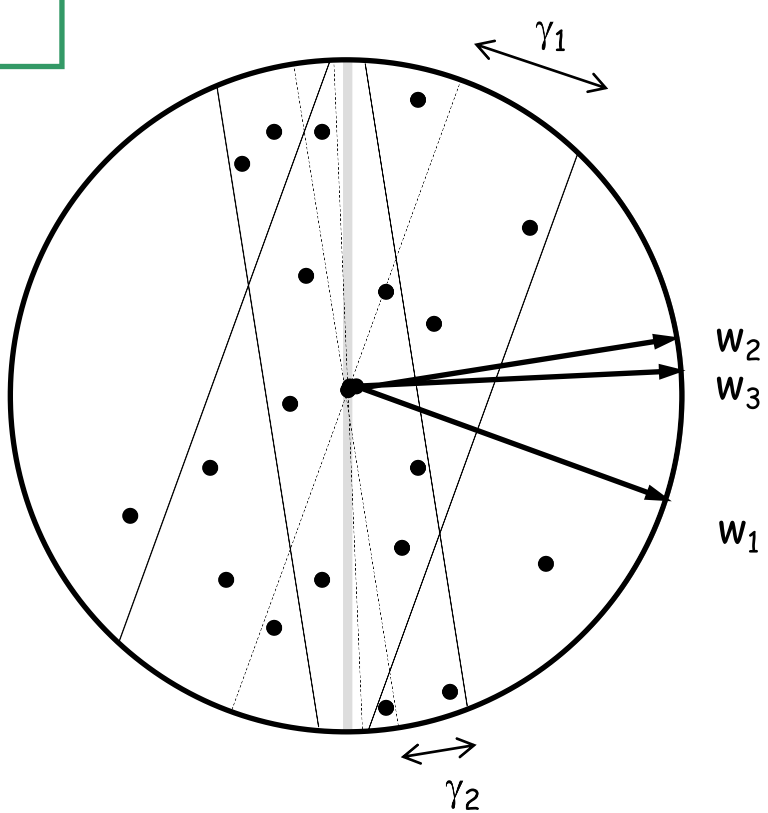
end iterate

Margin Based Active-Learning, Realizable Case

Draw m_1 unlabeled examples, label them, add them to $W(1)$.

iterate $k = 2, \ldots, s$

- find a hypothesis w_{k-1} consistent with $W(k-1)$.
- $W(k)=W(k-1)$.
- sample m_k unlabeled samples x satisfying $|w_{k-1}^T \cdot x| \leq \gamma_{k-1}$
- label them and add them to $W(k)$.



Margin Based Active-Learning, Realizable Case

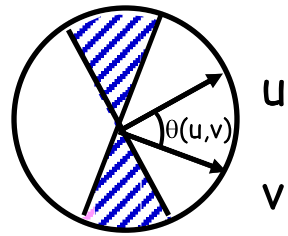
Theorem P_X log-concave in R^d.

If $\gamma_k = O\left(\frac{c}{2^k}\right)$ and $m_k = O\left(d + \log\log(1/\epsilon)\right)$ then after $s = \log\left(\frac{1}{\epsilon}\right)$ iterations w_s has error $\leq \varepsilon$.

Linear Separators, Log-Concave Distributions

Fact 1

$$d(u, v) \approx \frac{\theta(u,v)}{\pi}$$

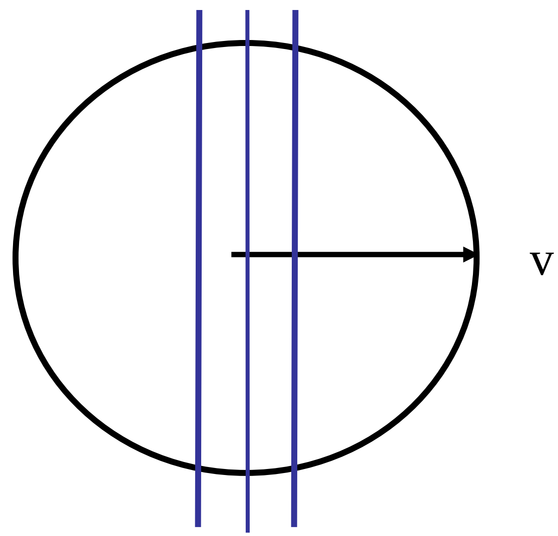


Proof idea:

- project the region of disagreement in the space given by u and v
- use properties of log-concave distributions in 2 dimensions.

