Foundations for Learning in the Age of Big Data

Maria-Florina (Nina) Balcan
Carnegie Mellon University

Today's topic: Active Learning (AL)

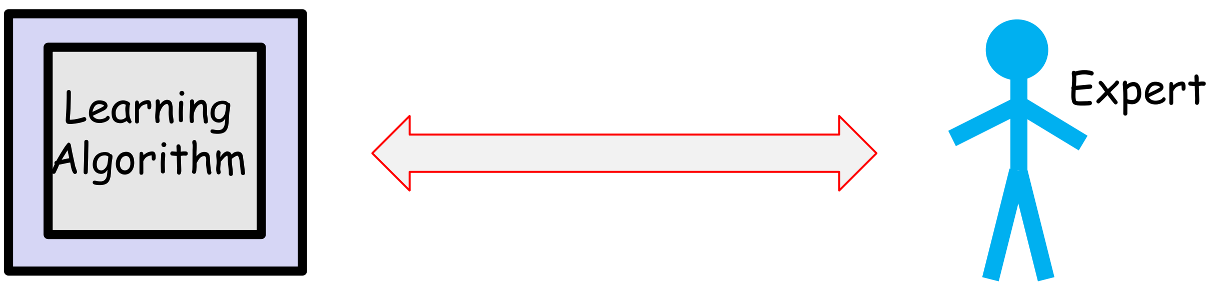
AL: learning algo takes a much more active role than in classic supervised learning in order to minimize the need for expert intervention.

Classic Fully Supervised Learning Paradigm Insufficient

- Modern applications: massive amounts of raw data.
  - E.g., billions of webpages; massive collections of images
- Only a tiny fraction can be annotated by human experts.

Modern ML: New Learning Approaches

- Modern applications: massive amounts of raw data.
- **Techniques that best utilize data, minimizing need for expert/human intervention.**
- Paradigms where there has been great progress.
  - Semi-supervised Learning, (Inter)active Learning.

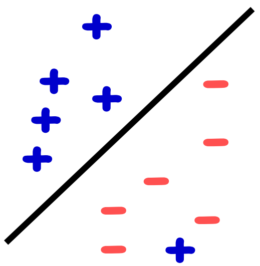


Active Learning

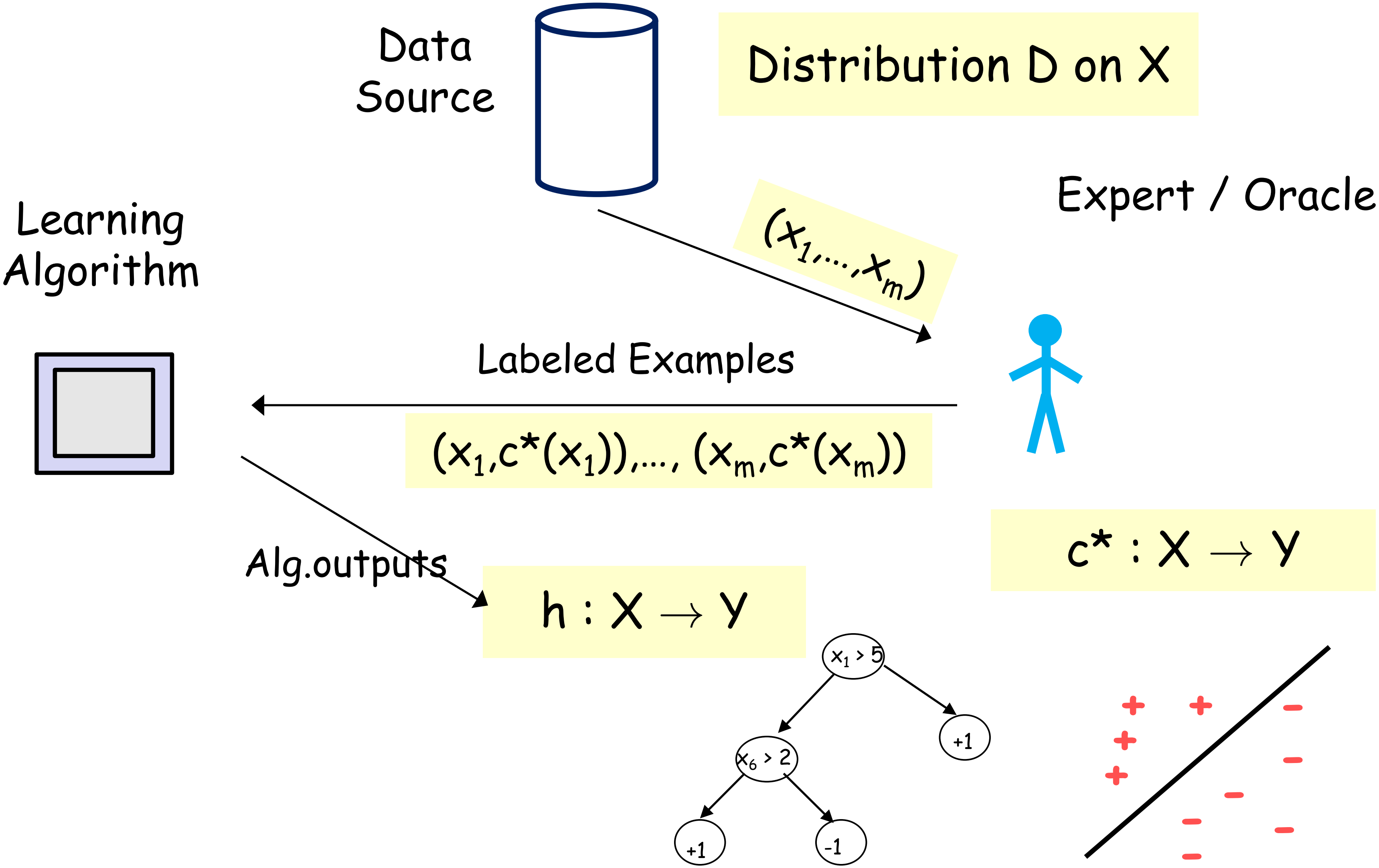
Lots of exciting activity in recent years on understanding the power of active learning. Mostly label efficiency.

