Active Statistical Query Learning

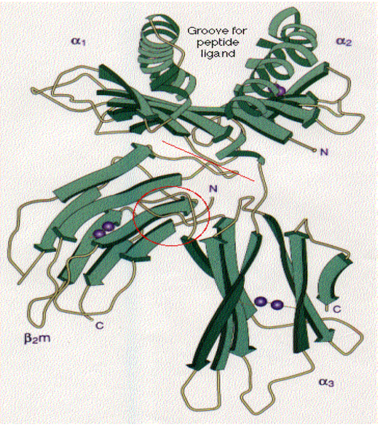
Maria-Florina Balcan

Vitaly Feldman, IBM Research

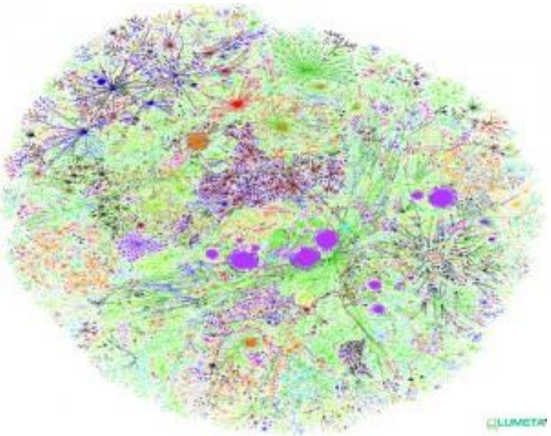
2-Minute Version

Modern applications: massive amounts of raw data.

Only a tiny fraction can be annotated by human experts.



Protein sequences



Billions of webpages

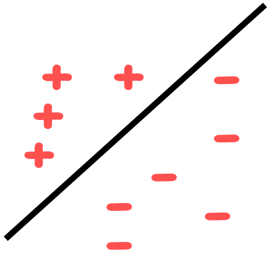
Images

Active learning: leverage available data, minimize need for expert intervention.

2-Minute Version

Model for designing Statistical Active Learning (AL) algos

- Poly time statistical AL algos → poly time algos tolerant to random classification noise.
  - thresholds, rectangles, lin. separators.

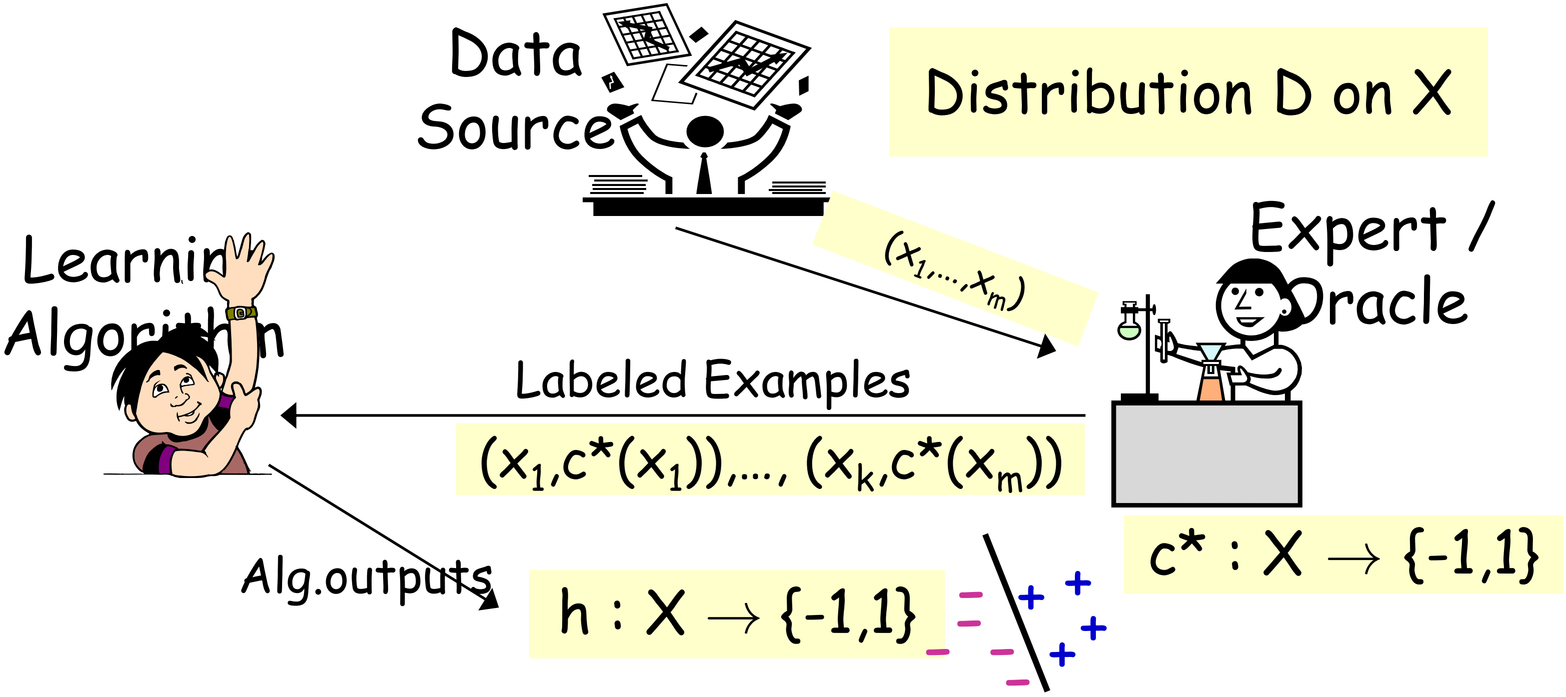


- Naturally lead to differentially private AL algorithms.

