Active Learning from Single and Multiple Annotators

Kamalika Chaudhuri
University of California, San Diego

Joint work with Chicheng Zhang
Classification

Given: \((x_i, y_i)\)

Vector of features \hspace{2cm} \text{Discrete Labels}

Find: Prediction rule in a class to predict \(y\) from \(x\)
Challenge: Acquiring Labeled Data

Unlabeled data is cheap

Labels are expensive
Active Learning

Given: \((x_i, y_i)\)

Find: Prediction rule to predict \(y\) from \(x\)
Active Learning

Given: \( (x_i, y_i) \)

Find: Prediction rule to predict \( y \) from \( x \)

Interactive Label Queries
Active Learning

Given: \( (x_i, y_i) \)

Find: Prediction rule to predict \( y \) from \( x \) using few label queries
Why Active Learning Helps?

Given: Unlabeled data, interactive label queries  
Find: Good prediction rule using few label queries
Why Active Learning Helps?

Given: Unlabeled data, interactive label queries
Find: Good prediction rule using few label queries
Why Active Learning Helps?

Given: Unlabeled data, interactive label queries
Find: Good prediction rule using few label queries
Why Active Learning Helps?

Given: Unlabeled data, interactive label queries
Find: Good prediction rule using few label queries
Why Active Learning Helps?

Given: Unlabeled data, interactive label queries
Find: Good prediction rule using few label queries
Challenge: Agnostic Active Learning
PAC Model: Realizable vs. Agnostic

Given: Concept class C, Samples \((x_i, y_i)\) from D
Find: \(c\) in C with low error

\[
\exists c^* \in C \quad \text{such that} \\
c^*(x) = y, \forall (x, y) \sim D
\]

Realizable

Agnostic

No Assumptions on D
Agnostic Active Learning

Given: Concept class $C$ (best $c$ in $C$ has error $\nu^*$)

$(x_i, y_i)$ drawn from $D$

Find: $c$ in $C$ with error $\leq \nu^* + \epsilon$

using few label queries

with no assumptions on $D$
Why is Agnostic Active Learning hard?

Given: Unlabeled data, interactive label queries
No assumptions on data distribution
Find: Good prediction rule using few label queries
Why is Agnostic Active Learning hard?

Given: Unlabeled data, interactive label queries
No assumptions on data distribution
Find: Good prediction rule using few label queries
Why is Agnostic Active Learning hard?

Given: Unlabeled data, interactive label queries
No assumptions on data distribution
Find: Good prediction rule using few label queries

Statistically Inconsistent: finds local minimum!
Talk Outline

What is Active Learning?

Active Learning from Single Annotator [NIPS 14]

Active Learning from Multiple Annotators [NIPS 15]
Active Learning from a Single Annotator
Previous Work: Disagreement-based Active Learning

1. Maintain candidate set $V$ (that contains best $c$ in $C$)

2. For unlabeled $x$, if there exist $c_1, c_2 \in V$ s.t

\[ c_1(x) \neq c_2(x) \]

then, query label of $x$, and update $V$

[CAL94, BBL06, DHM07, H07, BDL09, BHLZ10, …]
Previous Work: Disagreement-based Active Learning

1. Maintain candidate set $V$ (that contains best $c$ in $C$)
2. For unlabeled $x$, if there exist $c_1, c_2$ in $V$ s.t
   \[ c_1(x) \neq c_2(x) \]
   then, query label of $x$, and update $V$

**Pro:** Very general

**Con:** High label requirement
Previous Work: Margin-based Active Learning

Geometric algorithm for linear classification wrt uniform and log-concave distribution on sphere

[BZ07, BL13, ABL14]

**Pro:** Label requirement

**Con:** Not general
Is there a general algorithm with better label complexity?
Talk Outline

What is Active Learning?

Active Learning from Single Annotator [NIPS 14]
  — Previous Work
  — Confidence Based Active Learning
Confidence-Rated Predictor (CRP)

Classifiers that can Abstain
Confidence-Rated Predictor (CRP)

Classifiers that can Abstain

Performance Metrics:
Error: Rate of mistakes
Abstention Rate: Rate of Don’t Know predictions
CRP with Guaranteed Error [EW10]

Input: Concept set $V$
Unlabeled data $U$

Error guarantee $\eta$
CRP with Guaranteed Error [EW10]

Input: Concept set $V$  
Unlabeled data $U$

Goal: Assign label $P(z)$ in $\{+1, -1, 0\}$ to each $z$ in $U$

Guarantee: For all $c$ in $V$,  
$$\Pr_{z \sim U}\left( P(z) = -c(z) \right) \leq \eta$$
Confidence-Based Active Learning

