Active Learning Drifting Distributions, and Convex Surrogate Losses

Liu Yang

Carnegie Mellon University

Outline

- Active Learning with a Drifting Distribution ([Yang11 NIPS])

Active Learning with a Drifting Distrib: Model

- Scenario:

  - Unobservable seq. of distrib.s $D_1, D_2, \ldots$ with each $D_t \in \mathcal{D}$

  - Unobservable time-indep. regular cond. distrib. represent by fn

$$\eta : X \to [0, 1]$$

  - $\mathcal{Z} = \{(X_t, Y_t)\}_{t=1}^{\infty}$: an infinite seq. of indep. r. v., s.t., $\forall t, X_t \sim D_t$

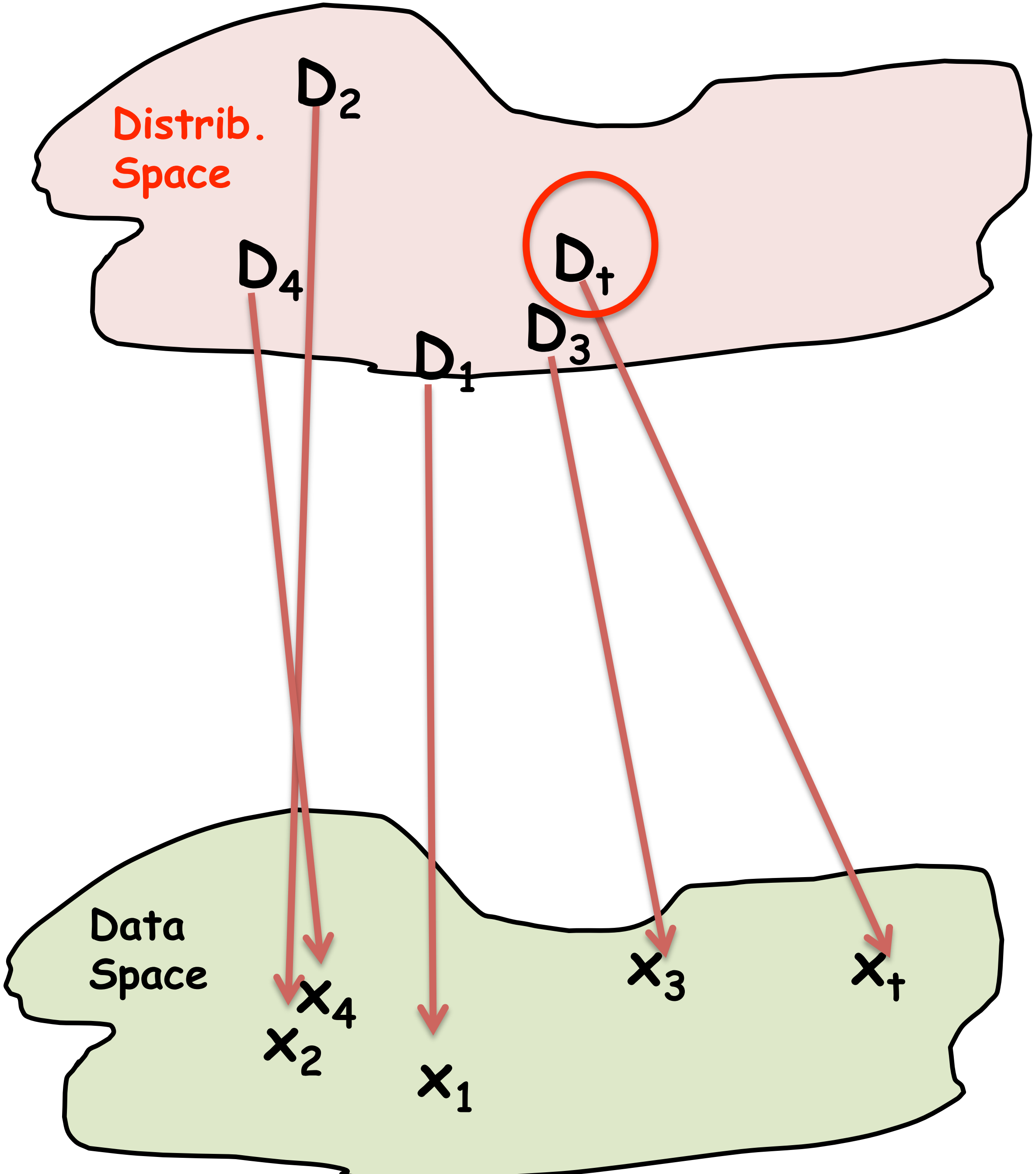
and cond. distrib. Of Y_t given X_t satisfies