Fact 2

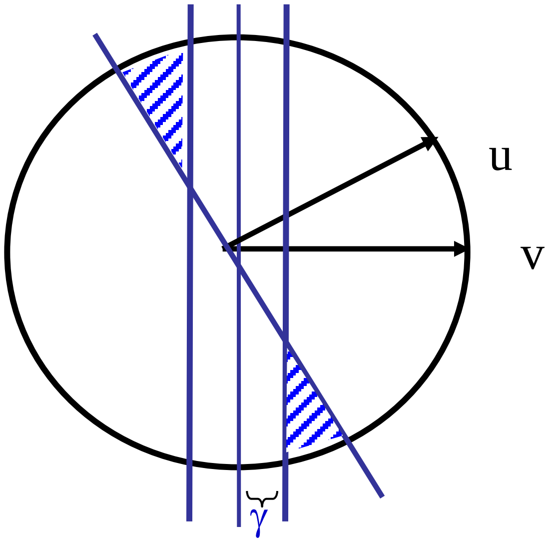
$$\Pr_X[|v \cdot x| \leq \gamma] \leq \gamma.$$



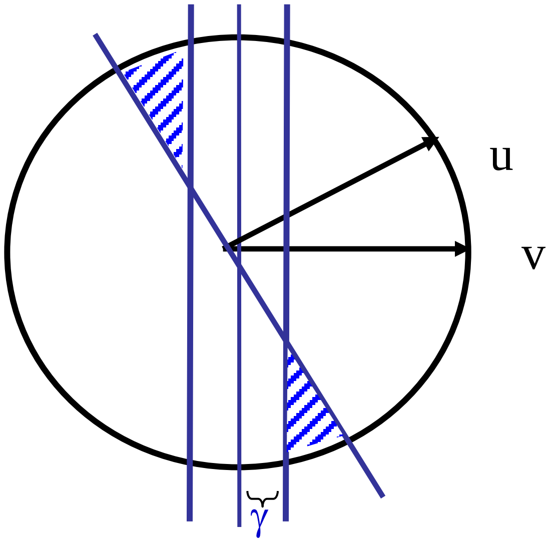
Linear Separators, Log-Concave Distributions

$\boxed{\text{Fact 3}}$ If $\theta(u, v) = \beta$ and $\gamma = C\beta$

$$\Pr_x\left[(u \cdot x)(v \cdot x) < 0, |v \cdot x| \geq \gamma\right] \leq \frac{\beta}{4}.$$



Linear Separators, Log-Concave Distributions



$\boxed{\text{Fact 3}}$ If $\theta(u, v) = \beta$ and $\gamma = C\beta$

E

$$\Pr_X [(u \cdot x)(v \cdot x) < 0, |v \cdot x| \geq \gamma] \leq \frac{\beta}{4}.$$

Proof idea:

- project the region of disagreement in the space given by u and v
- Note that each x in E has $||x|| \geq \gamma/\beta = c_2$