This lecture: provable guarantees for active learning.

- Disagreement based active learning.
- Power of aggressive localization for label efficient and poly time active learning for linear separators.



PAC/SLT models for Supervised Learning



Two Main Aspects in Classic Machine Learning

Algorithm Design. How to optimize?

Automatically generate rules that do well on observed data.

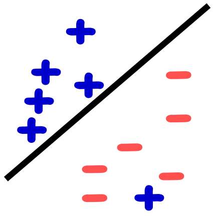
Runing time: $\text{poly}\left(d, \frac{1}{\epsilon}, \frac{1}{\delta}\right)$

Generalization Guarantees, Sample Complexity

Confidence for rule effectiveness on future data.

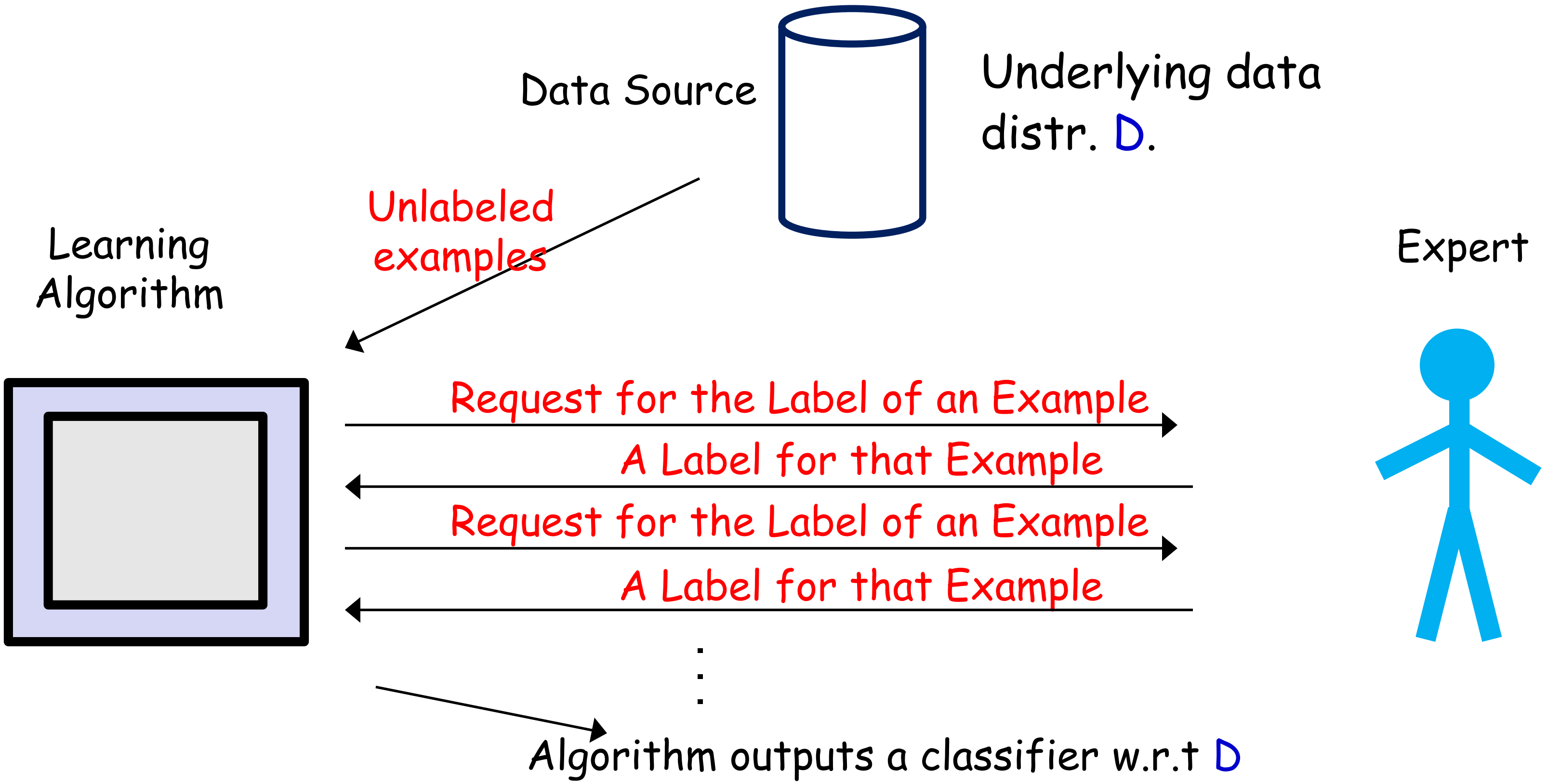
Realizable: $O\left(\frac{1}{\epsilon}\left(\text{VCdim}(H)\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$

E.g, C= linear separators in R^d: $O\left(\frac{1}{\epsilon}\left(d\ \log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right)\right)$



Agnostic: $O\left(\frac{1}{\epsilon^2}\left(\text{VCdim}(H) + \log\left(\frac{1}{\delta}\right)\right)\right)$

Active Learning



- Learner can choose specific examples to be labeled.
- Goal: use fewer labeled examples [pick informative examples to be labeled].

What Makes a Good Active Learning Algorithm?

- Guaranteed to output a good classifier for most learning problems.

- Doesn't make too many label requests.

  Hopefully a lot less than fully supervised passive learning.

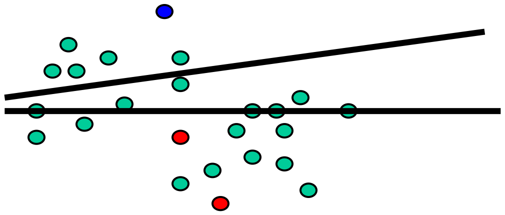
- Need to choose the label requests carefully, to get informative labels.