$$\forall x \in X, P(Y_t = +1 | X_t = x) = \eta(x)$$

- Active learning protocol

  At each time t, alg is presented with X_t, and is required to predict a label $\hat{Y}_t \in \{-1, +1\}$, then it may optionally request to see true label value Y_t

- Interested in cumulative #mistakes up to time T and total #labels requested up to time T



Distrib.
Space
D_2
D_4
D_t
D_3
D_1
Data
Space
x_4
x_2
x_1
x_3
x_t

Definition and Notations

- Instance space $X = R^n$
- Distribution space $\mathcal{D}$ of distributions on X
- Concept space C of classifiers h: X -> {-1,1}
  - Assume C has VC dimension vc < ∞
- D_t: Data distrib. on X at t
- Unknown target fn h*: true labeling fn
- $Err_t(h) = P_{x \sim D_t}[h(x) \neq h^*(x)]$
- In realizable case, h* in C and $err_t(h^*) = 0$.
- For $V \subseteq C$, $diam_t(V) = \sup_{h,g \in V} D_t(\{x : h(x) \neq g(x)\})$

Def: disagreement coefficient, tvd

- The disagreement coefficient of h* under a distri. P on X, is define as, (r > 0)

$$\theta_P(\epsilon) = \sup_{r>\epsilon} P(DIS(B_P(h^*, r)))/r.$$
$$DIS(V) = \{x \in \mathcal{X} : \exists h, g \in V \text{ s.t. } h(x) \neq g(x)\}$$
$$B_P(h, r) = \{g \in C : P(x : h(x) \neq g(x)) \leq r\}$$

- Total variation distance of probability measures P and Q on a sigma-algebra $\mathcal{G}$ of subsets of the sample space is defined via

$$||P - Q|| = sup_{A \in \mathcal{G}} |P(A) - Q(A)|$$

Assumptions

- Independence of the X_t variables
- Vc-dim < ∞
- Assumption 1 (totally bounded) : $\mathcal{D}$ is totally bounded (i.e. satisfies $\forall \epsilon > 0, |\mathcal{D}_\epsilon| < \infty$)

  - For each ε > 0, $\mathcal{D}_\epsilon$ denote a minimal subset of $\mathcal{D}$ s.t.

  $\forall D \in \mathcal{D}, \exists D' \in \mathcal{D}_\epsilon$ s.t. $\|D - D'\| < \epsilon$ (i.e. a minimal ε-cover of $\mathcal{D}$)
- Assumption 2 (poly-covers)

$$\forall \epsilon > 0, |\mathcal{D}_\epsilon| < c \cdot \epsilon^{-m}$$

  where c,m ≥ 0 are constants.

Realizable-case Active Learning CAL

CAL

1. $t \leftarrow 0$, $\mathcal{Q}_0 \leftarrow \emptyset$, and let $\hat{h}_0 = \mathcal{A}(\emptyset)$
2. Do
3. $\quad t \leftarrow t + 1$
4. $\quad$ Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$
5. $\quad$ If $\max_{y \in \{-1,+1\}} \min_{h \in \mathbb{C}} \hat{\text{er}}(h; \mathcal{Q}_{t-1} \cup \{(X_t, y)\}) = 0$
6. $\quad\quad$ Request Y_t, let $\mathcal{Q}_t = \mathcal{Q}_{t-1} \cup \{(X_t, Y_t)\}$
7. $\quad$ Else let $Y_t' = \operatorname{argmin}_{y \in \{-1,+1\}} \min_{h \in \mathbb{C}} \hat{\text{er}}(h; \mathcal{Q}_{t-1} \cup \{(X_t, y)\})$, and let $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1} \cup \{(X_t, Y_t')\}$
8. $\quad$ Let $\hat{h}_t = argmin_{h \in C} \hat{er}(h; Q_t)$

