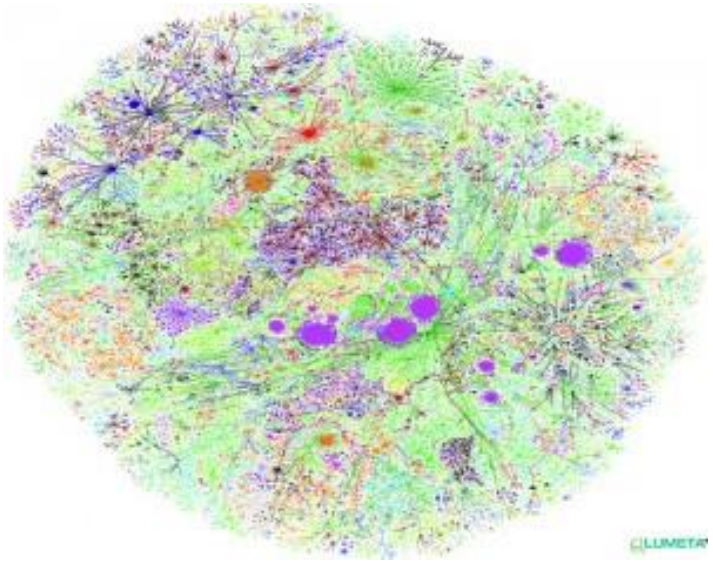
Foundations For Learning in the Age of Big Data

Maria-Florina Balcan

Modern Machine Learning

New applications

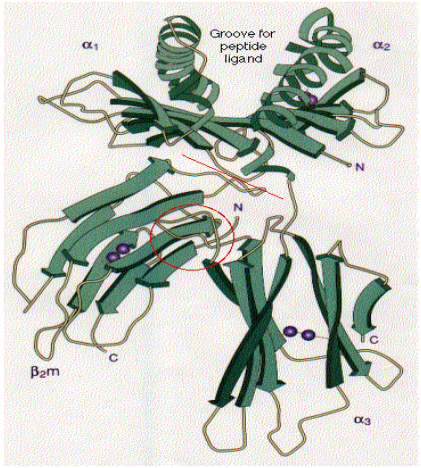
Explosion of data



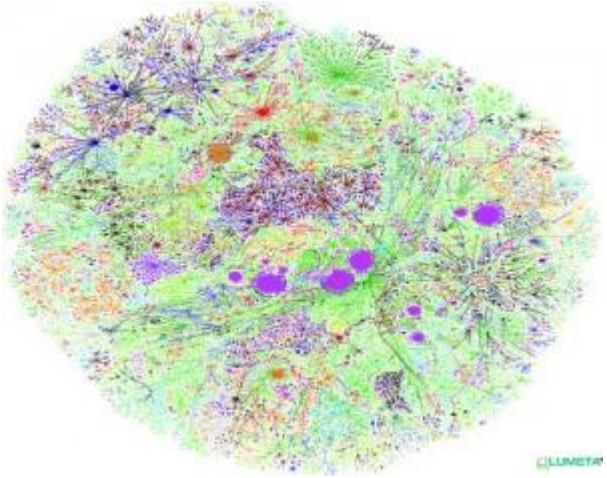
Classic Paradigm Insufficient Nowadays

Modern applications: massive amounts of raw data.

Only a tiny fraction can be annotated by human experts.



Protein sequences



Billions of webpages

Images

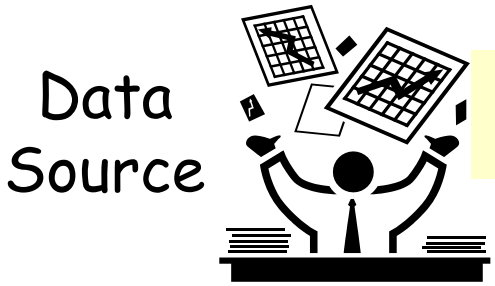
Classic Paradigm Insufficient Nowadays

Modern applications: massive amounts of data distributed across multiple locations

Outline of the talk

- Interactive Learning
  - Noise tolerant poly time active learning algos.
  - Implications to passive learning.
  - Learning with richer interaction.
- Distributed Learning
  - Model communication as key resource.
  - Communication efficient algos.