Can adaptive querying really do better than passive sampling?

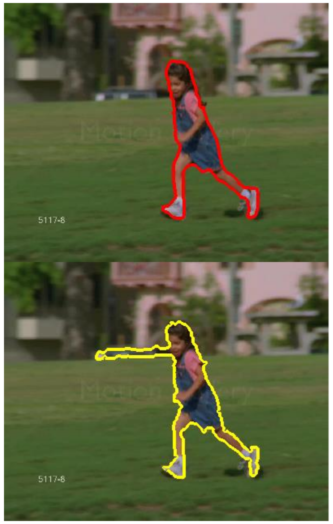
- YES! (sometimes)
- We often need far fewer labels for active learning than for passive.
- This is predicted by theory and has been observed in practice.

Active Learning in Practice

- Text classification: active SVM (Tong-Koller, ICML2000).
  - e.g., request label of the example closest to current separator.

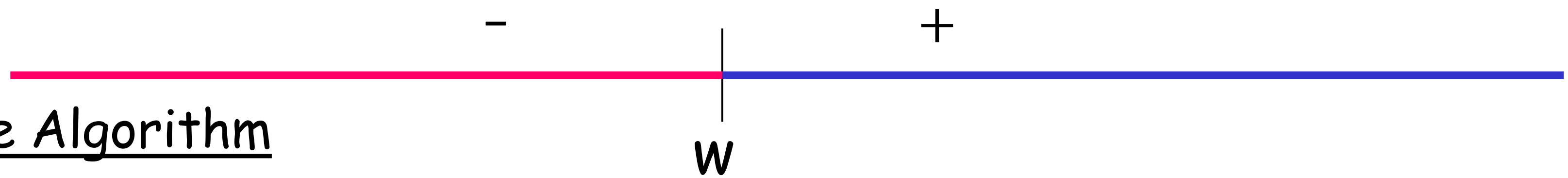


- Video Segmentation (Fathi-Balcan-Ren-Regh, BMVC 11).



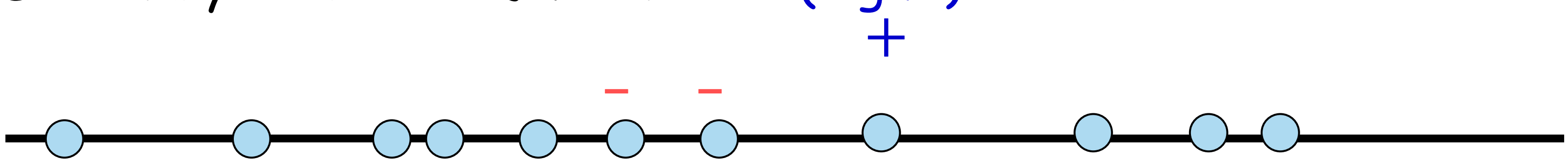
Can adaptive querying help? [CAL92, Dasgupta04]

- Threshold fns on the real line: $h_w(x) = 1(x \geq w)$, $H = \{h_w : w \in R\}$



Active Algorithm

- Get N unlabeled examples
- How can we recover the correct labels with $\ll N$ queries?
- Do binary search! Just need O(log N) labels!



- Output a classifier consistent with the N inferred labels.

- $N = O(1/\epsilon)$ we are guaranteed to get a classifier of error $\leq \epsilon$.

Passive supervised: $\Omega(1/\epsilon)$ labels to find an ε-accurate threshold.

Active: only $O(\log 1/\epsilon)$ labels. Exponential improvement.

Active Learning, Provable Guarantees

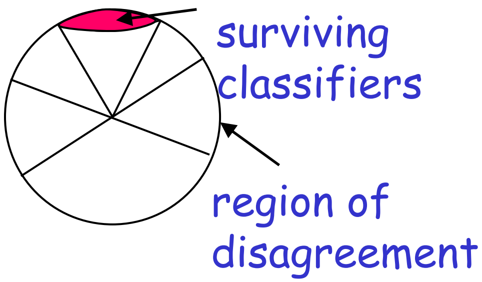
Lots of exciting results on sample complexity. E.g.,

- DasguptaKalaiMonteleoni'05, CastroNowak'07, CavallantiCesa-BianchiGentile'10, YanChaudhuriJavidi'16

- DasguptaHsu'08, UrnerWulffBenDavid'13

- "Disagreement based" algorithms

Pick a few points at random from the current region of disagreement (uncertainty), query their labels, throw out hypothesis if you are statistically confident they are suboptimal.

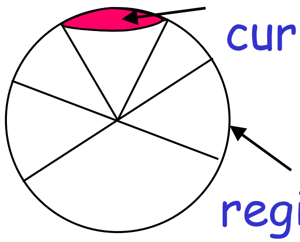


[BalcanBeygelzimerLangford'06, Hanneke07, DasguptaHsuMontleoni'07, Wang'09, Fridman'09, Koltchinskii10, BHW'08, BeygelzimerHsuLangfordZhang'10, Hsu'10, Ailon'12, …]

Disagreement Based Active Learning

A^2 Agnostic Active Learner

[Balcan-Beygelzimer-Langford, ICML 2006]



Let $H_1 = H$.

For $\mathbf{t = 1}$, ….,

- Pick a few points at random from current region of disagreement $DIS(H_t)$ and query their labels.
- Throw out hypothesis if statistically confident they are suboptimal.

Disagreement Based Active Learning

A^2 Agnostic Active Learner [Balcan-Beygelzimer-Langford, ICML 2006]

Pick a few points at random from the current region of disagreement (uncertainty), query their labels, throw out hypothesis if you are statistically confident they are suboptimal.