Sublinear Result: Realizable Case

Theorem. If $\mathcal{D}$ is totally bounded, then CAL, achieves an expected mistake bound $\bar{M}_T = o(T)$
And if $\theta_{\mathcal{D}}(\epsilon) = o(1/\epsilon)$, then CAL makes an E[#queries]
$$\bar{Q}_T = o(T)$$

[Proof Sketch]:
Partition D into buckets of diam < eps.
Pick a time T_eps past all indices from finite buckets and all the infinite bucket has at least

$$L(\epsilon) = \lceil \frac{8}{\sqrt{\epsilon}} \left(d \ln \frac{24}{\sqrt{\epsilon}} + \ln \frac{4}{\sqrt{\epsilon}} \right) \rceil$$

Number of Mistakes

- Alternative scenario:
  - Let P_i be in bucket i
  - Swap the L(ε) samples for bucket i with L(ε) samples from P_i
  - L(ε) large enough so $E[diam(V)]_{alternative}$ < sqrt{eps}.

  - Note: E[diam(V)] ≤ $E[diam(V)]_{alternative}$ + $sum_{L(\varepsilon)\ t\ values}$||P_i - D_t|| < √ε + L(ε)*ε.

    So E[diam] -> 0 as T -> ∞
  - E[#mistake] $\leq \sum_{t=1}^{T} E[diam(V_{t-1})]$
  - Since $E[diam(V_{t-1})] \to 0, \sum_{t=1}^{T} E[diam(V_{t-1})] = o(T)$

Number of Queries

- E[#queries] $= \sum_{t=1}^{T} P(\text{make query})$
- P(make query) = E[P(DIS(V$_{t\text{-}1}$))]
  $E[\theta(r)\max\{diam, r\}] \leq \theta(r)E[diam] + \theta(r) \cdot r$
- Let $r_T = \frac{1}{T}\sum_{t=1}^{T} E[diam_t(V_{t-1})]$
  then $r_t \rightarrow 0$ and
  E[#queries]$\leq \theta(r_T)\sum_{t=1}^{T} E[diam_t(V_{t-1})] + \theta(r_T)r_T = \theta(r_T)r_T(T+1)$
- $\theta(\epsilon) = o(1/\epsilon) => \theta(r_T)r_T \rightarrow 0$ => $\theta(r_T)r_T(T+1) = o(T)$

Explicit Bound: Realizable Case

Theorem. If poly-covers assumption is satisfied ($|\mathcal{D}_\epsilon| < (1/\epsilon)^m$) then CAL achieves an expected mistake bound $\bar{M}_T$ and E[#queries] $\bar{Q}_T$ such that

$$\bar{M}_T = O\left(T^{\frac{m}{m+1}} d^{\frac{1}{m+1}} \log^2 T\right)$$
$$\bar{Q}_T = O\left(\theta_{\mathcal{D}}\left(\epsilon_T\right) T^{\frac{m}{m+1}} d^{\frac{1}{m+1}} \log^2 T\right)$$

where $\epsilon_T = (d/T)^{\frac{1}{m+1}}$

[Proof Sketch]

Fix any ε >0, and enumerate $\mathcal{D}_\epsilon = \{P_1, P_2, \cdots, P_{|\mathcal{D}_\epsilon|}\}$

For t in N, let K(t) be the index k of the closest $P_k \in \mathcal{D}_\epsilon$ to D_t.

Alternative data sequence:

Let $\{X'_t\}_{t=1}^{\infty}$ be indep., with $X_t \sim P_{K(t)}$

This way all samples corresp. to distrib.s in a given bucket all came from same distri.

Let V'_t be the corresponding version spaces.