Outline of the talk

- Passive Learning. Statistical Query Learning
- Active Learning
- Active Statistical Query Learning

Statistical / PAC learning model



Data Source

Distribution D on X

$(x_1,\ldots,x_m)$

Expert / Oracle

Learning Algorithm

Labeled Examples

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

$c^* : X \to \{-1,1\}$

Alg.outputs

$h : X \to \{-1,1\}$

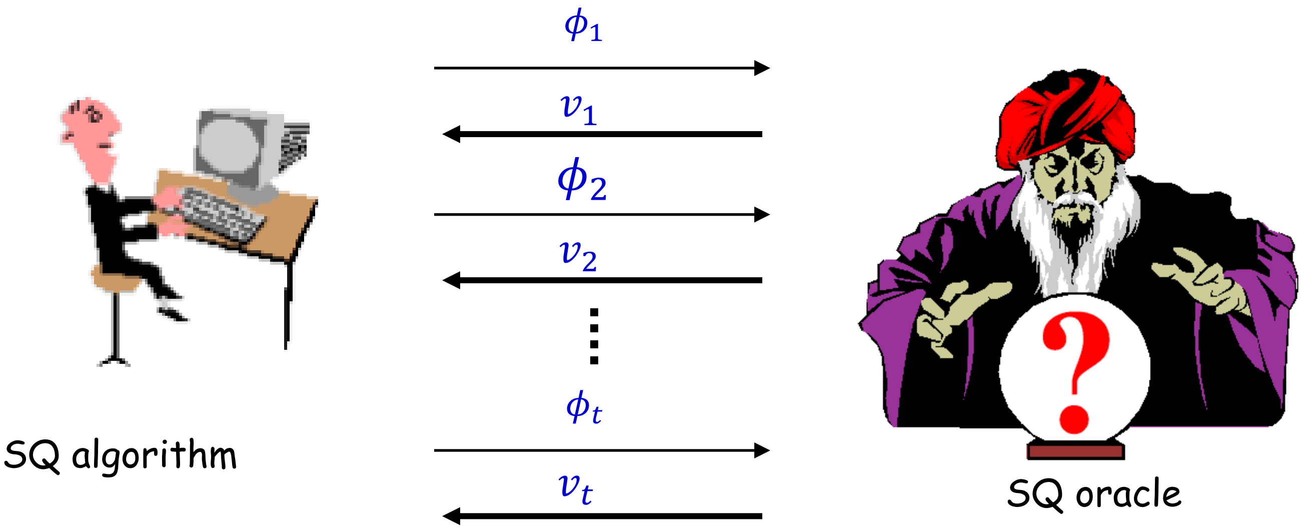
- Algo sees $(x_1,c^*(x_1)),\ldots,(x_k,c^*(x_m))$, x_i i.i.d. from D; $c^* \in C$ [induce P]
- Do optimization over S, find hypothesis $h \in C$.
- Goal: h has small error over D.

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

- PAC model: poly time algo.

Statistical Query (SQ) Model [Kearns 93]

- Only statistical properties (not individual examples).
- Algo asks: "what is prob. a (labeled) example has property ϕ? Pls. tell me up to additive error τ."



ϕ_1

v_1

ϕ_2

v_2

$\vdots$

ϕ_t

v_t

SQ algorithm

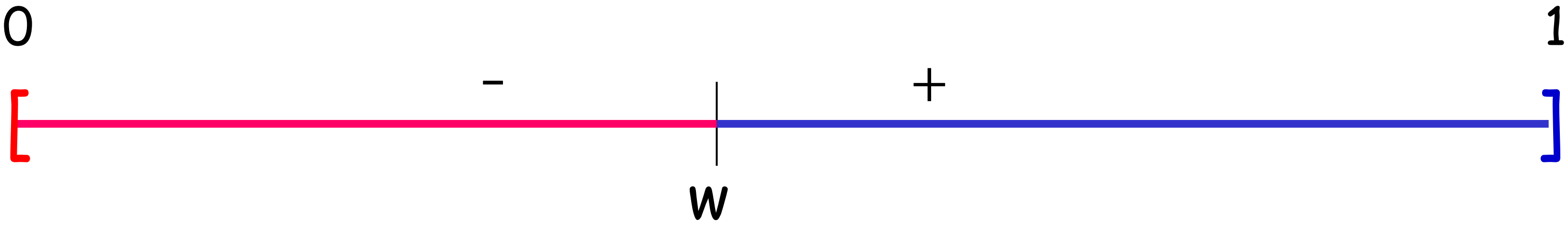
SQ oracle

$$|v_i - \boldsymbol{E}_D[\phi_i(x, c^*(x))]| \le \tau_i$$

τ_i is tolerance of the query

Must output h of error $\le \varepsilon$.

Simple Example: Threshold Fns



If D uniform

- Ask p = Pr(x is positive) up to tolerance ϵ.

  [use query $\phi(x,l) = \frac{1+l}{2}; \quad \tau = \epsilon$]

- Output 1-p.