Supervised Classification

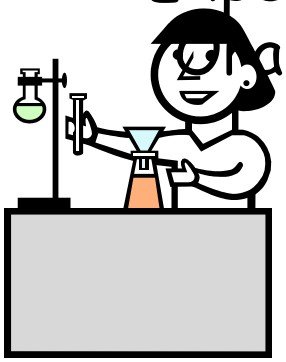


Data Source

Distribution D on X

$(x_1,\ldots,x_m)$

Expert / Oracle



Learning Algorithm

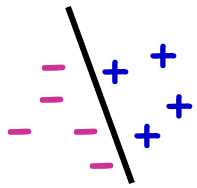
Labeled Examples

$(x_1,c^*(x_1)),\ldots,(x_m,c^*(x_m))$

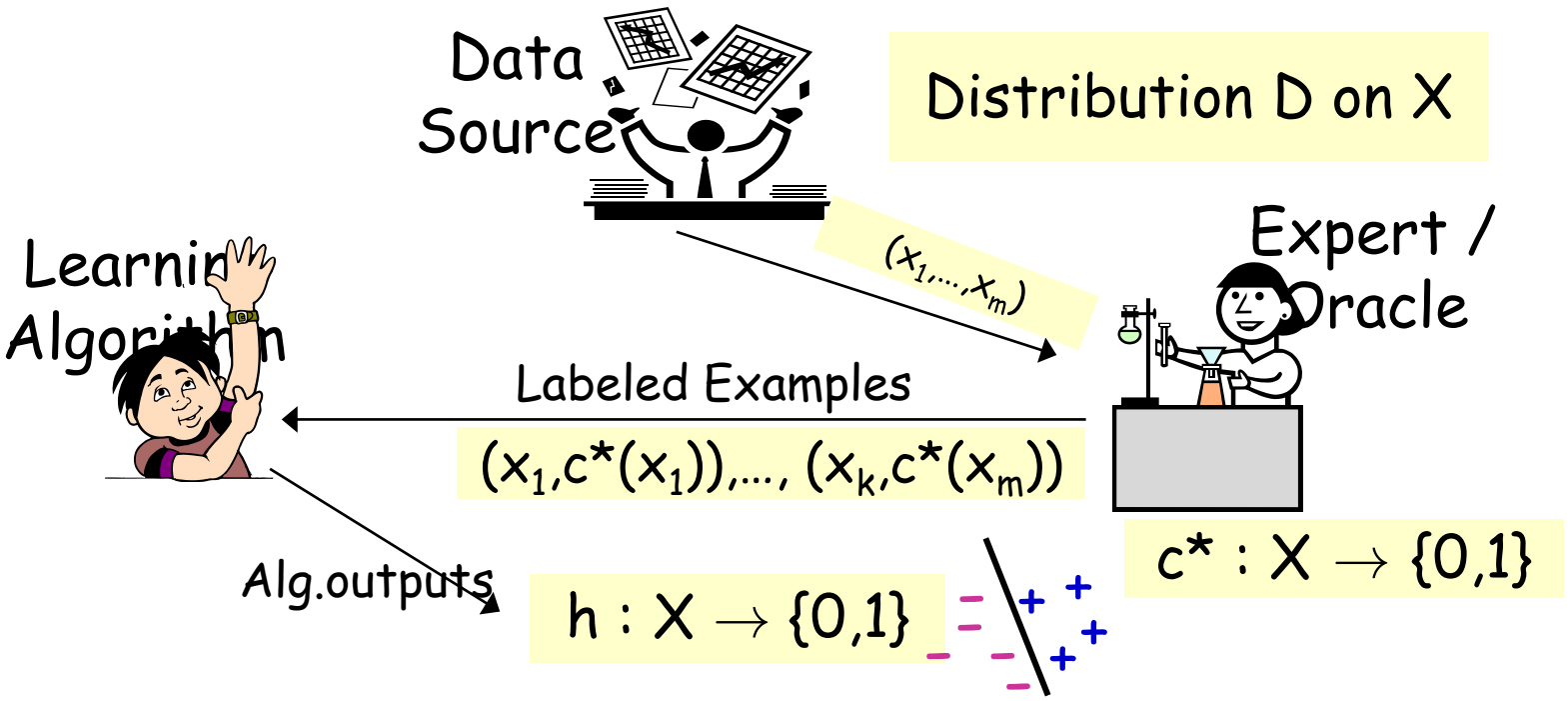
$c^* : X \rightarrow \{0,1\}$

Alg.outputs

$h : X \rightarrow \{0,1\}$



Statistical / PAC learning model



- Algo sees $(x_1,c^*(x_1)),\ldots,(x_k,c^*(x_m))$, x_i i.i.d. from D
- 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

Two Main Aspects in Classic Learning Theory

Algorithm Design. How to optimize?

Automatically generate rules that do well on observed data.

E.g., Boosting, SVM, etc.

Confidence Bounds, Generalization Guarantees

Confidence for rule effectiveness on future data.

Sample Complexity Results

Confidence Bounds, Generalization Guarantees

Confidence for rule effectiveness on future data.

Theorem

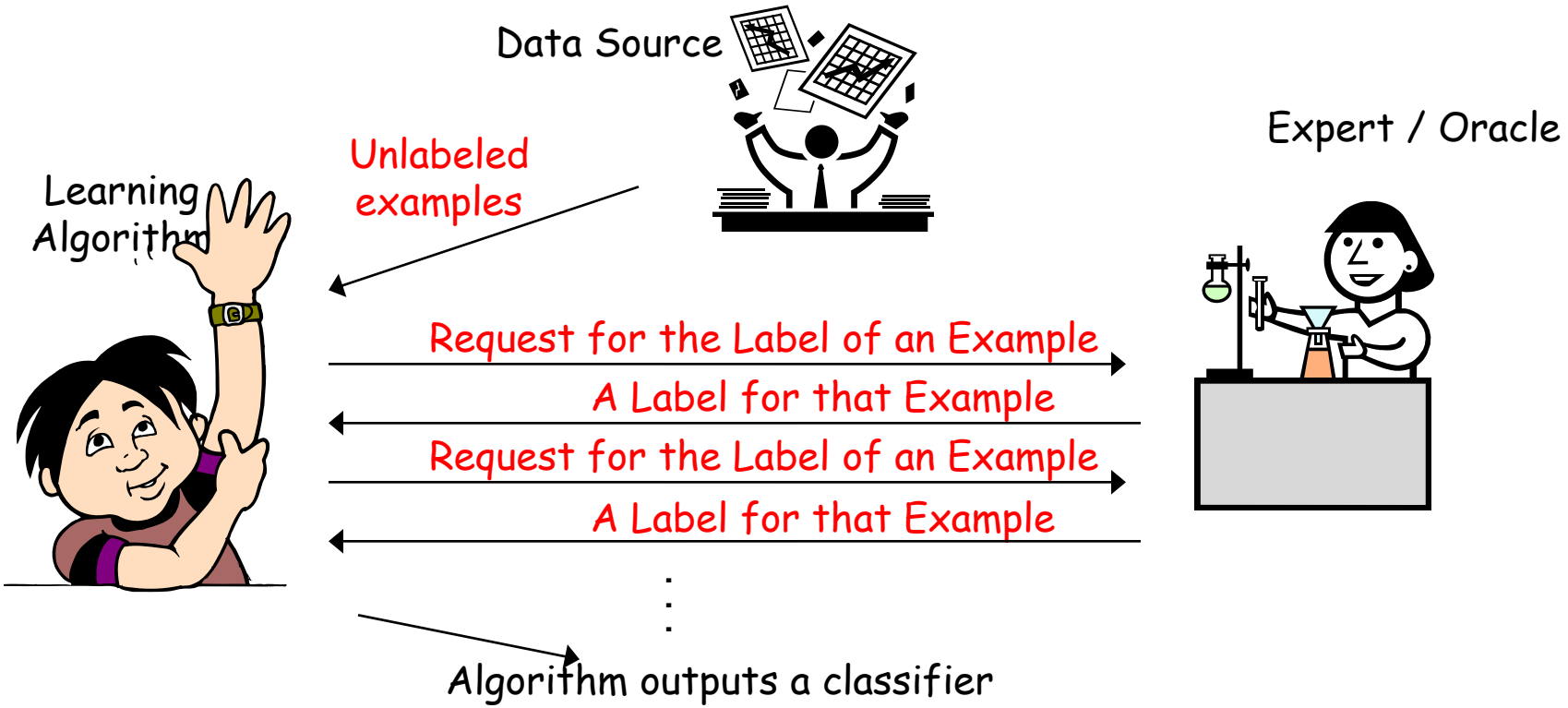
$$m \geq \frac{1}{\varepsilon}\left[VCdim(C)\log(\frac{1}{\varepsilon}) + \ln\left(\frac{1}{\delta}\right)\right]$$

labeled examples are sufficient s.t. with prob. at least $1-\delta$, all $h \in C$ with $\hat{err}(h) = 0$ have $err(h) \leq \varepsilon$.

- Agnostic – replace ε with ε^2.

Interactive Machine Learning

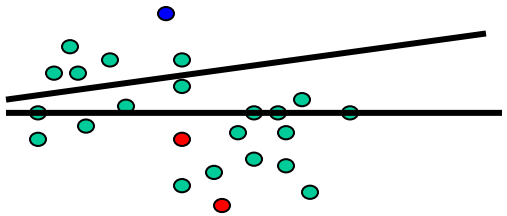
Active Learning



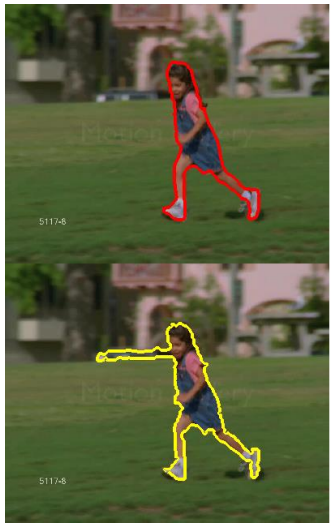
- Learner can choose specific examples to be labeled.
- Goal: use fewer labeled examples.
  - Need to pick informative examples to be labeled.