E[#mistakes] $\leq E[\sum_{t=1}^{T} diam_{P_{K(t)}}(V'_{t-1})] + \sum_{t=1}^{T} \|D_t - P_{K(t)}\|$

$$\leq \sum_{t=1}^{T} E[diam_{P_{K(t)}}(V'_{t-1})] + \epsilon T$$

Classic PAC bound => $E[diam_{P_{K(t)}}(V'_{t-1})] \leq O(\frac{d \log t}{|\{i \leq t : K(i) = K(t)\}|})$

(#previous distrib.s in D_t's bucket)

So $\sum_{t=1}^{T} E[diam_{P_{K(t)}}(V'_{t-1})] \leq O(d \log T) \sum_{t=1}^{T} \frac{1}{|\{i \leq t : K(i) = K(t)\}|}$

$$\leq O\left(d \log T) |\mathcal{D}_\epsilon| \sum_{u=1}^{T} \frac{1}{u} \leq O(d |\mathcal{D}_\epsilon| \log^2(T)\right)$$

(each bucket has at most T samples)

So E[#mistakes] $\leq O(d(\frac{1}{\epsilon})^m \log^2(T) + \epsilon T)$

Take $\epsilon = (T/d)^{-\frac{1}{m+1}}$ to get the stated theorem.

To bound E[#queries], again it is

$$\leq E[\sum_{t=1}^{T} D_t(DIS(V_{t-1}))] \leq E[\sum_{t=1}^{T} \theta(\epsilon) \max\{diam_t(V_{t-1}, \epsilon)\}$$

$$\leq \theta(\epsilon) E[\sum_{t=1}^{T} diam_t(V_{t-1})] + \theta(\epsilon)\epsilon T$$

just showed this is $\leq O\left(d(\frac{1}{\epsilon})^m \log^2(T) + \theta(\epsilon)\epsilon T\right)$

So

$$O(\theta(\epsilon) d \left(\frac{1}{\epsilon}\right)^m \log^2(T) + \theta(\epsilon)\epsilon T)$$

Again, taking $\epsilon = (T/d)^{-\frac{1}{m+1}}$ gives the stated result.

Learning with Noise
Noise conditions

- Strictly benign noise condition:
  $h^* = \text{sign}(\eta - 1/2) \in C$ and $\forall x, \eta(x) \neq 1/2$
- Special case: Tsybakov's noise conditions
- η satisfies strictly benign noise condition and for some c > 0 and α≥0, $\forall t > 0, P(|\eta(x) - 1/2| < t) < c \cdot t^\alpha$

$$P(h(x) \neq h^*(x)) \leq c'(er(h) - er(h^*))^{\frac{\alpha}{\alpha+1}}$$

- Unif Tsybakov assumption: Tsybakov Assumption is satisfied for all $D \in \mathcal{D}$ with the same c and α values.

Agnostic CAL [DHM]

ACAL

1. $t \leftarrow 0$, $\mathcal{L}_t \leftarrow \emptyset$, $\mathcal{Q}_t \leftarrow \emptyset$, let $\hat{h}_t$ be any element of $\mathbb{C}$
2. Do
3. $\quad t \leftarrow t+1$
4. $\quad$ Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$
5. $\quad$ For each $y \in \{-1,+1\}$, let $h^{(y)} = \text{LEARN}(\mathcal{L}_{t-1} \cup \{(x_t, y)\}, \mathcal{Q}_{t-1})$
6. $\quad$ If either y has $h^{(-y)} = \varnothing$ or
   $\hat{\text{er}}(h^{(-y)}; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) - \hat{\text{er}}(h^{(y)}; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) > \hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$
7. $\quad\quad \mathcal{L}_t \leftarrow \mathcal{L}_{t-1} \cup \{(X_t, y)\}$, $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1}$
8. $\quad$ Else Request Y_t, and let $\mathcal{L}_t \leftarrow \mathcal{L}_{t-1}$, $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1} \cup \{(X_t, Y_t)\}$
9. $\quad$ Let $\hat{h}_t = \text{LEARN}(\mathcal{L}_t, \mathcal{Q}_t)$
10. $\quad$ If t is a power of 2
11. $\quad\quad \mathcal{L}_t \leftarrow \emptyset$, $\mathcal{Q}_t \leftarrow \emptyset$

Based on subroutine: $\text{LEARN}(\mathcal{L}, \mathcal{Q}) = \underset{h \in \mathbb{C}: \hat{\text{er}}(h;\mathcal{L})=0}{\text{argmin}} \hat{\text{er}}(h; \mathcal{Q})$ if $\min_{h \in \mathbb{C}} \hat{\text{er}}(h; \mathcal{L}) = 0$, and otherwise $\text{LEARN}(\mathcal{L}, \mathcal{Q}) = \varnothing$.