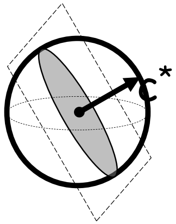
Guarantees for A^2 [Hanneke'07]:

Disagreement coefficient: $\theta_{c^*} = \sup_{r > \eta + \epsilon} \frac{P(DIS(B(c^*, r)))}{r}$

Realizable: $m = VCim(C)\theta_{c^*}\log(1/\epsilon)$

Agnostic: $m = \frac{\eta^2}{\epsilon^2} VCim(C)\theta_{c^*}^2\log(1/\epsilon)$

Linear separators, uniform distr.: $\theta_{c^*} = \sqrt{d}$



Disagreement Based Active Learning

Pick a few points at random from the current region of disagreement (uncertainty), query their labels, throw out hypothesis if you are statistically confident they are suboptimal.

[BalcanBeygelzimerLangford'06, Hanneke07, DasguptaHsuMontleoni'07, Wang'09, Fridman'09, Koltchinskii10, BHW'08, BeygelzimerHsuLangfordZhang'10, Hsu'10, Ailon'12, …]

Positives

- Generic (any class),
- adversarial label noise.

Negatives

- suboptimal in label complexity
- computationally prohibitive.

Key Question:

Poly Time, Noise Tolerant/Agnostic, Label Optimal AL Algos?

Margin Based Active Learning

Margin based algo for learning linear separators.

[Balcan-Long, COLT 2013] [Awasthi-Balcan-Long, STOC 2014] [Awasthi-Balcan-Haghtalab-Urner, COLT15] [Awasthi-Balcan-Haghtalab-Zhang, COLT16] [Awasthi-Balcan-Long, JACM 2017]

- Realizable: exponential improvement, only $O(d \log 1/\epsilon)$ labels to find w error ε, when D logconcave

- Agnostic: poly-time AL algo outputs w with $err(w) = O(\eta + \epsilon)$, η=err(best lin. sep), $O(d \log 1/\epsilon)$ labels when D logconcave.

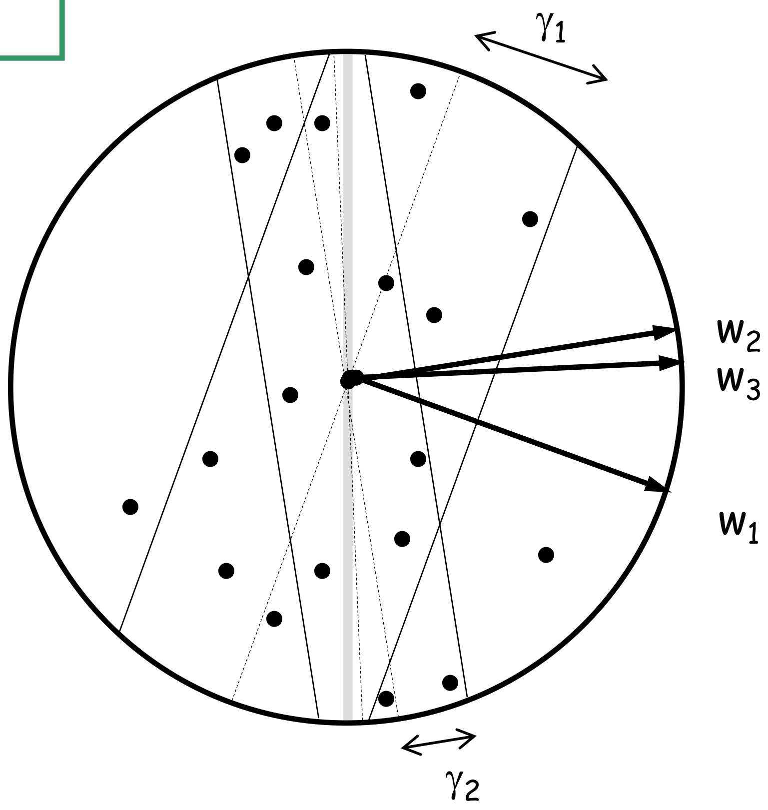
- Improves on noise tolerance of previous best passive [KKMS'05], [KLS'09] algos too!

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 $|w_{k-1} \cdot x| \le \gamma_{k-1}$
- label them and add them to $W(k)$.



Margin Based Active-Learning, Realizable Case

Log-concave distributions: log of density fnc concave.

- wide class: uniform distr. over any convex set, Gaussian, etc.

$$f(\lambda x_1 + (1 - \lambda x_2)) \geq f(x_1)^{\lambda} f(x_2)^{1-\lambda}$$

Theorem D log-concave in R^d. If $\gamma_k = O\left(\frac{1}{2^k}\right)$ then $err(w_s) \leq \epsilon$ after $s = \log\left(\frac{1}{\epsilon}\right)$ rounds using $\tilde{O}(d)$ labels per round.

Active learning

$O\left(d \log\left(\frac{1}{\epsilon}\right)\right)$ label requests

$\Theta\left(\frac{d}{\epsilon}\right)$ unlabeled examples

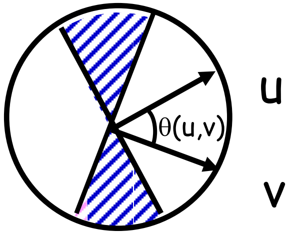
Passive learning

$\Theta\left(\frac{d}{\epsilon}\right)$ label requests

Linear Separators, Log-Concave Distributions

Fact 1

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

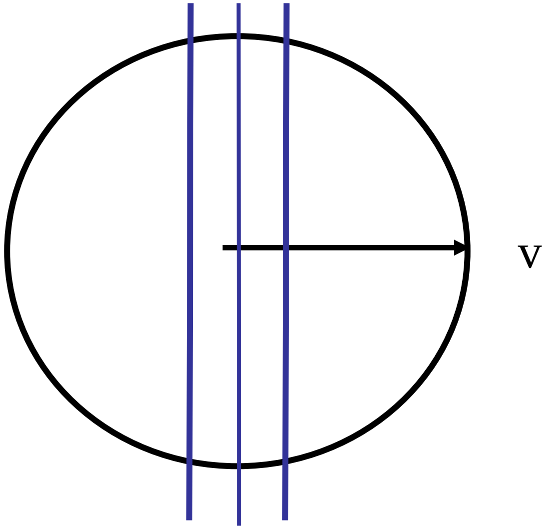


Proof idea:

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

Fact 2

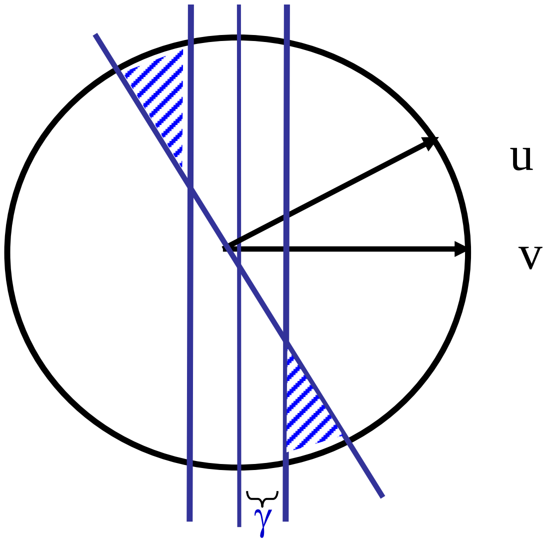
$$\Pr_X\left[|v \cdot x| \leq \gamma\right] \leq \gamma.$$



Linear Separators, Log-Concave Distributions

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

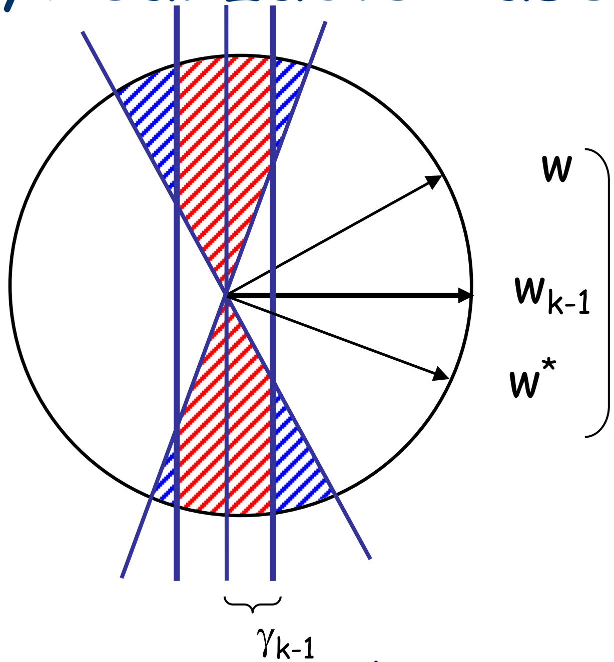
$$\Pr_x [(u \cdot x)(v \cdot x) < 0, |v \cdot x| \geq \gamma] \leq \frac{\beta}{4}.$$



Margin Based Active-Learning, Realizable Case

iterate k=2, … ,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).



Proof Idea

Induction: all w consistent with W(k) have error at most $1/2^k$;
so, w_k has error at most $1/2^k$.

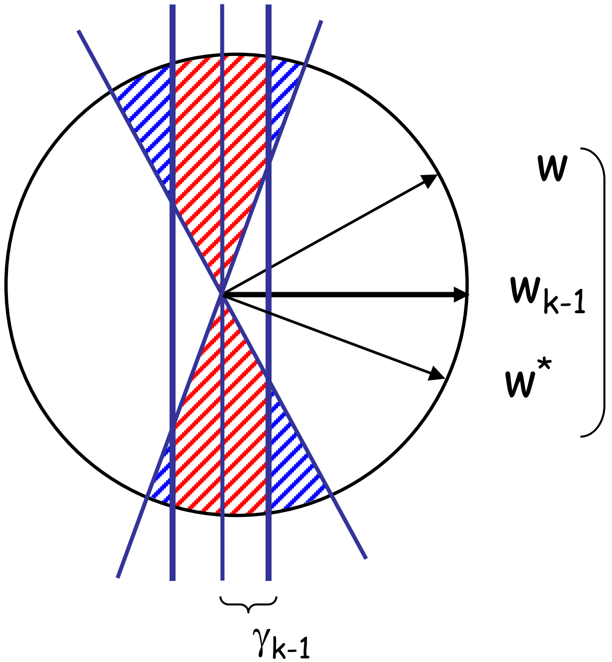
For $\gamma_k = O\left(\frac{c}{2^k}\right)$