Active Learning in Practice

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

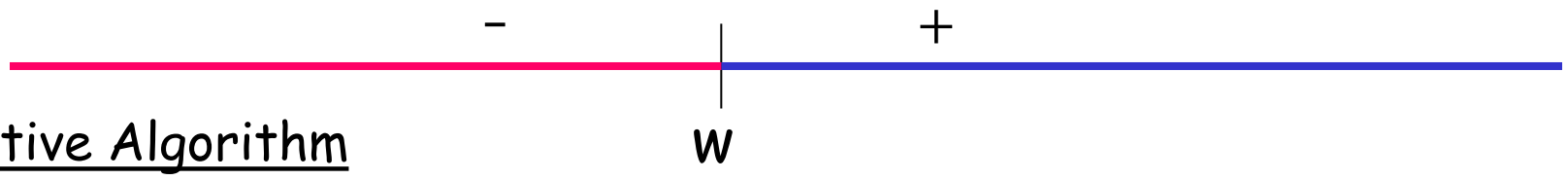


- Video Segmentation (Fathi-Balcan-Ren-Regh, BMVC 11).

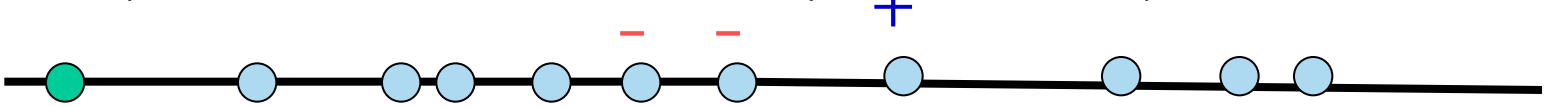


Provable Guarantees, Active Learning

- Canonical theoretical example [CAL92, Dasgupta04]



- Sample with $1/\varepsilon$ unlabeled examples; do binary search.



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

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

Disagreement Based Active Learning

"Disagreement based " algos: query points from current region of disagreement, throw out hypotheses when statistically confident they are suboptimal.

First analyzed in [Balcan, Beygelzimer, Langford'06] for A^2 algo.

Lots of subsequent work: [Hanneke07, DasguptaHsuMontleoni'07, Wang'09 , Fridman'09, Koltchinskii10, BHW'08, BeygelzimerHsuLangfordZhang'10, Hsu'10, Ailon'12, …]

Generic (any class), adversarial label noise.

Suboptimal in label complex & computationally prohibitive.

Poly Time, Noise Tolerant, Label Optimal AL Algos.

Margin Based Active Learning

Margin based algo for learning linear separators

- Realizable: exponential improvement, only $O(d \log 1/\varepsilon)$ labels to find w error ε when D logconcave. [Balcan-Long COLT'13]

- Resolves an open question on sample complex. of ERM.

- Agnostic & malicious noise: poly-time AL algo outputs w with $err(w) = O(\eta)$, η =err(best lin. sep). [Awasthi-Balcan-Long 2013]

  - First poly time AL algo in noisy scenarios!
  - First for malicious noise [Val85] (features corrupted too).

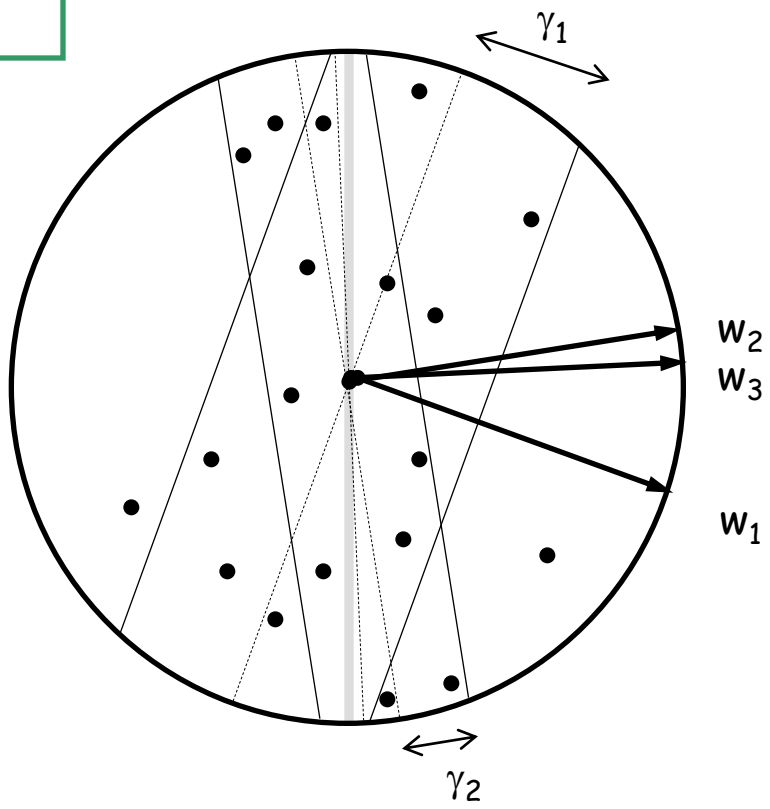
- Improves on noise tolerance of previous best passive [KKMS'05], [KLS'09] algos too!

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



Margin Based Active-Learning, Realizable Case

Theorem D log-concave in R^d.

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

Log-concave distributions: log of density fnc concave

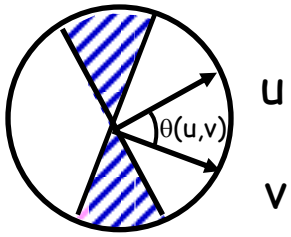
$$f(\lambda x_1 + (1 - \lambda x_2)) \geq f(x_1)^{\lambda} f(x_2)^{1-\lambda}$$

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

Linear Separators, Log-Concave Distributions

Fact 1

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

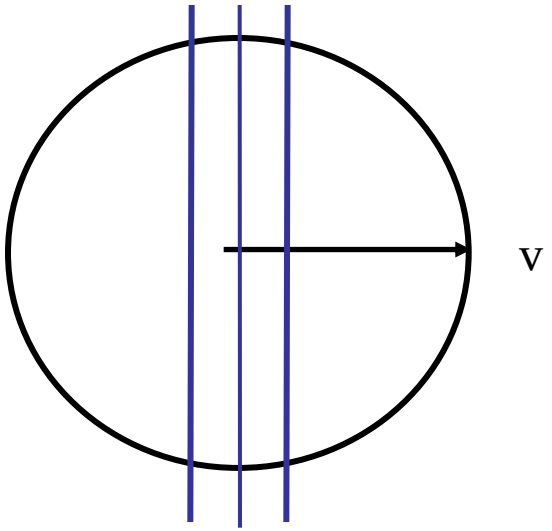


Proof idea:

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

Fact 2

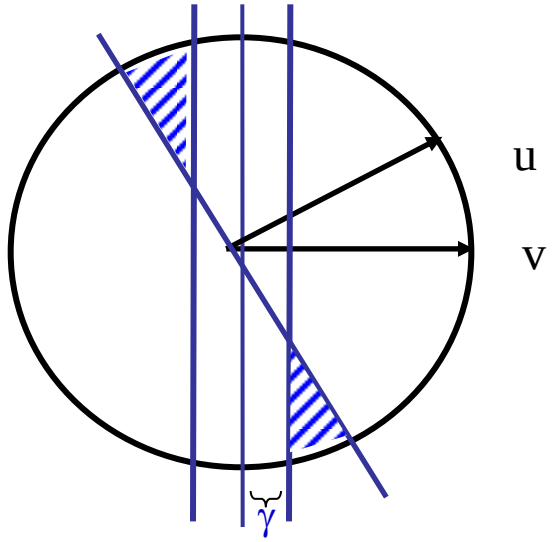
$$Pr_x[|v \cdot x| \leq \gamma] \leq \gamma.$$



Linear Separators, Log-Concave Distributions

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

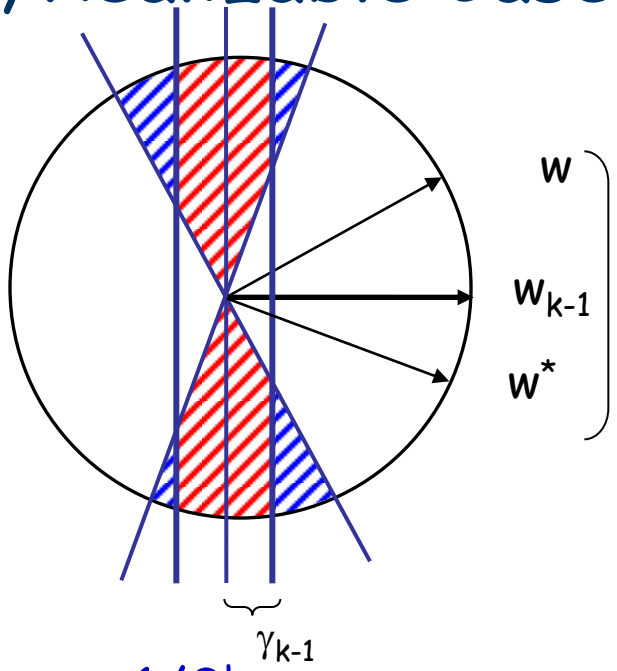
$$\Pr_x[(u \cdot x)(v \cdot x) < 0, |v \cdot x| \geq \gamma] \leq \frac{\beta}{4}.$$



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 $\leq 1/2^k$;

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

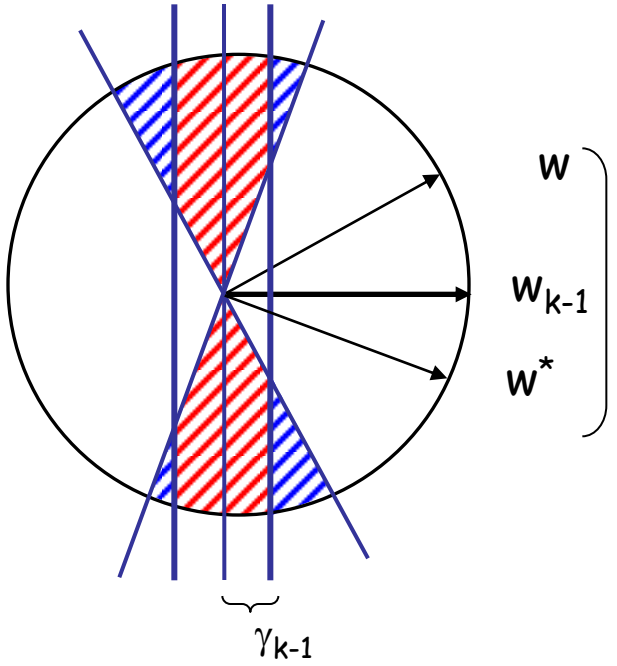
For $\gamma_k = O\left(\frac{c}{2^k}\right)$

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

Proof Idea

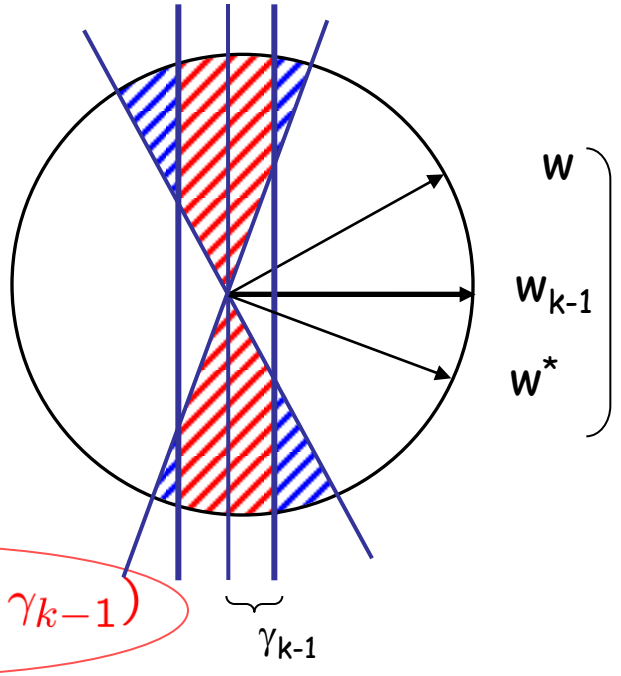
Under logconcave distr. for $\gamma_k = O\left(\frac{c}{2^k}\right)$

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



Proof Idea

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



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

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

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

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.

Margin Based Analysis [Balcan-Long, COLT13]

Theorem: (Active, Realizable)

D log-concave in R^d only $O(d \log 1/\varepsilon)$ labels to find w, err(w) $\leq \epsilon$.

Also leads to optimal bound for ERM passive learning

Theorem: (Passive, Realizable)

Any w consistent with $m = O\left(\frac{d}{\epsilon} + \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)\right)$

labeled examples satisfies err(w) $\leq \epsilon$, with prob. 1-δ.

- 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]

Our Results [Awasthi-Balcan-Long'2013]

Theorem: (Active, Agnostic or Malicious)

Poly(d, $1/\varepsilon$) time algo, uses only $O(d^2 \log 1/\varepsilon)$ labels to find w with $err(w) \leq \eta \log(1/\eta) + \varepsilon$ when D log-concave in R^d .

- First poly time algo for agnostic case.
- First analysis for malicious noise [Val85] [features corrupted too].