In general,

- Ask p = Pr(x is positive) up to tolerance $\epsilon/2$.
- Ask unlabeled SQs (binary search) to find z s.t.

  $\Pr(x \in (z,1]) \in [p - \epsilon/2, p + \epsilon/2]$.

- Output z.

Properties of SQ model

- Can simulate SQ algos from random examples.

  [Result of query ϕ, τ whp $1-\delta$ from empirical expectation of $O\left(\frac{1}{\tau^2}\log\left(\frac{1}{\delta}\right)\right)$ random examples.]

- Can automatically convert to work in presence of random classification noise!

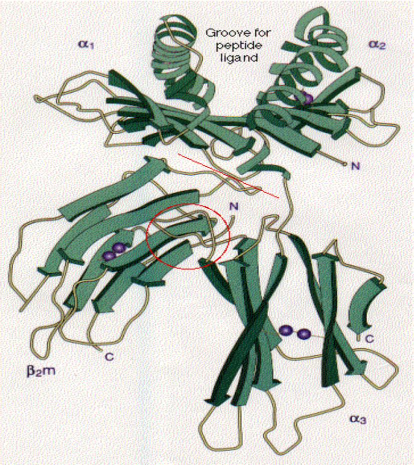
- Many ML algorithms have SQ analogues.
  - E.g, Perceptron, BFKV'96,DV'06 for linear separators.

- Can be made differentially private [BDMN'05]!

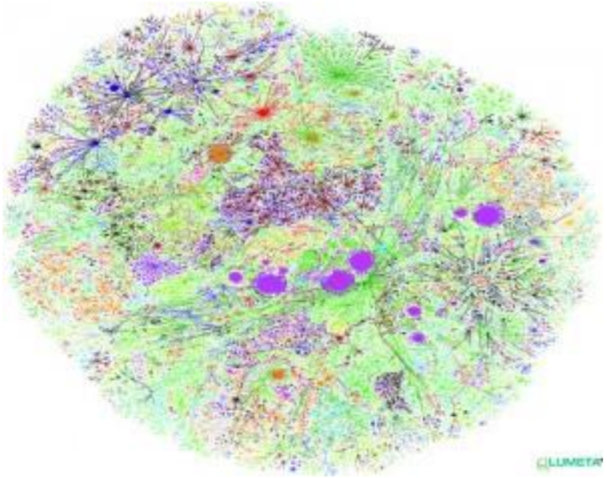
Classic Paradigm Insufficient Nowadays

Modern applications: massive amounts of raw data.

Only a tiny fraction can be annotated by human experts.