Tsybakov Noise: Sublinear Results & Explicit Bound

Theorem. If $\mathcal{D}$ is totally bounded and η satisfies strictly benign noise condition, then ACAL achieves an excess expected mistake bound

$$\bar{M}_T - M_T^* = o(T)$$

and if additionally $\theta_{\mathcal{D}}(\epsilon) = o(1/\epsilon)$, then ACAL makes an expected number of queries $\bar{Q}_T = o(T)$

Theorem. If poly-covers Assumption and Unif Tsybakov assumption are satisfied, then ACAL achieves an expected excess number of mistakes ACAL achieves expected #mistakes $\bar{M}$ and expected #queries $\bar{Q}_T$ such that, for $\epsilon_T = T^{-\frac{\alpha}{(\alpha+2)(m+1)}}$

$$\bar{M}_T - M_T^* = \tilde{O}\left(T^{\frac{(\alpha+2)m+1}{(\alpha+2)(m+1)}}\right)$$

$$\bar{Q}_T = \tilde{O}\left(\theta_{\mathcal{D}}(\epsilon_T) \cdot T^{\frac{(\alpha+2)(m+1)-\alpha}{(\alpha+2)(m+1)}}\right)$$

Outline

- Convex Losses in Active Learning
  (Joint work with Steve Hanneke)

Negative Results for AL with Convex Losses [AISTATS'10]