$1/2^{k+1}$

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

Proof Idea

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

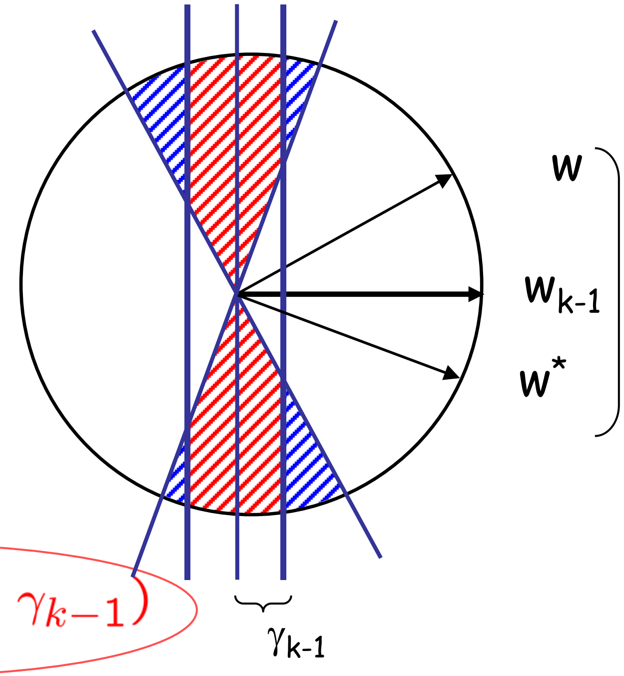


$1/2^{k+1}$

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

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

Proof Idea



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

$1/2^{k+1}$

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

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

$$\leq C\gamma_{k-1}.$$

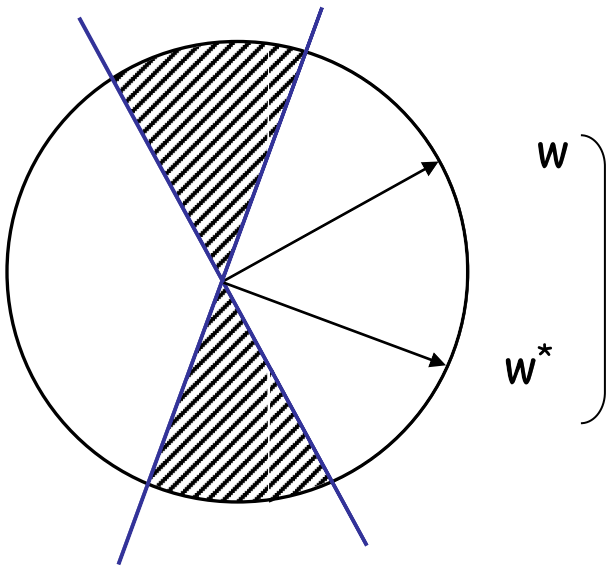
Enough to ensure

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

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

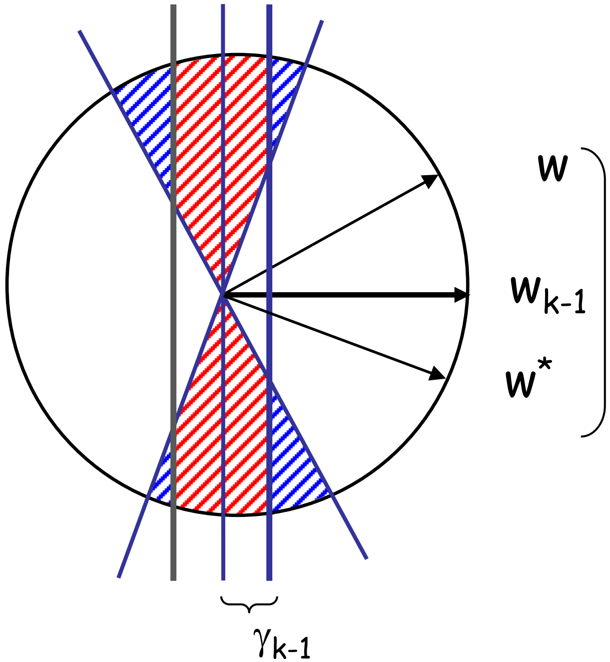
Key Insight: Aggressive Localization

Induction: all w consistent with $W(k)$, $err(w) \leq 1/2^k$

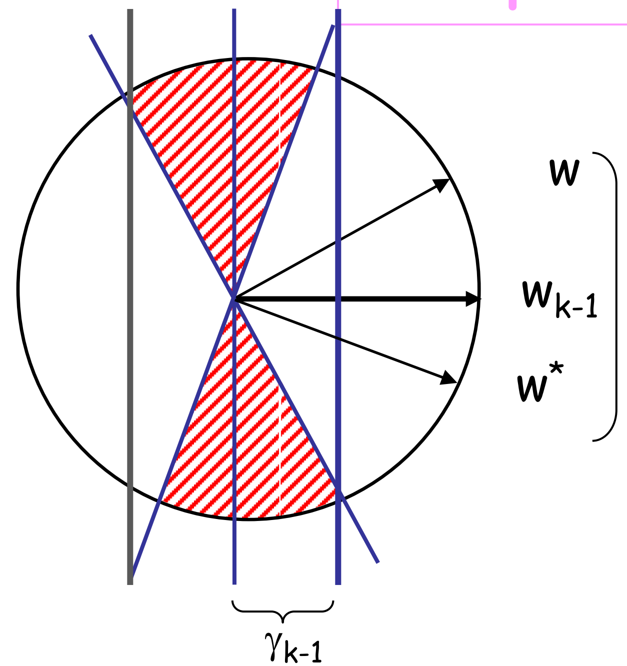


Key Insight: Aggressive Localization

Induction: all w consistent with W(k), err(w) ≤ 1/2^k

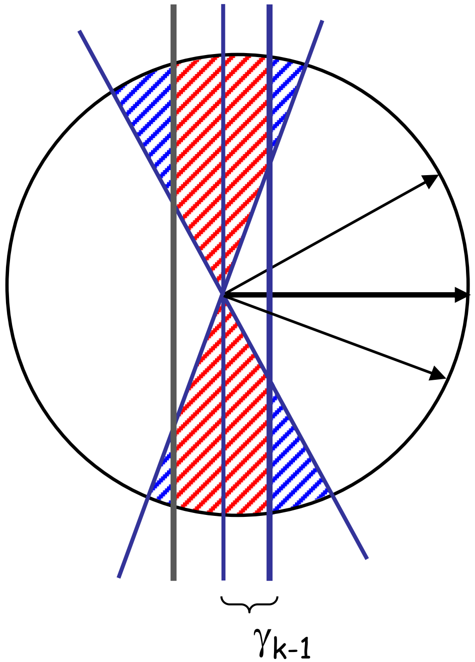


Suboptimal



Key Insight: Aggressive Localization

Induction: all w consistent with W(k), err(w) ≤ $1/2^k$

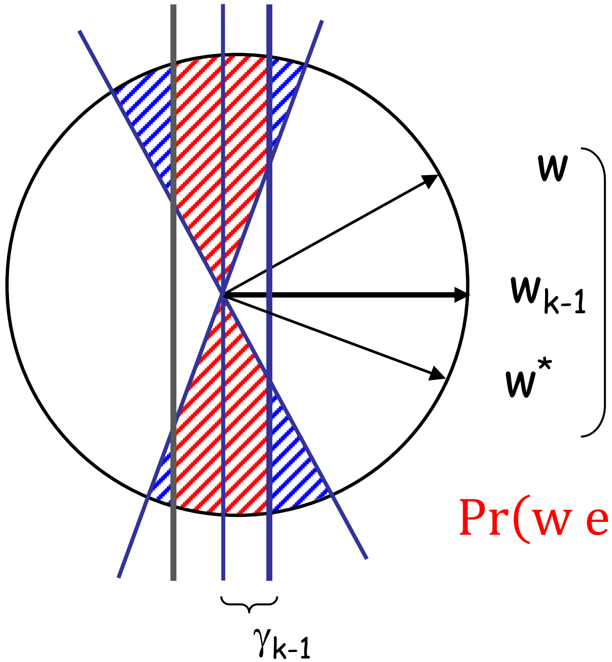


$1/2^{k+1}$

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

Key Insight: Aggressive Localization