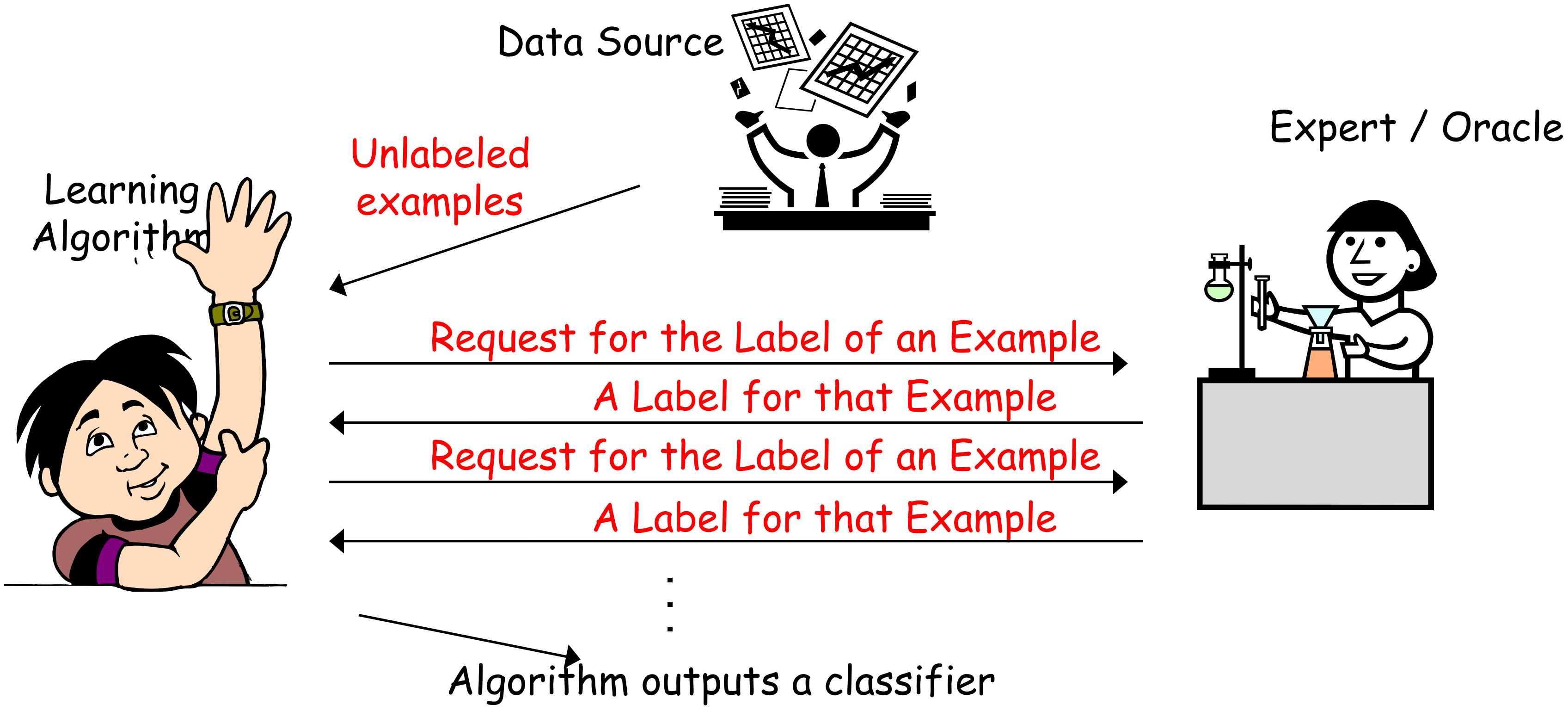
Protein sequences



Billions of webpages

Images

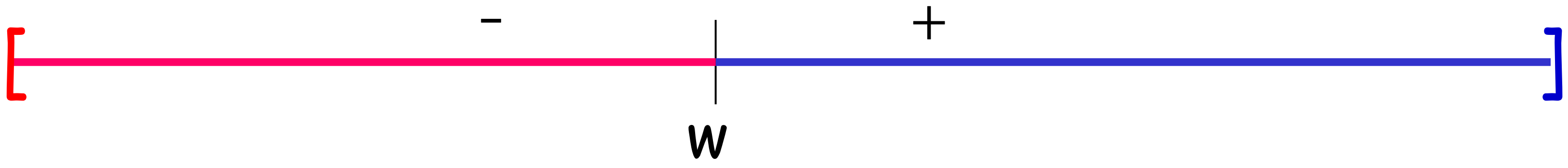
Active Learning: Major Area in Modern ML



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

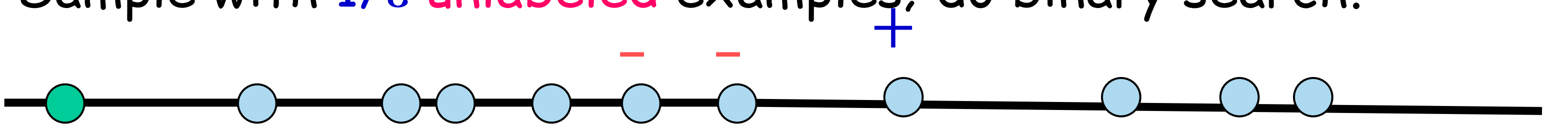
Provable Guarantees, Active Learning

- Canonical theoretical example [CAL92, Dasgupta04]



Active Algorithm

- Sample with 1/ε unlabeled examples; do binary search.



Passive supervised: Ω(1/ε) labels to find an ε-accurate threshold.

Active: only O(log 1/ε) labels. Exponential improvement.

Lots of exciting activity in recent years

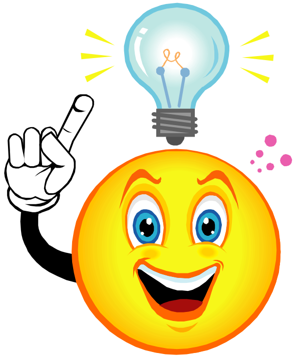
- Very general "disagreement based" algos [quer pts from region of disagreement, throw out hyp. when statistically con ey are suboptimal]
  - First analyzed in [Balcan, Beyg ford'06].
  - [Hanneke07, Dasgu oni'07, Wang'09, Fridman'09, Koltchinskii1 eygelzimerHsuLangfordZhang'10, Hsu'1 ...]

Computationally inefficient

- Algos for specific (noise free) case .g., linear separators.
  - QBC [Freund et al., '97]
  - Active Pe upta, Kalai, Monteleoni'05]
  - Margin Bas AL [Balcan BroderZhang'07] [BalcanLong'13]

Very specific

Open: poly time, noise tolerant AL algos.



This work: framework for designing poly time AL algos tolerant to random classification noise that satisfy DP naturally.

Active Statistical Query Model

Instead of access to random examples, algo only gets active estimates of statistical properties.

Query (χ, ϕ), $\chi: X \rightarrow [0,1]$ filter [prob. of querying label of x]

"What is prob. a labeled example from $P_{|\chi}$ has property ϕ?.

Pls. tell me up to additive error τ if $\mathrm{E_D}[\chi(x)] \geq \tau_f$ "

If $\mathrm{E_D}[\chi(x)] \geq \tau_f$ then $\left|v - \boldsymbol{E}_{P_{|\chi}}[\phi(x, f(x))]\right| \leq \tau$

τ_f filter tolerance; τ tolerance of query (χ, ϕ)