Fn class $\mathcal{F} of fns : f : \mathcal{X} \rightarrow R$

Loss fn $l : R \rightarrow [0, \infty)$

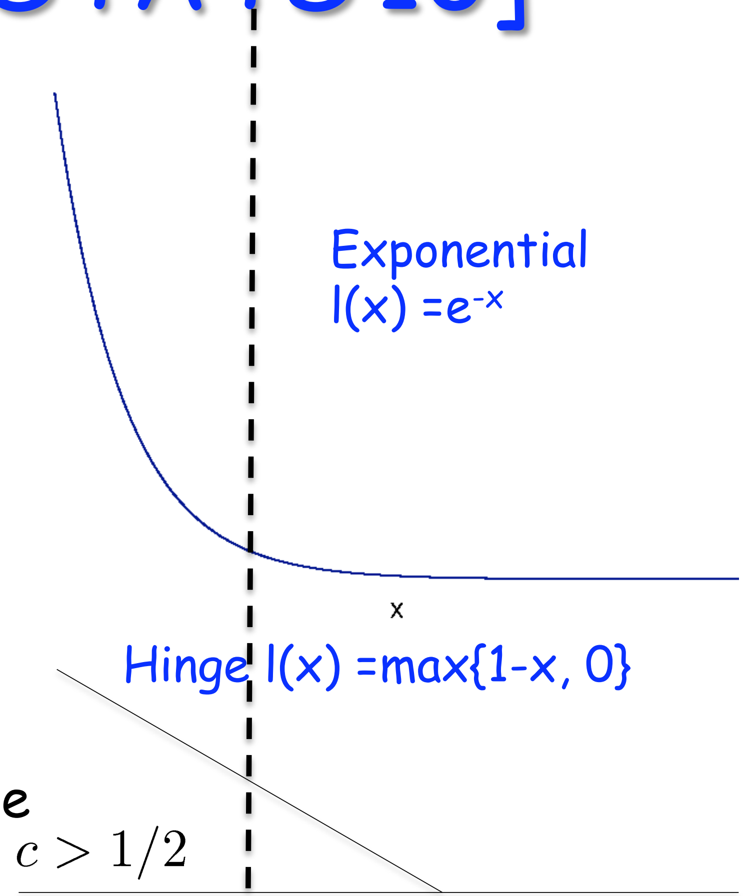
Interested in convex nonincreasing loss

Data distri. D still on $\mathcal{X} \times -1, +1$

Risk $R_l(f) = E[l(f(x)Y)] \text{ for } (X, Y) \sim \mathcal{D}_{XY}$

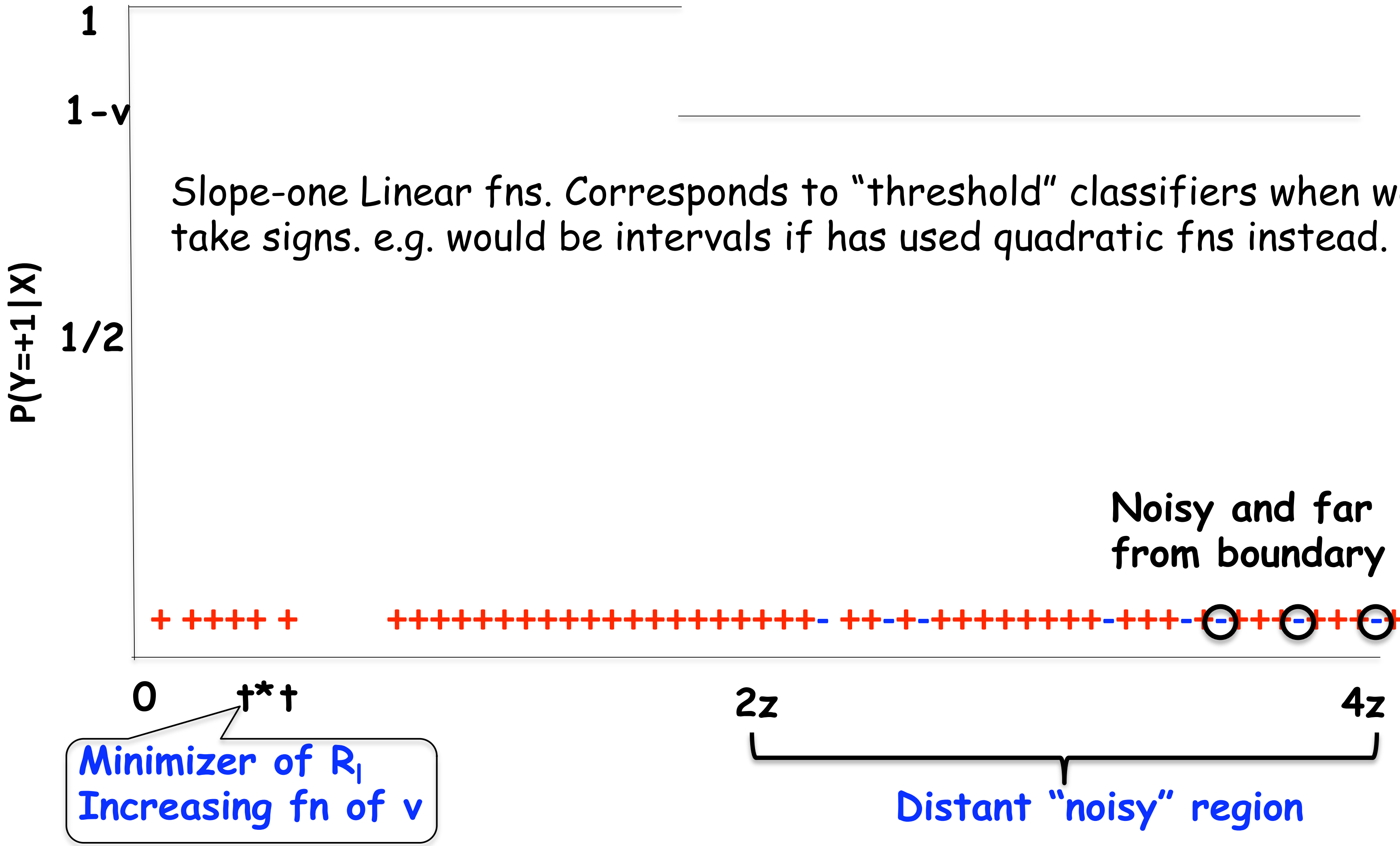
Question: How many labels needed to find $\hat{f} \in \mathcal{F} \text{ with } R_l(\hat{f}) - \inf_{f \in \mathcal{F}} R_l(f) \leq \epsilon$?