Theorem: (Active, Agnostic or Malicious)

Poly(d, $1/\varepsilon$) time algo, uses only $O(d^2 \log 1/\varepsilon)$ labels to find w with $err(w) \leq c\,(\eta + \varepsilon)$ when D uniform in R^d .

- Significantly improves over previously known results fo passive learning too! [KKMS'05, KLS'09]

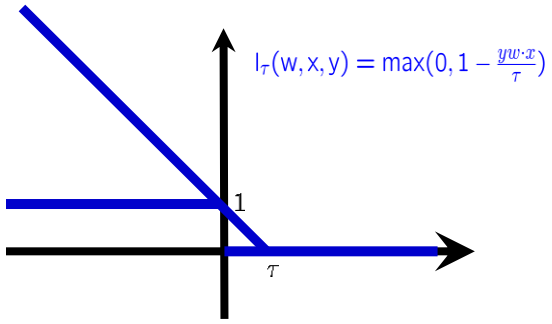
Margin Based Active-Learning, Agnostic Case

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



Margin Based Active-Learning, Agnostic Case

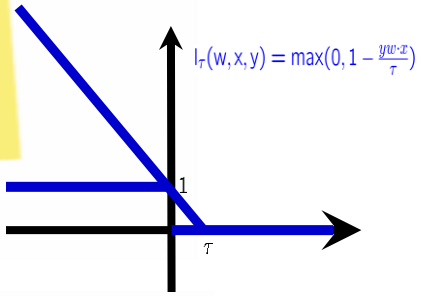
Draw m_1 unlabeled examples, label them, add them to W.

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

Localization in concept space.



Localization in instance space.

Analysis: the Agnostic Case

Theorem D log-concave in R^d. $\eta = \Omega(\epsilon/\log^2(1/\epsilon))$

If $\gamma_k = O\left(\frac{c}{2^k}\right)$, $\tau_k = O\left(\frac{c}{2^k}\right)$, $r_k = O\left(\frac{c}{2^k}\right)$, $s = \log\left(\frac{1}{\epsilon}\right)$, **err($w_s$)$\leq \varepsilon$.**

Key ideas:

- As before need $\Pr(w \text{ errs on } x \mid\mid |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq C$
- For w in $B(w_{k-1}, r_{k-1})$ we have $l(w,$ Infl. Noisy points

$$\text{err}(w_k) \leq E[l(w_k, x, y)] \leq \tau_k/\gamma_k + \eta 2^k \sqrt{d} \leq C$$

- sufficient to set $\eta \leq \epsilon/\sqrt{d}$ Hinge loss over clean examples

- Careful variance analysis leads $\eta = \Omega(\epsilon/\log^2(1/\epsilon))$

Analysis: Malicious Noise

Theorem D log-concave in R^d. $\eta = \Omega(\epsilon/\log^2(1/\epsilon))$

If $\gamma_k = O\left(\frac{c}{2^k}\right)$, $\tau_k = O\left(\frac{c}{2^k}\right)$, $r_k = O\left(\frac{c}{2^k}\right)$, $s = \log\left(\frac{1}{\epsilon}\right)$, **err($w_s$)$\leq \varepsilon$.**

The adversary can corrupt both the label and the feature part.

Key ideas:

- As before need $\Pr(w \text{ errs on } x \mid |w_{k-1} \cdot x| \leq \gamma_{k-1}) \leq C$
- Soft localized outlier removal and careful variance analysis.

Improves over Passive Learning too!

Passive Learning	**Prior Work**	**Our Work**
Malicious	$\eta = \Omega\left(\frac{\epsilon^3}{\log^2(d/\epsilon)}\right)$ [KLS'09]	$\eta = \Omega(\epsilon/\log^2(1/\epsilon))$
Agnostic	$\eta = \Omega\left(\frac{\epsilon^3}{\log(1/\epsilon)}\right)$ [KLS'09]	$\eta = \Omega(\epsilon/\log^2(1/\epsilon))$
Active Learning [agnostic/malicious]	NA	$\eta = \Omega(\epsilon/\log^2(1/\epsilon))$

Improves over Passive Learning too!

Passive Learning	**Prior Work**	**Our Work**
Malicious	$\eta = \Omega\left(\frac{\epsilon}{d^{1/4}}\right)$ [KKMS'05] $\eta = \Omega\left(\frac{\epsilon^2}{\log(d/\epsilon)}\right)$ [KLS'09]	$\eta = \Omega(\epsilon)$ Info theoretic optimal
Agnostic	$\eta = \Omega\left(\frac{\epsilon}{\sqrt{\log(1/\epsilon)}}\right)$ [KKMS'05]	$\eta = \Omega(\epsilon)$
Active Learning [agnostic/malicious]	NA	$\eta = \Omega(\epsilon)$ Info theoretic optimal

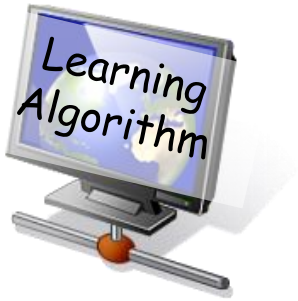
Slightly better results for the uniform distribution case.

Localization both algorithmic and analysis tool!

Useful for active and passive learning!

Important direction: richer interactions with the expert.