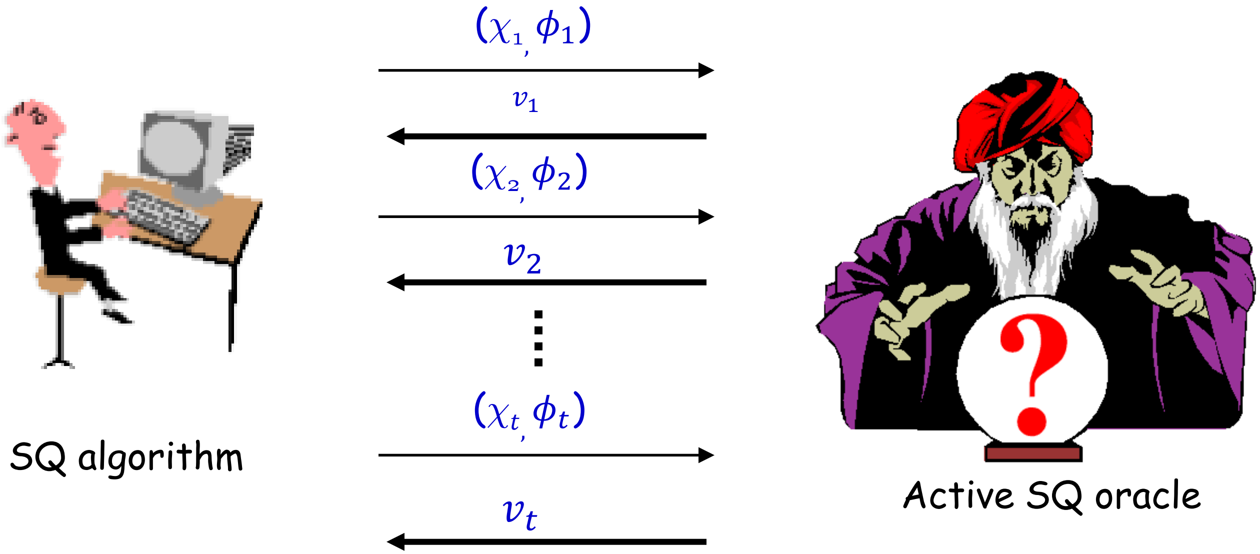
Algo gets an estimate of the prob that ϕ satisfied cond. on x satisfying χ.

Active Statistical Query Model

Query (χ, ϕ), $\chi: X \to [0,1]$ filter [prob. of querying label of x]

“What is the prob. a labeled example from $P_{|\chi}$ has property ϕ?.

Pls. tell me up to additive error τ if $\mathrm{E}_{\mathrm{D}}[\chi(x)] \geq \tau_f$ ”



(χ_1, ϕ_1)

v_1

(χ_2, ϕ_2)

v_2

⋮

(χ_t, ϕ_t)

v_t

Active SQ algorithm

Active SQ oracle

Simulating Active Statistical Queries

Query (χ, ϕ) , τ_f , τ : if $\mathrm{E_D}[\chi(x)] \geq \tau_f$ then $\left|v - \boldsymbol{E}_{p_{|\chi}}[\phi(x, f(x))]\right| \leq \tau$

Fact Can be simulated with $\frac{1}{\tau^2} \log\left(\frac{1}{\delta}\right)$ labeled examples and $\frac{1}{\tau_f}\frac{1}{\tau^2} \log\left(\frac{1}{\delta}\right)$ unlabeled samples.

Design algo with τ large [only τ_f small], much less labeled data.

<u>Notes</u>:

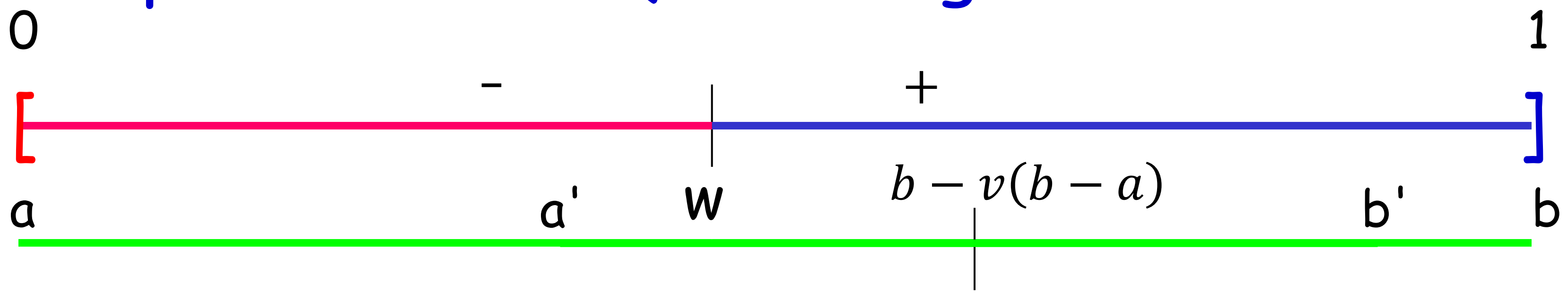
1. Generalizes SQ model ($\chi = 1$, $\tau_f = 1$).

2. Since $E_{P_{|\chi}}[\phi(x, l)] = \frac{E_P[\phi(x,l)\ \chi(x)]}{E_P[\phi(x,l)]}$, can use 2 passive SQs.

   Need to estimate $E_P[\phi(x, l)\ \chi(x)]$ within $\tau E_P[\phi(x, l)]$

   Too much labeled data.

Example: Active SQ Learning of Thresholds



Passive SQ: Ask query $\phi(x,l) = \frac{1+l}{2}$ with tolerance ϵ ; so $1/\epsilon^2$ labels

Active SQ: Key : localize/filter and use only constant tolerance

Assume w ∈ [a,b]

Ask query $\phi(x,l) = \frac{1+l}{2}$; $\chi(x) = I_{x \in [a,b]}$; $\tau = \frac{1}{4}, \tau_f = b - a$; get v.