We'll study "Bounded Noise" Scenario where $\exists f \in \mathcal{F} \text{ s.t. } P(Y = \text{sign}(f(x))|x) > c \text{ for some } c > 1/2$



These are easy for active learning under 0-1 loss.
Now let us see about under convex losses.

$\mathcal{F} = \{f_t(x) = x - t : t \in R\} \quad \mathcal{D}_X = Uniform(0, 4z)$



Calculus + convexity $R_l(f_t) - R_l(f_{t^*}) \geq c(t - t^*)^2$

Let v_t be the v that would make t* = t

More calculus => $(t - t^*)^2 \geq (\nu_t - \nu_{t^*})^2$

So $R_l(f_{\hat{t}}) - R_l(f_{t^*}) \leq \epsilon$ **=>** $(\nu_{\hat{t}} - \nu_{t^*})^2 < c\epsilon$

estimating a Bernoulli mean requires Ω(1/ε) samples

Definition: Surrogate Losses

[BJM06]: For η_0 in [0,1], define

$$l^*(\eta_0) = \inf_{z \in R}(\eta_0 l(z) + (1-\eta_0) l(-z))$$
$$l^*_-(\eta_0) = \inf_{z \in R: z(2\eta_0 - 1) \le 0}(\eta_0 l(z) + (1-\eta_0) l(-z))$$

- Loss l is classification-calibrated if, for every η_0 in [0,1]\{1/2},

$$l^*_-(\eta_0) > l^*(\eta_0)$$

Calibration means: fn with minimal surrogate loss => fn with minimal err

$l^*_-(\eta(X))$: minimum value of conditional-risk at X s.t.

$$\text{sign}(h(X)) \neq \text{ sign } (\eta(X) - 1/2)$$

$l^*(\eta(X))$: minimum conditional l-risk at X, s.t. $E[l^*(\eta(X))] = \inf_h R_l(h)$

- Ψ-transform of a loss fn:

-BJM06 defined a loss-dependent function Ψ to convert excess surrogate risk bounds into excess error rate bounds, specifically,

$$(er(h) - er(h^*))^{\alpha/(1+\alpha)} \Psi((er(h) - er(h^*))^{1/(1+\alpha)}) \le R_l(h) - R_l(h^*)$$

- Modulus of convexity:

$$\delta(\epsilon) = \max\{(f(x) + f(y))/2 f((x+y)/2) : |x-y| > \epsilon\}$$

suppose $\delta(\epsilon) \ge \epsilon^p$

Alg: A modification on ACAL stream-based

ACAL

1. $t \leftarrow 0$, $\mathcal{L}_t \leftarrow \emptyset$, $\mathcal{Q}_t \leftarrow \emptyset$, let $\hat{h}_t$ be any element of $\mathbb{C}$
2. Do
3. $\quad t \leftarrow t+1$
4. $\quad$ Predict $\hat{Y}_t = \hat{h}_{t-1}(X_t)$
5. $\quad$ For each $y \in \{-1,+1\}$, let $h^{(y)} = \text{LEARN}(\mathcal{L}_{t-1} \cup \{(x_t, y)\}, \mathcal{Q}_{t-1})$
6. $\quad$ If either y has $h^{(-y)} = \varnothing$ or
   $\hat{\text{er}}(h^{(-y)}; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) - \hat{\text{er}}(h^{(y)}; \mathcal{L}_{t-1} \cup \mathcal{Q}_{t-1}) > \hat{\mathcal{E}}_{t-1}(\mathcal{L}_{t-1}, \mathcal{Q}_{t-1})$
7. $\quad\quad \mathcal{L}_t \leftarrow \mathcal{L}_{t-1} \cup \{(X_t, y)\}$, $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1}$
8. $\quad$ Else Request Y_t, and let $\mathcal{L}_t \leftarrow \mathcal{L}_{t-1}$, $\mathcal{Q}_t \leftarrow \mathcal{Q}_{t-1} \cup \{(X_t, Y_t)\}$
9. $\quad$ Let $\hat{h}_t = \text{LEARN}(\mathcal{L}_t, \mathcal{Q}_t)$
10. $\quad$ If t is a power of 2
11. $\quad\quad \mathcal{L}_t \leftarrow \emptyset$, $\mathcal{Q}_t \leftarrow \emptyset$