Better Accuracy

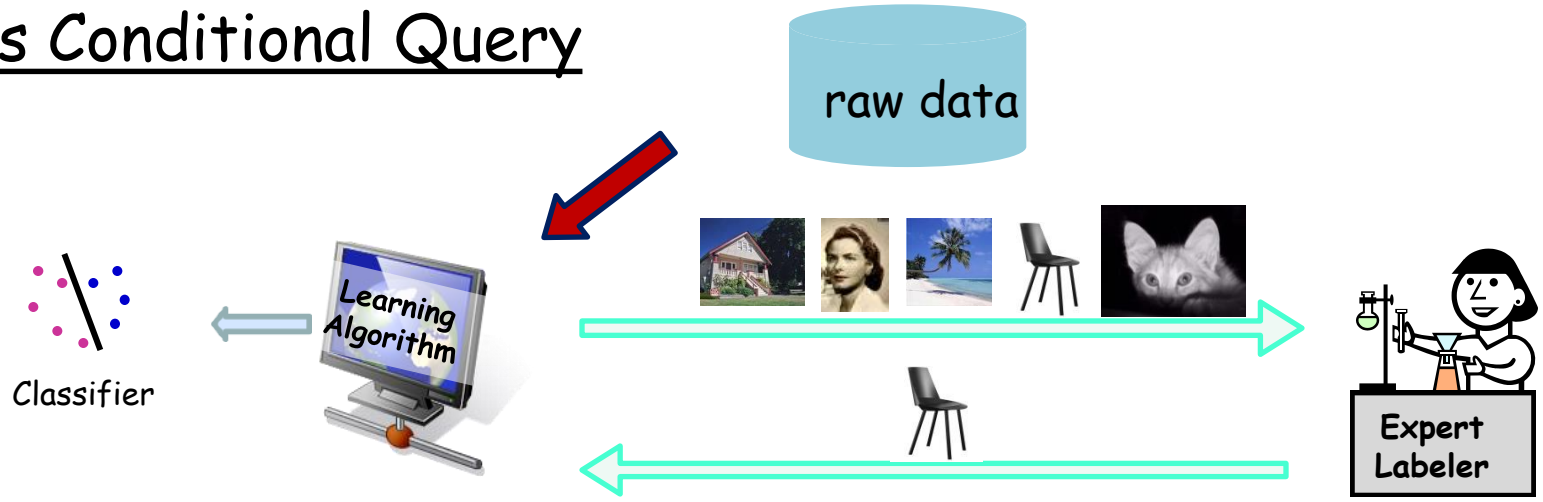


Natural interaction

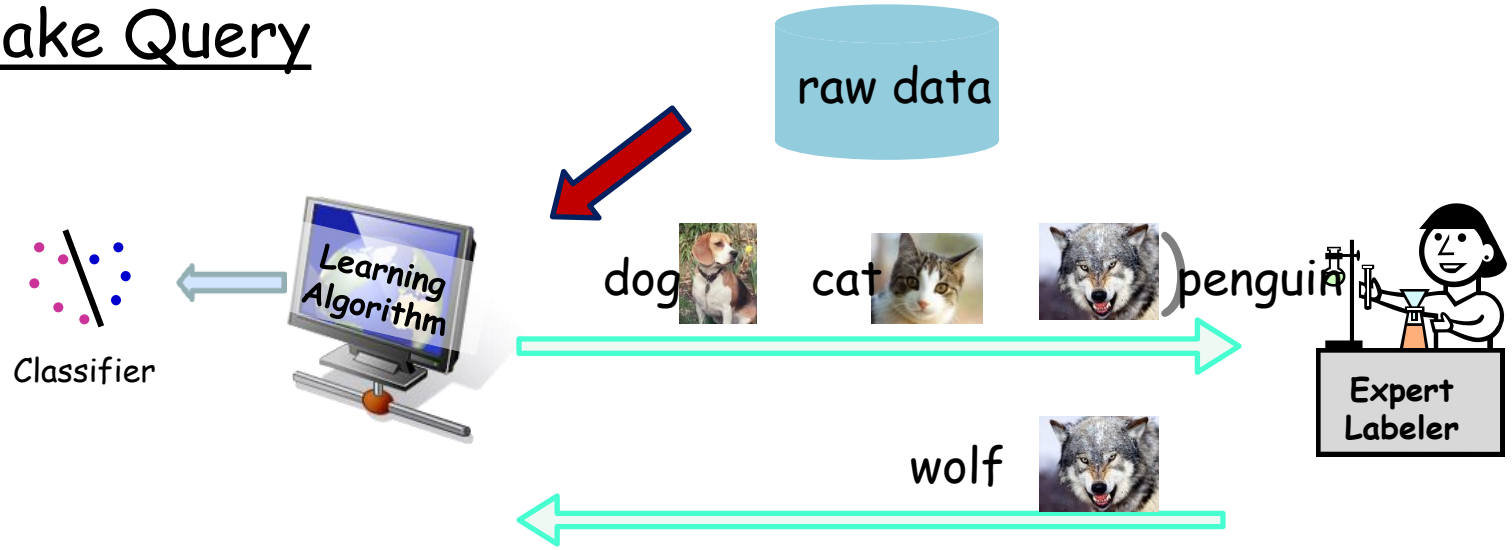
Fewer queries

New Types of Interaction [Balcan-Hanneke COLT'12]

Class Conditional Query

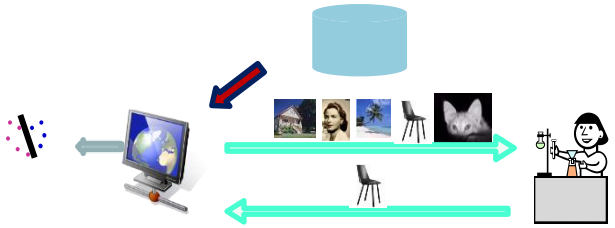


Mistake Query



Class Conditional & Mistake Queries

- Used in practice, e.g. Faces in IPhoto.
- Lack of theoretical understanding.



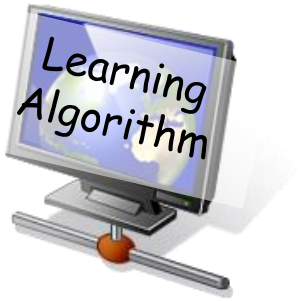
- Realizable (Folklore): much fewer queries than label requests.

Balcan-Hanneke, COLT'12

Tight bounds on the number of CCQs to learn in the presence of noise (agnostic and bounded noise)

Important direction: richer interactions with the expert.

Better Accuracy



Natural interaction

Fewer queries

Distributed Learning

Distributed Learning

Data distributed across multiple locations.

E.g., medical data

Distributed Learning

- Data distributed across multiple locations.
- Each has a piece of the overall data pie.
- To learn over the combined D, must communicate.
- Communication is expensive.

President Obama cites Communication-Avoiding Algorithms in FY 2012 Department of Energy Budget Request to Congress

Important question: how much communication?

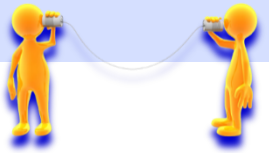
Plus, privacy & incentives.

Distributed PAC learning

[Balcan-Blum-Fine-Mansour, COLT 2012]
Runner UP Best Paper

- X – instance space. k players.
- Player i can sample from D_i, samples labeled by c*.
- Goal: find h that approximates c* w.r.t. D=1/k (D_1 + … + D_k)

Goal: learn good h over D, as little communication as possible



Main Results

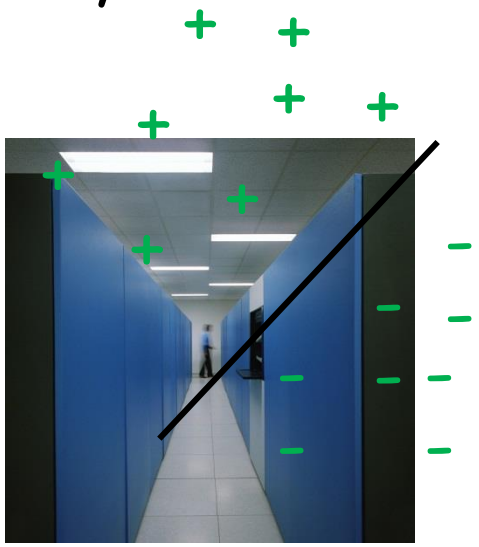
- Generic bounds on communication.
- Broadly applicable communication efficient distr. boosting.
- Tight results for interesting cases [intersection closed, parity fns, linear separators over "nice" distrib].
- Privacy guarantees.