Know $|v - E[\phi(x,l)|\, x \ \in \ [a,b]| \le \frac{1}{4}$; $|E[\phi(x,l)|\, x \ \in \ [a,b]| = \frac{b-w}{b-a}$

So $\mathrm{w} \in \left[b - \left(v + \frac{1}{4}\right)(b-a), b - \left(v - \frac{1}{4}\right)(b-a)\right]$ twice smaller than [a,b]

Only $\log\left(\frac{1}{\epsilon}\right)$ rounds, and $\log\left(\frac{1}{\epsilon}\right) \log\left(\frac{\log\left(\frac{1}{\epsilon}\right)}{\delta}\right)$ labeled examples

Noise Tolerance

Fact Query (χ, ϕ) , τ_f , τ

Under RCN given access to P^η estimate $E_{P_{|\chi}}[\phi(x,l)]$ within τ using $\frac{1}{\tau^2}\frac{1}{(1-2\eta)^2}\log\left(\frac{1}{\delta}\right)$ labeles and $\frac{1}{\tau_f}\frac{1}{\tau^2}\frac{1}{(1-2\eta)^2}$ $\log\left(\frac{1}{\delta}\right)$ unlabeled examples.

[Active SQs can be simulated from examples corrupted with RCN noise.]

Key points: Break into part affected by noise, and part unaffected; estimate each within $\frac{\tau}{2}$

$$\phi(x,l) = \frac{\phi(x,1) - \phi(x,-1)}{2}\ l + \frac{\phi(x,1) + \phi(x,-1)}{2}$$

$$E_{P^\eta_{|\chi}}\left[\frac{\phi(x,1) - \phi(x,-1)}{2}\ l\right] = (1-2\eta)E_{P_{|\chi}}\left[\frac{\phi(x,1) - \phi(x,-1)}{2}\ l\right]$$

sufficient to estimate it within $(1-2\eta)\tau/2$

Active SQ Learning of Linear Separators

Run a passive SQ algo to get w_0 with $\mathrm{err}(w_0) < C$.

iterate $k = 2, \ldots, s$

- let μ_k be the indicator fnc of being within $\gamma_{k\text{-}1}$ of w_{k-1}.
- Let $\chi_k = \frac{\sum_{\{i \le k\}} \mu_i}{k}$
- Run passive SQ over $\mathrm{D}_{|\chi_k}$ to output w_k of error $\frac{c}{k}$ over $\mathrm{D}_{|\chi_k}$.

[passive SQs over $\mathrm{D}_{|\chi_k}$ implemented as active SQs with $\tau_f = \mathrm{C}\epsilon$]

Theorem D log-concave in R^d.

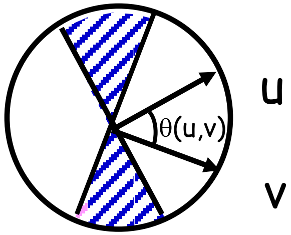
If $\gamma_k = O\left(\frac{c}{2^k}\right)$ then after $s = \log\left(\frac{1}{\epsilon}\right)$ iterations $\mathrm{err}(w_s) \le \epsilon$

Total number of labeled examples is $\mathrm{poly}\left(d, \log\left(\frac{1}{\epsilon}\right)\right)$

Linear Separators, Log-Concave Distributions

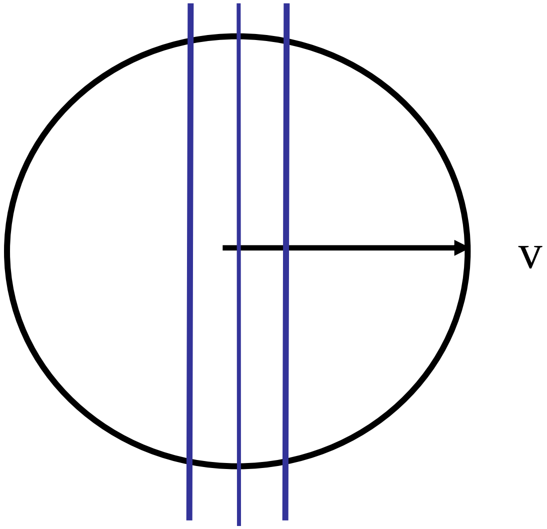
Fact 1

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



Fact 2

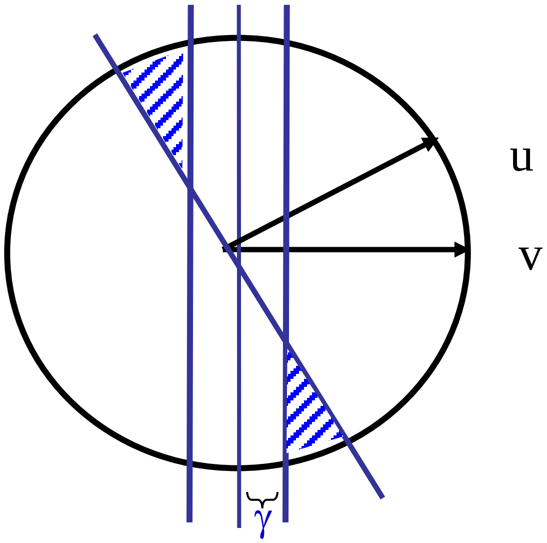
$$\Pr_X\left[|v \cdot x| \leq \gamma\right] \leq \gamma.$$