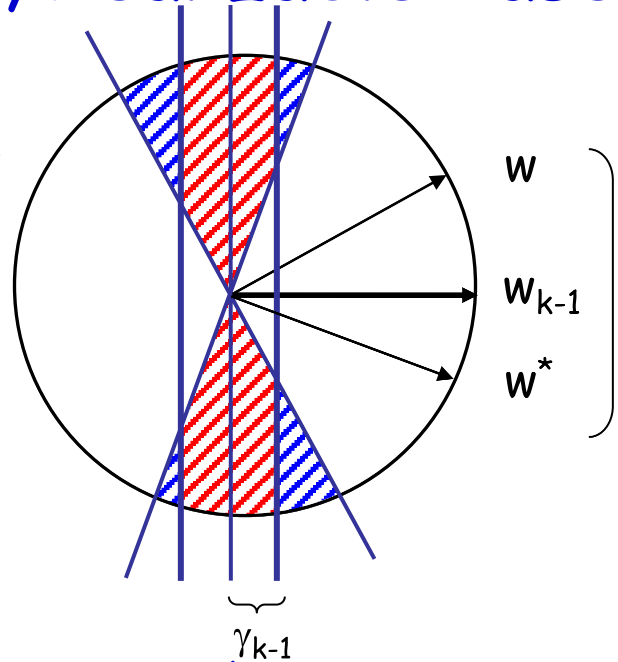
$$\Pr_X [x \in E] = \sum_{i=1}^{\infty} \Pr [E \cap (B((i+1)c_2) - B(ic_2))]$$

$$\leq C\beta(i+1)^2 \exp[-Ci]$$

Margin Based Active-Learning, Realizable Case

iterate $k=2, \ldots, s$

- find a hypothesis w_{k-1} consistent with $W(k-1)$.
- $W(k)=W(k-1)$.
- sample m_k unlabeled samples x satisfying $|w_{k-1}^T \cdot x| \leq \gamma_{k-1}$
- label them and add them to $W(k)$.



Proof Idea

Induction: all w consistent with $W(k)$ have error $\leq 1/2^k$;

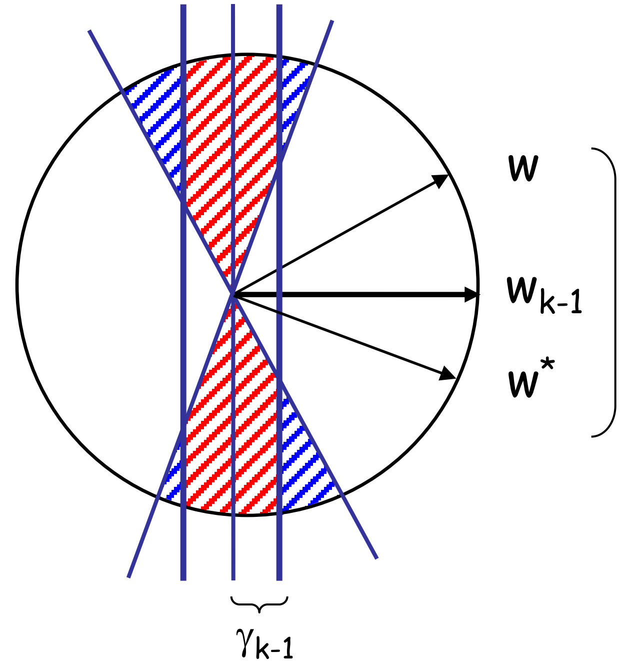
so, w_k has error $\leq 1/2^k$.

For $\gamma_k = O\left(\frac{c}{2^k}\right)$

$$\mathrm{err}(w) = \underbrace{\Pr(w \text{ errs on x }, |w_{k-1} \cdot x| \geq \gamma_{k-1})}_{\leq 1/2^{k+1}} + \Pr(w \text{ errs on x }, |w_{k-1} \cdot x| \leq \gamma_{k-1})$$