Interesting special case to think about

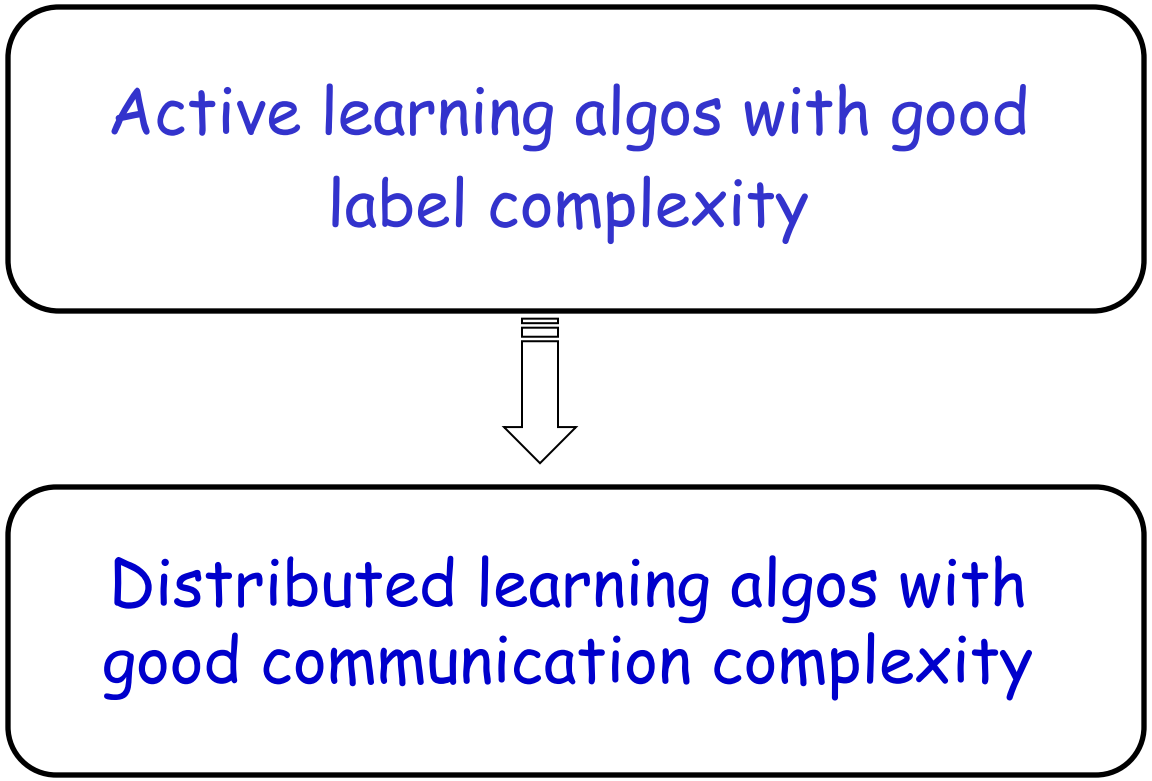
k=2. One has the positives and one has the negatives.

- How much communication, e.g., for linear separators?

Player 1



Player 2



So, if linear sep., log-concave distr. only d log(1/ϵ) examples communicated.

Generic Results

Baseline $d/\epsilon \log(1/\epsilon)$ examples, 1 round of communication

- Each player sends $d/(\epsilon k) \log(1/\epsilon)$ examples to player 1.
- Player 1 finds consistent $h \in C$, whp error $\leq \epsilon$ wrt D

Distributed Boosting

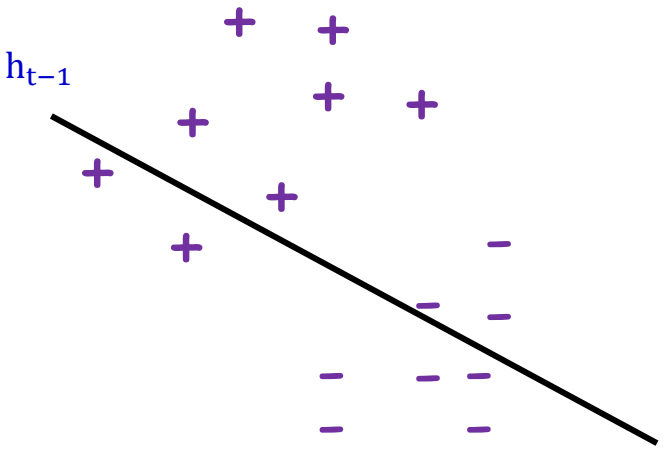
Only $O(d \log 1/\epsilon)$ examples of communication

Key Properties of Adaboost

Input: $S=\{(x_1, y_1), \ldots, (x_m, y_m)\}$

- For $t=1,2, \ldots, T$
  - Construct D_t on $\{x_1, \ldots, x_m\}$
  - Run weak algo A on D_t , get h_t

Output $H\_final=sgn(\sum \alpha_t \ h_t)$



- D_1 uniform on $\{x_1, \ldots, x_m\}$
- D_{t+1} increases weight on x_i if h_t incorrect on x_i ; decreases it on x_i if h_t correct.

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{\{-\alpha_t\}} \text{ if } y_i = h_t(x_i)$$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} e^{\{\alpha_t\}} \text{ if } y_i \neq h_t(x_i)$$

Key points:

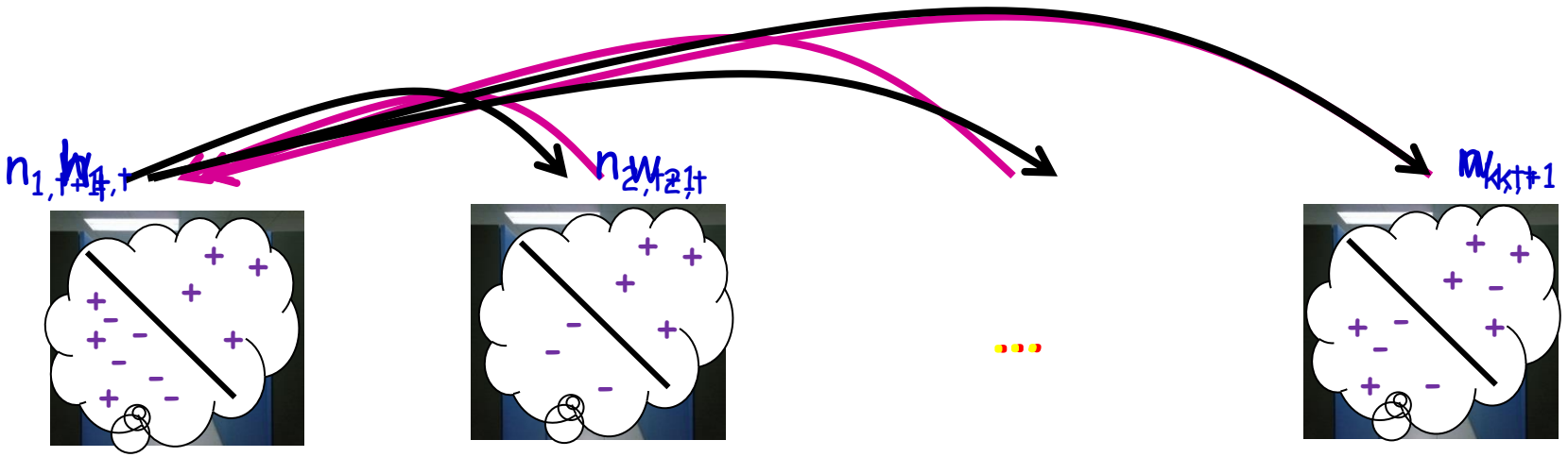
- $D_{t+1}(x_i)$ depends on $h_1(x_i), \ldots, h_t(x_i)$ and normalization factor that can be communicated efficiently.
- To achieve weak learning it suffices to use $O(d)$ examples.