Two Key Ideas:

1. A reduction from Active Learning to CRP with guaranteed error

2. Construction of a CRP with guaranteed error
Algorithm Outline

1. Active Learning via CRP

2. How to construct a CRP with guaranteed error
Active Learning via CRP

Given:  
Unlabeled data U  
Concept class C  
Confidence-rate Predictor $P$  
Target excess error $\epsilon$

Goal: Output $c$ in $C$ with target excess error $\epsilon$ with min #label queries

Problem: Which labels to query?
Definition: Confidence Sets for $c^*$

Recall: $c^*$ is the concept in $C$ with min error

Confidence Set: Set $S$ of concepts s.t. w.p. $\geq 1 - \delta$, $c^*$ is in $S$

Similar to confidence intervals
How to construct confidence sets?

Use generalization bounds:

**Known:** \( \forall c, \text{w.p. } \geq 1 - \delta, |err(c) - \hat{err}(c)| \leq \sqrt{\frac{VC(C) \log(n/\delta)}{n}} \)

**Choose:** All \( c \) with:

\[
\hat{err}(c) \leq \min_{c' \in C} err(c') + \sqrt{\frac{VC(C) \log(n/\delta)}{n}}
\]

(A bit more complicated for active learning)
Active Learning via CRP

**Given:**  Unlabeled data $U$

Concept class $C$

Confidence-rate Predictor $P$

Target error $\epsilon$

1. Initialize: Confidence set for $c^*$ $V_1 = C$, $d = VCdim(C)$.

2. For epoch $k = 1, 2, 3, \ldots$
   (a) Target error for epoch $k$: $\epsilon_k = \frac{1}{2^k}$

**Problem:** How to query labels?
Active Learning via CRP

Given: Unlabeled data $U$
Concept class $C$
Confidence-rate Predictor $P$
Target error $\epsilon$

1. Initialize: Confidence set for $c^*$ $V_1 = C$, $d = \text{VCDim}(C)$.

2. For epoch $k = 1, 2, 3, \ldots$
   (a) Target error for epoch $k$: $\epsilon_k = \frac{1}{2^k}$

Key Idea 1: Run $P$ with error guarantee $\epsilon_k/64$, and query label when $P$ abstains
Active Learning via CRP

Given: Unlabeled data $U$
Concept class $C$
Confidence-rate Predictor $P$
Target error $\epsilon$

1. Initialize: Confidence set for $c^*$ $V_1 = C, d = V\text{Cdim}(C)$.

2. For epoch $k = 1, 2, 3, \ldots$
   (a) Target error for epoch $k$: $\epsilon_k \leq \frac{1}{2^k}$

Key Idea 1: Run $P$ with error guarantee $\epsilon_k/64$, and query label when $P$ abstains

Key Idea 2: Query just enough labels to get target excess error $\epsilon_{k+1}/\phi_k$ on the Don’t Knows ( $\phi_k$ = $P$’s abstention prob.)
Active Learning via CRP

Given: Unlabeled data $U$  
Concept class $C$  
Confidence-rate Predictor $P$  
Target error $\epsilon$

1. Initialize: Confidence set for $c^*$ $V_1 = C$, $d = VCdim(C)$.

2. For epoch $k = 1, 2, 3, \ldots$  
   (a) Target error for epoch $k$: $\epsilon_k = \frac{1}{2^k}$

Key Idea 1: Run $P$ with error guarantee $\epsilon_k/64$, and query label when $P$ abstains

Key Idea 2: Query just enough labels to get target excess error $\epsilon_{k+1}/\phi_k$ on the Don’t Knows ( $\phi_k = P$’s abstention prob.)

Low abstention prob. means less labels queried
Active Learning via CRP

Given: Unlabeled data $U$  
Concept class $C$  
Confidence-rate Predictor $P$  
Target error $\epsilon$

1. Initialize: Confidence set for $c^*$ $V_1 = C$, $d = VCdim(C)$.

2. For epoch $k = 1, 2, 3, \ldots$  
   (a) Target error for epoch $k$: $\epsilon_k = \frac{1}{2^k}$  
   (b) Run CRP $P$ with $V = V_k$, unlabeled data, error guarantee $\eta = \epsilon_k / 64$
Active Learning via CRP

Given: Unlabeled data $U$  
       Concept class $C$  
       Confidence-rate Predictor $P$  
       Target error $\epsilon$