Linear Separators, Log-Concave Distributions

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

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

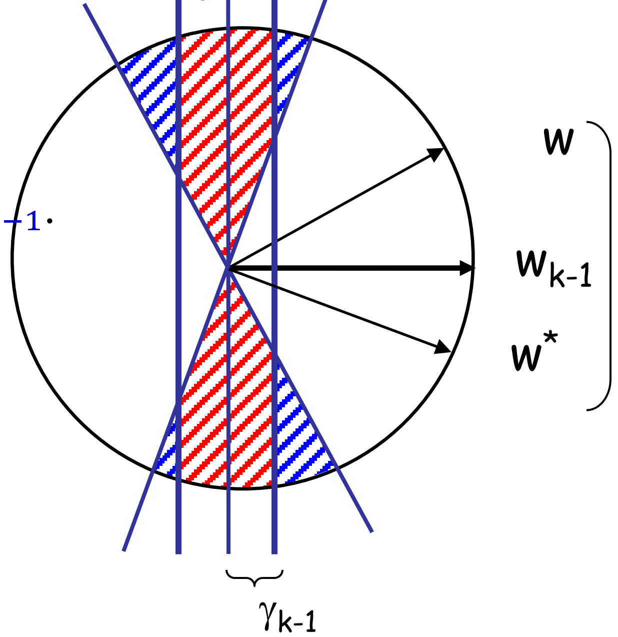


Margin Based Active-Learning, Realizable Case

Run a passive SQ algo to get w_0 with $\mathrm{err}(w_0)<C$.

iterate $k = 2, \ldots, s$

- let μ_k indicator fnc of being within margin γ_{k-1} of w_{k-1}.
- Let $\chi_k = \frac{\sum_{\{i \le k\}} \mu_i}{k}$
- Run passive SQ over $D_{|\chi_k}$ with error $\frac{c}{k}$ and filter tolerance $C\epsilon$ to obtain w_k



Proof Idea

Induction: all w s.t. $\mathrm{err}_{D_{|\mu_i}}(w) \le C$, $i \le k$, $\mathrm{err}_D(w) \le \frac{1}{2^k}$. So, $\mathrm{err}_D(w_k) \le \frac{1}{2^k}$

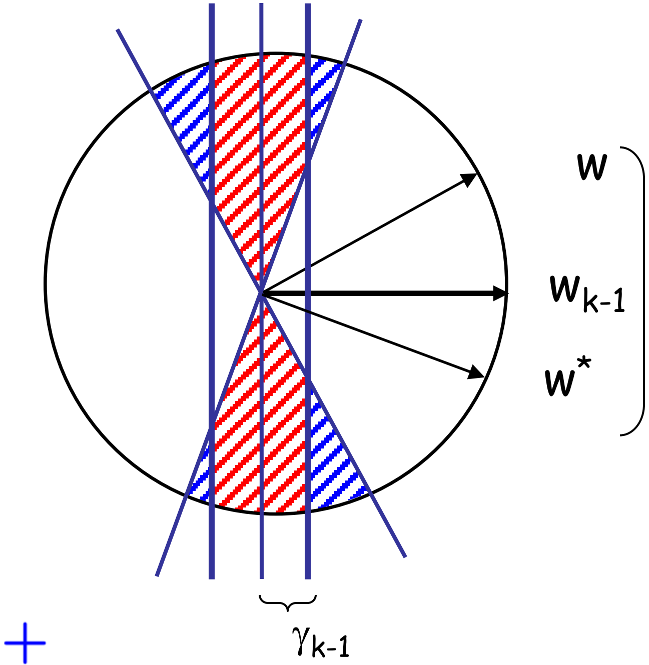
Let w s. t. $\mathrm{err}_{D_{|\mu_i}}(w) \le C$ for $i \le k$.

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

Proof Idea

By induction $\mathrm{err}_{\mathrm{D}}(\mathrm{w}) \leq \frac{1}{2^{k-1}}$, so $\theta(w, w^*) \leq 2^{-k+1}$

Also $\theta(w_{k-1}, w^*) \leq 2^{-k+1}$

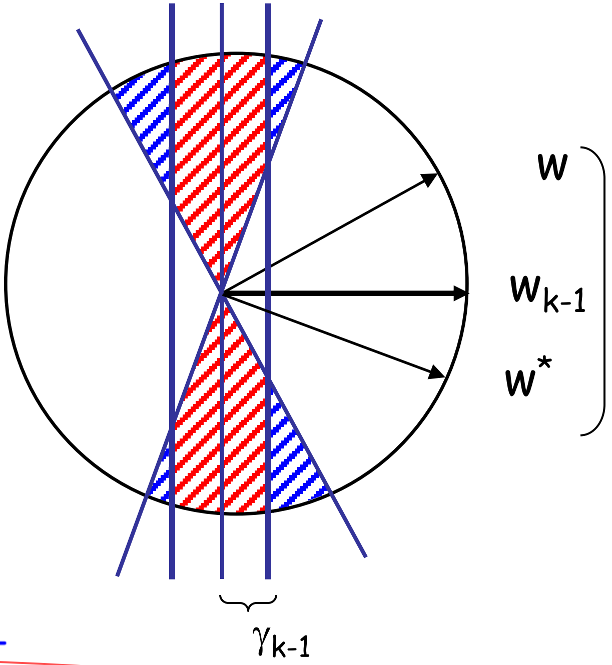


For $\gamma_k = O\left(\frac{c}{2^k}\right)$

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

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

Proof Idea



By induction $\mathrm{err}_{\mathrm{D}}(w) \le \frac{1}{2^{k-1}}$, so $\theta(w, w^*) \le 2^{-k+1}$

Also $\theta(w_{k-1}, w^*) \le 2^{-k+1}$

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

$\le 1/2^{k+1}$

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

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

By assumption $\le C$ $\le C\gamma_{k-1}.$

$\le 1/2^{k+1}$

So $\mathrm{err}_{\mathrm{D}}(\mathrm{w}) \le \frac{1}{2^k}$, as desired.