Induction: all w consistent with W(k), err(w) $\leq$ $1/2^k$



$1/2^{k+1}$

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

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

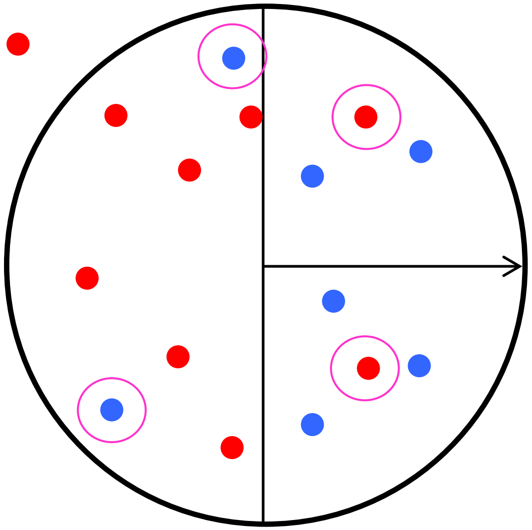
Enough to ensure $\Pr(w \text{ errs on } x \mid |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq C$

Need only $m_k = \tilde{O}(d)$ labels in round k.

Key point: localize aggressively, while maintaining correctness.

The Agnostic Case

- No linear separator can separate ● and ●
- Best linear separator error η



Algorithm still margin based

Draw m_1 unlabeled examples, label them, add them to W.

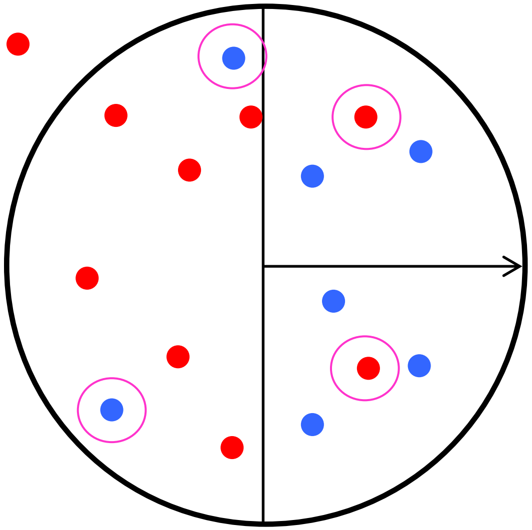
iterate k=2, …, s

- find w_{k-1} in $B(w_{k-1}, r_{k-1})$ of small τ_{k-1} hinge loss wrt W.
  - Clear working set.
  - sample m_k unlabeled samples x
    satisfying $|w_{k-1} \cdot x| \leq \gamma_{k-1}$;
  - label them and add them to W.

end iterate

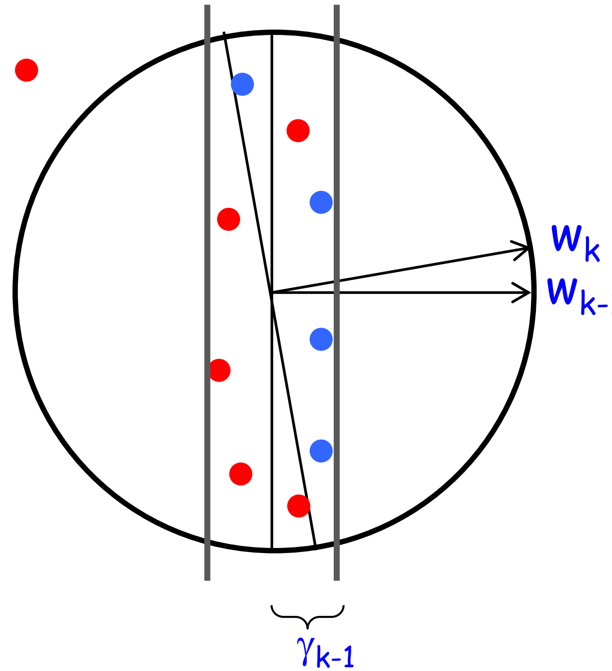
The Agnostic Case

- No linear separator can separate ● and ●
- Best linear separator error η



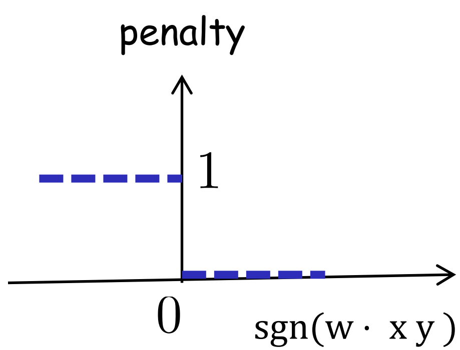
The Agnostic Case

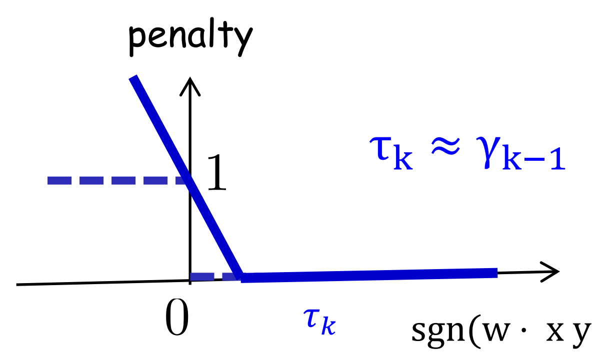
- No linear separator can separate ● and ●
- Best linear separator error η