1. Initialize: Confidence set for $c^*$ $V_1 = C$, $d = VCdim(C)$.

2. For epoch $k = 1, 2, 3, \ldots$
   (a) Target error for epoch $k$: $\epsilon_k = \frac{1}{2^k}$
   (b) Run CRP $P$ with $V = V_k$, unlabeled data, error guarantee $\eta = \epsilon_k / 64$
   (c) Whenever $P$ abstains, query enough labels to get excess error $\epsilon_k / \phi_k$ on the Don’t Knows ($\phi_k$ is $P$’s abstention rate).
Active Learning via CRP

Given: Unlabeled data \( U \)  

Concept class \( C \)  

Confidence-rate Predictor \( P \)  

Target error \( \epsilon \)

1. Initialize: Confidence set for \( c^* \) \( V_1 = C \), \( d = \text{VCdim}(C) \).

2. For epoch \( k = 1, 2, 3, \ldots \).  
   (a) Target error for epoch \( k \): \( \epsilon_k = \frac{1}{2^k} \)
   (b) Run CRP \( P \) with \( V = V_k \), unlabeled data, error guarantee \( \eta = \epsilon_k / 64 \)
   (c) Whenever \( P \) abstains, query enough labels to get excess error \( \epsilon_k / \phi_k \) on the Don’t Knows (\( \phi_k \) is \( P \)’s abstention rate).
   (d) Update: Confidence set \( V_{k+1} \) (based on new labeled samples)
Active Learning via CRP

Given: Unlabeled data \( U \)  
  Concept class \( C \)  
  Confidence-rate Predictor \( P \)  
  Target error \( \epsilon \)

1. Initialize: Confidence set for \( c^* \) \( V_1 = C, d = \text{VCdim}(C) \).
2. For epoch \( k = 1, 2, 3, \ldots \):
   (a) Target error for epoch \( k \): \( \epsilon_k = \frac{1}{2^k} \)
   (b) Run CRP \( P \) with \( V = V_k \), unlabeled data, error guarantee \( \eta = \epsilon_k / 64 \)
   (c) Whenever \( P \) abstains, query enough labels to get excess error \( \epsilon_k / \phi_k \) on the Don’t Knows (\( \phi_k \) is \( P \)'s abstention rate).
   (d) Update: Confidence set \( V_{k+1} \) (based on new labeled samples)
3. Output any \( c \) in \( V_{\log(1/\epsilon)} \)
Algorithm Outline

1. Active Learning via CRP

2. How to construct a CRP with guaranteed error
CRP with Guaranteed Error [EW10]

Input: Concept set $V$  
Unlabeled data $U$

Goal: Assign label $P(z)$ in $\{+1, -1, 0\}$ to each $z$ in $U$

Guarantee: For all $c$ in $V$,  
$$\Pr_{z \sim U} (P(z) = -c(z)) \leq \eta$$
CRP with Guaranteed Error

Given:  Concept set $V$  Error guarantee $\eta$  Unlabeled data $U$
CRP with Guaranteed Error

Given: Concept set $V$  Error guarantee $\eta$  Unlabeled data $U$

1. Construct and solve Linear Program:

   For each $z_i$ in $U$, variables: $\alpha_i$ : Prob. of predicting $+1$ on $z_i$
   $\beta_i$ : Prob. of predicting $-1$ on $z_i$
CRP with Guaranteed Error

Given: Concept set $V$ Error guarantee $\eta$ Unlabeled data $U$

1. Construct and solve Linear Program:

For each $z_i$ in $U$, variables:

- $\alpha_i$: Prob. of predicting +1 on $z_i$
- $\beta_i$: Prob. of predicting -1 on $z_i$

$$\max \sum_{i} \alpha_i + \beta_i \quad |U| \times \text{Non-abstention rate}$$
CRP with Guaranteed Error

**Given:** Concept set $V$  
Error guarantee $\eta$  
Unlabeled data $U$

I. Construct and solve Linear Program:

For each $z_i$ in $U$, variables:

- $\alpha_i$: Prob. of predicting +1 on $z_i$
- $\beta_i$: Prob. of predicting -1 on $z_i$

\[
\max \sum_i \alpha_i + \beta_i \quad \text{max} \quad |U| * \text{Non-abstention rate}
\]

\[
\text{s.t:} \forall c \in V, \sum_{i : c(z_i) = 1} \beta_i + \sum_{i : c(z_i) = -1} \alpha_i \leq \eta |U| \quad \text{Error guarantee}
\]
CRP with Guaranteed Error