Active SQ Learning of Linear Separators

Theorem D log-concave in R^d.

If $\gamma_k = O\left(\frac{c}{2^k}\right)$ then after $s = \log\left(\frac{1}{\epsilon}\right)$ iterations $\mathrm{err}(w_s) \leq \epsilon$

Total number of labeled examples $\mathrm{poly}\left(d, \log\left(\frac{1}{\epsilon}\right)\right)$

<u>Label complexity</u>:

- Round k, passive SQ over $D_{|\chi_k}$, get $\mathrm{err}_{D_{|\chi_k}}(w_k) \leq \epsilon' = \frac{c}{k}$.
- Only $\mathrm{poly}(d, \epsilon')$ passive SQs over $D_{|\chi_k}$ with $\tau = 1/\mathrm{poly}(d, \epsilon')$

  [can be implemented as active SQs with $\tau = 1/\mathrm{poly}(d, \epsilon')$, $\tau_f = C\epsilon$].

Active Differential Privacy

Learner has full access to unlabeled portion of database S.

For every element of S can request the label.

Goal: 1. Do learning while minimize # label request

2. Ensure differential privacy [modifying a record in S does not affect much prob. that any h is output]

x_1	
…	
x_i	
…	
x_n	

Active Differential Privacy

- A is α-*differentially private* if for any two neighbor datasets S, S′ (differ in just one element $(x_i, y_i) \rightarrow (x_i', y_i')$).

x_1	
...	
x_i	y_i
...	
x_n	

x_1	
...	
x_i'	y_i'
...	
x_n	

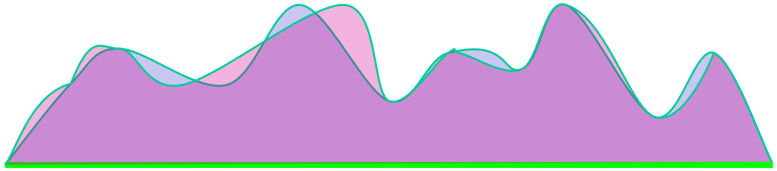
For all outcomes v,

$$e^{-\alpha} \leq \frac{\Pr(A(S) = v)}{\Pr(A(S') = v)} \leq e^{\alpha}$$

$\approx 1 + \alpha$

$\approx 1 - \alpha$

Prob. over randomness in A



Active Differential Privacy

Theorem Any active SQ alg with M queries of tolerance τ, filter-tolerance τ_f, can be made to preserve α-Diff Privacy using $O(\left(\frac{M}{\alpha\tau} + \frac{M}{\tau^2}\right)\log(M))$ label requests, $O(\left(\frac{M}{\alpha\tau\tau_f} + \frac{M}{\tau^2\tau_f}\right)\log(M))$ unlabeled examples.

Privacy cost original

Privacy cost original

Implications:

- For $\alpha \geq \tau$, privacy "for free" in terms of # of labeled requests.
- For lin. sep. & thresholds, can learn and preserve DP with much fewer label requests than non-private passive as long as α is large compared to ϵ.

Active Differential Privacy

Theorem Any active SQ alg with M queries of tolerance τ, filter-tolerance τ_f, can be made to preserve α-Diff Privacy using $\mathrm{O}\left(\left(\underbrace{\frac{M}{\alpha\tau}}_{\text{Privacy cost}}+\underbrace{\frac{M}{\tau^2}}_{\text{original}}\right)\log(M)\right)$ label requests, $O\left(\left(\underbrace{\frac{M}{\alpha\tau\tau_f}}_{\text{Privacy cost}}+\underbrace{\frac{M}{\tau^2\tau_f}}_{\text{original}}\right)\log(M)\right)$ unlabeled examples.

Proof sketch:

- Answer each query using disjoint set of $O\left(\left(\frac{1}{\alpha\tau\tau_f}+\frac{1}{\tau^2\tau_f}\right)\log(M)\right)$ unlabeled exs.
- Will query the T examples that pass the filter and add Laplace noise.
- Sets are disjoint so suffices to satisfy α-DP per query.
- Query sensitivity is $\frac{1}{T}$, so suffices to add $\frac{1}{\alpha T}$ Laplace noise per query.
- Sample size large enough so that whp, noise added is $\leq \tau/2$. Combine with $\tau/2$ from sample size to get whp overall error $\leq \tau$ per query.

Discussion

Model for designing Statistical Active Learning (AL) algos

- Poly time statistical AL algos → poly time algos tolerant to random classification noise.
- Naturally lead to differentially private AL algorithms.

Open questions

Deal with more general types of noise [ABL'13].

Practical Implications?