Distributed Adaboost

Each player i has a sample S_i from D_i.

For $t=1,2,\ldots,T$

- Each player sends player 1 data to produce weak h_t.
  [For $t=1$, $O(d/k)$ examples each.]
- Player 1 broadcasts h_t to others.
- Player i reweights its own distribution on S_i using h_t and sends the sum of its weights $w_{i,t}$ to player 1.
- Player 1 determines # of samples to request from each i
  [samples $O(d)$ times from the multinomial given by $w_{i,t}/W_t$ to get $n_{i,t+1}$].



Communication: fundamental resource in DL

Theorem Learn any class C with $O(\log(1/\epsilon))$ rounds using $O(d)$ examples + 1 hypothesis per round.

- Key: in Adaboost, $O(\log 1/\epsilon)$ rounds to achieve error ϵ.

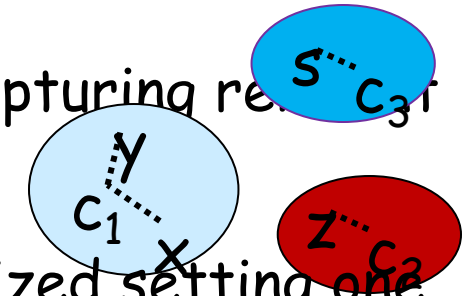
Theorem In the agnostic case, can learn to error $O(OPT)+\epsilon$ using only $O(k \log|C| \log(1/\epsilon))$ examples.

- Key: distributed implementation of Robust Halving developed for learning with mistake queries [Balcan-Hanneke'12].

Distributed Clustering [Balcan-Ehrlich-Liang, NIPS 2013]

k-median: find center pts $c_1, c_2, \ldots, c_r$ to minimize $\sum_x \min_i d(x,c_i)$

k-means: find center pts $c_1, c_2, \ldots, c_r$ to minimize $\sum_x \min_i d^2(x,c_i)$

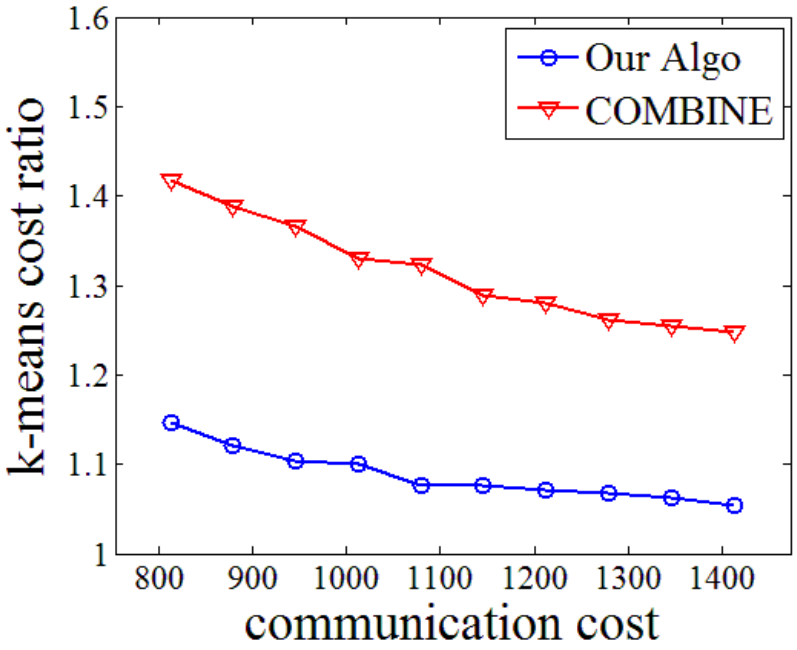


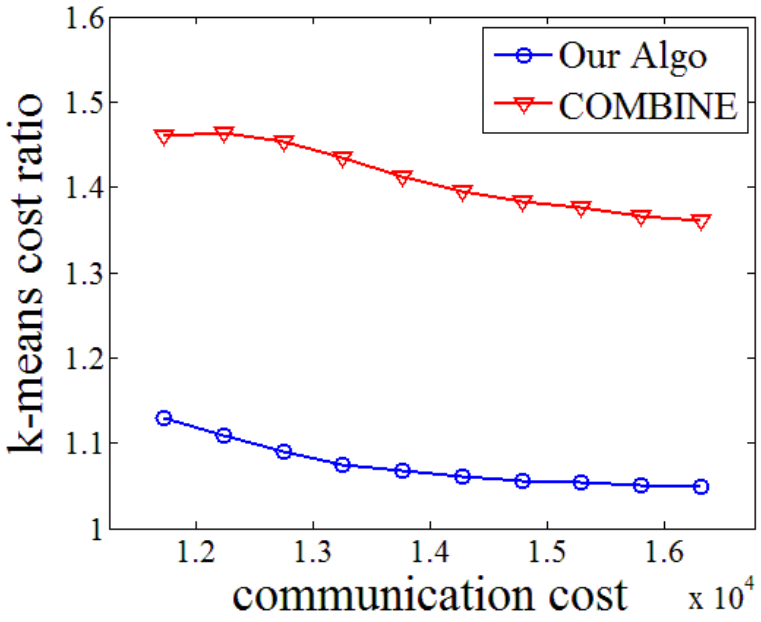
- Key idea: use coresets, short summaries capturing re[illegible] info w.r.t. all clusterings.
- [Feldman-Langberg STOC'11] show that in centralized setting one can construct a coreset of size $\tilde{O}(rd/\epsilon^2)$
- By combining local coresets, we get a global coreset – the size goes up multiplicatively by #sites.
- In [Balcan-Ehrlich-Liang, NIPS 2013] show a 2 round procedure with communication only $\tilde{O}(rd/\epsilon^2)$

  [As opposed to $\tilde{O}(rd/\epsilon^2)$#sites]

Distributed Clustering [Balcan-Ehrlich-Liang, NIPS 2013]

k-means: find center pts $c_1, c_2, \ldots, c_k$ to minimize $\sum_x \min_i d^2(x,c_i)$





Discussion

- Communication as a fundamental resource.

Open Questions

- Other learning or optimization tasks.
- Refined trade-offs between communication complexity, computational complexity, and sample complexity.
- Analyze such issues in the context of transfer learning of large collections of multiple related tasks (e.g., NELL).