Key idea 1:

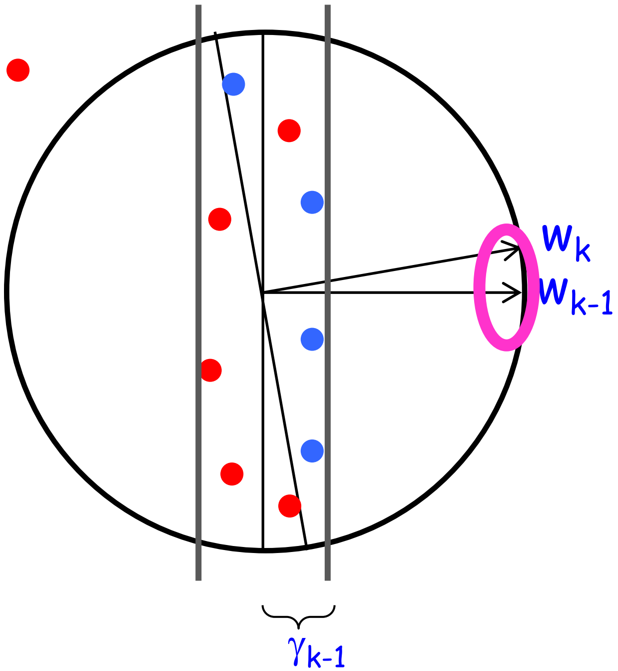
Replace error minimization with hinge loss minimization





The Agnostic Case

- No linear separator can separate ● and ●
- Best linear separator error η



Key idea 2:

Stay close to the current guess: w_k in small ball around w_{k-1}

Margin Based Active-Learning, Agnostic Case

Draw m_1 unlabeled examples, label them, add them to W.

iterate $k=2, \ldots, s$

- find w_{k-1} in $B(w_{k-1}, r_{k-1})$ of small τ_{k-1} hinge loss wrt W.
- Clear working set.
- sample m_k unlabeled samples x satisfying $|w_{k-1} \cdot x| \leq \gamma_{k-1}$;
- label them and add them to W.

end iterate

Localization in concept space.

Localization in instance space.

<u>Analysis, key idea:</u>

- Pick $\tau_k \approx \gamma_k$
- Localization & variance analysis control the gap between hinge loss and 0/1 loss (only a constant).

Improves over Passive Learning too!

Passive Learning	**Prior Work**	**Our Work**
Malicious	$\mathrm{err}(w) = O\left(\eta^{2/3}\log^{2/3}(d/\eta)\right)$ [KLS'09]	$\mathrm{err}(w) = O(\eta)$ Info theoretic optimal
Agnostic	$\mathrm{err}(w) = O\left(\eta^{1/3}\log^{1/3}(1/\eta)\right)$ [KLS'09]	$\mathrm{err}(w) = O(\eta)$
Bounded Noise $\lvert P(Y=1\lvert x) - P(Y=-1\lvert x)\rvert \geq \beta$	NA	$\eta + \epsilon$
Active Learning [agnostic/malicious/bounded]	NA	same as above! Info theoretic optimal

Key insights:

Localization both algorithmic and analysis tool!

Useful for active and passive learning!

- Well known for analyzing sample complexity.

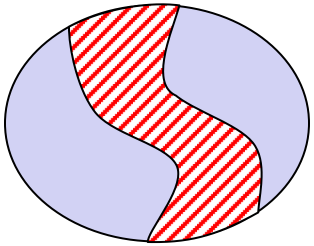
  [Bousquet-Boucheron-Lugosi'04], [BBL'06], [Hanneke'07], …

- We show useful for noise tolerant poly time algorithms.
  - Previously observed in practice, e.g. for SVMs [Bottou'07]

Further Margin Based Refinements

- Efficient algorithms
  - [Hanneke-Kanade-Yang, ALT'15] $O(d \log 1/\epsilon)$ label complexity for our poly time agnostic algorithm.
  - [Daniely, COLT '15] achieve $C\,\eta$ for any constant C in poly time in the agnostic setting [tradeoff running time with accuracy].
  - [Awasthi-Balcan-Haghtalab-Urner, COLT15], [Awasthi-Balcan-Haghtalab-Zhang, COLT16]

    Bounded noise, get error $\eta + \epsilon$ in time $\mathrm{poly}(d, 1/\epsilon)$
  - [YanZhang, NIPS'17] modified Perceptron enjoys similar guarantees.
- Active & Differentially Private [Balcan-Feldman, NIPS'13]
- General concept spaces: [Zhang-Chaudhuri, NIPS'14].
  - Compute the region to localize in each round by using unlabeled data and writing an LP

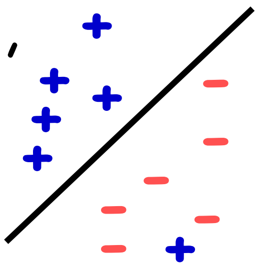


Discussion, Open Directions

AL: important paradigm in the age of Big Data. Lots of exciting developments.

- Disagreement based AL, general sample complexity.

- Margin based AL: label efficient, noise tolerant, poly time algo for learning linear separators.



  - Better noise tolerance than known for passive learning via poly time algos. [KKMS'05] [KLS'09]

  - Solve an adaptive sequence of convex optimization pbs on smaller & smaller bands around current guess for target.