$R_L(h^{(-y)}, Q) - R_L(h^{(y)}, Q) > \hat{E}_l(Q)$

LEARN(L, Q) =
$\text{argmin}_{f \text{ in } F;\ er\_L(f)=0}\ R_l(f; Q)$

Based on subroutine: $\text{LEARN}(\mathcal{L}, \mathcal{Q}) = \underset{h \in \mathbb{C}: \hat{\text{er}}(h;\mathcal{L})=0}{\text{argmin}} \hat{\text{er}}(h; \mathcal{Q})$ if $\min_{h \in \mathbb{C}} \hat{\text{er}}(h; \mathcal{L}) =$

0, and otherwise $\text{LEARN}(\mathcal{L}, \mathcal{Q}) = \varnothing$.

Can we do it efficiently ? General Results

- In general, we have results on how many labels are required to obtain a given excess error rate with this method, for general classification calibrated losses.
- Generally, if ε_t denotes the solution of

$$t = \tilde{O}\left(\left(\frac{1}{\epsilon^\alpha \Psi(\epsilon^{1-\alpha})}\right)^{2-2/p}\right)$$

for ε in terms of t, then

E[excess #mistakes] = $\tilde{O}\sum_{t=1}^{T}\epsilon_t$

E[#queries] = $\tilde{O}\left(\sum_{t=1}^{T}\theta(\epsilon_t^\alpha)\epsilon_t^\alpha\right)$

e.g., when l is squared loss = $(1-x)^2$, $\psi(x) = x^2$, p= 2

Can we do it efficiently ?
(Streamed-based, just for one distri.)

- Theorem. If loss is square loss, under surrogate loss assumption, optimal fn is in fn class, fn class is VC subgraph, satisifying Tsybkov noise with exponent
- α/(1−α), alg A' has excess #mistake $\sum_{i=1}^{T}(1/t)^{\frac{1}{2-\alpha}}$

  E[excess #mistake] = Õ($T^{\frac{1-\alpha}{2-\alpha}}$)

  E[#queries] = Õ($\theta(T^{\frac{-\alpha}{2-\alpha}})T^{\frac{2-2\alpha}{2-\alpha}}$)

E[excess #mistake] sublinear
- if θ = o(1/ε),
E[#queries] sublinear.

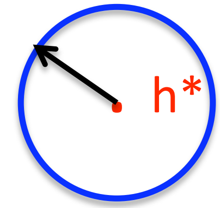
[Proof Sketch] By BJM06 analysis,

- If $t=(\frac{1}{\epsilon^{\alpha}\psi(1-\epsilon^{1-\alpha})})^{2-2/p}polylog(\log 1/\epsilon)$ then excess err rate < ε. This is sample complexity of passive learning with surrogate loss.
- E excess error under 0-1 loss

  Solve for t $\frac{1}{\epsilon^{\alpha}\psi(1-\epsilon^{1-\alpha})}=T$
- Get current excess error rate (as fn of t, bound on excess error rate, excess mistake = pr(make mistake but optimal fn doesn't)

  give excess err

Proof Sketch (cont.)



- If the loss is squared loss, fill in all the value, we get

$$\sum_{t=1}^{T} \left(\frac{1}{t}\right)^{\frac{1}{2-\alpha}} = T^{\frac{1-\alpha}{2-\alpha}}$$

- How to convert excess error to Pr(make a query)
- use Tsybakov noise condition

- Take $\left(\frac{1}{t}\right)^{\frac{1}{2-\alpha}}$, raise to the power of α, get diameter
- relate that to Pr(in DIS) by multiplying with θ (the disagreement coefficient, taking an argument)
- do that get $\sum_{t=1}^{T} \theta(t^{\frac{-\alpha}{2-\alpha}})\left(\frac{1}{t}\right)^{\frac{\alpha}{2-\alpha}} \leq \theta(T^{\frac{-\alpha}{2-\alpha}}) \sum_{t=1}^{T} t^{\frac{-\alpha}{2-\alpha}} = T^{\frac{2-2\alpha}{2-\alpha}}$
- plug in the bound on the diameter
- If θ is o(1/ε), this is sublinear

Thanks!