Proof Idea

Under the uniform distr. for $\gamma_k = O\left(\frac{c}{2^k}\right)$

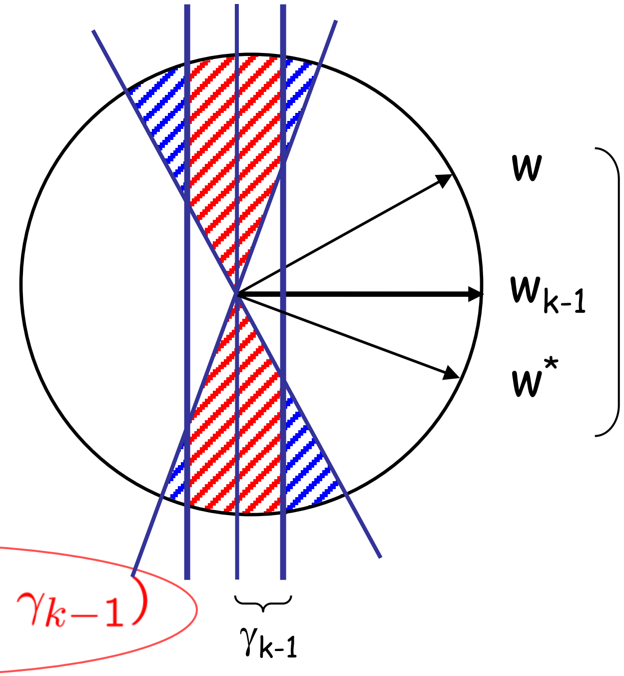


$$\mathrm{err}(w) = \underbrace{\Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \geq \gamma_{k-1})}^{\leq 1/2^{k+1}} +$$

$$\Pr(w \text{ errs on } x, |w_{k-1} \cdot x| \leq \gamma_{k-1})$$

Proof Idea

Under the uniform distr. for $\gamma_k = O\left(\frac{c}{2^k}\right)$



$$\mathrm{err}(w) = \overbrace{\Pr(w \text{ errs on } x,\ |w_{k-1}\cdot x| \geq \gamma_{k-1})}^{\leq 1/2^{k+1}} +$$

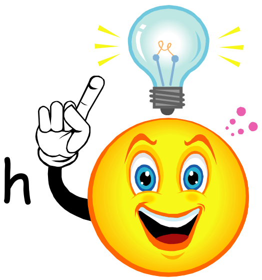
$$\Pr(w \text{ errs on } x \mid |w_{k-1}\cdot x| \leq \gamma_{k-1}) \underbrace{\Pr(|w_{k-1}\cdot x| \leq \gamma_{k-1})}_{\leq C\gamma_{k-1}.}$$

Enough to ensure

$$\Pr(w \text{ errs on } x \mid |w_{k-1}\cdot x| \leq \gamma_{k-1}) \leq C_1$$

Can do with only $m_k = O(d + \log\log(1/\epsilon))$ labels.

Passive Learning



Theorem

Any passive learning algo that outputs w consistent with

$$\frac{d}{\epsilon} + \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)$$ examples, satisfies err(w) $\leq \epsilon$, with proab. 1-δ.

High Level Idea

- Run algo online, use the intermediate w to track the progress
- Performs well even if it periodically builds w using some of the examples, and only uses borderline cases for preliminary classifiers for further training.
- Carefully distribute δ, allow higher prob. of failure in later stages [once w is already pretty good, it takes longer to get examples that help to further improve it]

Passive Learning

Theorem

Any passive learning algo that outputs w consistent with

$\frac{d}{\epsilon} + \frac{1}{\epsilon} \log\left(\frac{1}{\delta}\right)$ examples, satisfies err(w) $\leq \epsilon$, with proab. 1-δ.

High Level Idea

$$m_k = C_2 (d + \log((1 + s - k)/\delta)) \quad b_k = C_1/2^k$$

$$\sum_{k=1}^{s} 2^k (d + \log((1 + s - k)/\delta))) =$$

$$O(2^s (d + \log(1/\delta)) + \sum_{k=1}^{s} 2^k \log(1 + s - k)$$

$O(1/\epsilon)$

Discussion, Open Directions

- Broadens class of pbs for which AL provides exponential improvement in $1/\epsilon$ (without additional increase on d).
- First tight bound for a poly-time PAC algo for an infinite class of fns under a general class of distributions.
- Extensions to nearly log-concave distributions, noisy settings. Lower Bounds.

Open Directions

- Efficient query optimal algorithms for AL for more general settings.
- Close the existing gaps for passive learning for general classes and distributions.