Given:  Concept set V   Error guarantee \( \eta \)   Unlabeled data U

1. Construct and solve Linear Program:

For each \( z_i \) in U, variables:
- \( \alpha_i \): Prob. of predicting +1 on \( z_i \)
- \( \beta_i \): Prob. of predicting -1 on \( z_i \)

\[
\max \sum_i \alpha_i + \beta_i
\]

s.t:
- \( \forall c \in V, \sum_{i: c(z_i)=1} \beta_i + \sum_{i: c(z_i)=-1} \alpha_i \leq \eta |U| \)  Error guarantee
- \( \forall i, \alpha_i + \beta_i \leq 1, \alpha_i, \beta_i \geq 0 \)  Probability constraint

|U| * Non-abstention rate
CRP with Guaranteed Error

Given: Concept set $V$, Error guarantee $\eta$, Unlabeled data $U$

1. Construct and solve Linear Program:

$$\max \sum_i \alpha_i + \beta_i$$

s.t: $\forall c \in V, \sum_{i : c(z_i) = 1} \beta_i + \sum_{i : c(z_i) = -1} \alpha_i \leq \eta |U|$  
$\forall i, \alpha_i + \beta_i \leq 1, \alpha_i, \beta_i \geq 0$

2. For $z_i$ in $U$, output: 1 w.p. $\alpha_i$, $-1$ w.p. $\beta_i$, 0 ow

Note: Optimal abstention rate given concept set $V$
Algorithm Outline

1. Active Learning via CRP

2. How to construct a CRP with guaranteed error
Talk Outline

What is Active Learning?

Active Learning from Single Annotator [NIPS 14]
  — Previous Work
  — Confidence Based Active Learning
  — Performance Guarantees
Consistency Guarantees

Theorem: If P is a CRP with guaranteed error, then our active learning algorithm is consistent.

Note: Any CRP with guaranteed error gives consistency.
Label Complexity: Agnostic Case

Parameters: Target error $\epsilon$    VC dimension $d$
            Error of best $c$ in $C$ $\nu^*$
Label Complexity: Agnostic Case

Parameters: Target error $\epsilon$  VC dimension $d$
            Error of best $c$ in $C$ $\nu^*$

Disagreement-based: $\theta(2\nu^* + \epsilon)(\frac{d(\nu^*)^2 \log(1/\epsilon)}{\epsilon^2} + d \log^2 (1/\epsilon))$
Label Complexity: Agnostic Case

Parameters:  Target error $\epsilon$  VC dimension $d$

Error of best $c$ in $\mathbb{C}$ $\nu^*$

Disagreement-based: \( \theta(2\nu^* + \epsilon) \left( \frac{d(\nu^*)^2 \log(1/\epsilon)}{\epsilon^2} + d \log^2(1/\epsilon) \right) \)

Rate of change of disagreement region
Label Complexity: Agnostic Case

Parameters:
- Target error $\epsilon$
- VC dimension $d$
- Error of best $c$ in $C$, $\nu^*$

Disagreement-based: $\theta(2\nu^* + \epsilon) \left( \frac{d(\nu^*)^2 \log(1/\epsilon)}{\epsilon^2} + d \log^2(1/\epsilon) \right)$

Rate of change of disagreement region

Our Algorithm: $\sigma(2\nu^* + \epsilon, \epsilon/256) \left( \frac{d(\nu^*)^2 \log(1/\epsilon)}{\epsilon^2} + d \log^2(1/\epsilon) \right)$
Label Complexity: Agnostic Case

Parameters: Target error \( \epsilon \)  
VC dimension \( d \)
Error of best \( c \) in \( C \)  \( \nu^* \)

Disagreement-based:
\[
\theta(2\nu^* + \epsilon) \left( \frac{d(\nu^*)^2 \log(1/\epsilon)}{\epsilon^2} + d \log^2(1/\epsilon) \right)
\]
Rate of change of disagreement region

Our Algorithm:
\[
\sigma(2\nu^* + \epsilon, \epsilon/256) \left( \frac{d(\nu^*)^2 \log(1/\epsilon)}{\epsilon^2} + d \log^2(1/\epsilon) \right)
\]
Rate of change of DK region
Lower than \( \theta(2\nu^* + \epsilon) \)
Label Complexity: Agnostic Case

Parameters: 
\[ \text{Target error } \epsilon \quad \text{VC dimension } d \]

\[ \text{Error of best } c \text{ in } C \quad \nu^* \]

Disagreement-based: 
\[ \theta(2\nu^* + \epsilon) \left( \frac{d(\nu^*)^2 \log(1/\epsilon)}{\epsilon^2} + d \log^2(1/\epsilon) \right) \]

Rate of change of disagreement region

Our Algorithm: 
\[ \sigma(2\nu^* + \epsilon, \epsilon/256) \left( \frac{d(\nu^*)^2 \log(1/\epsilon)}{\epsilon^2} + d \log^2(1/\epsilon) \right) \]

Rate of change of DK region
Lower than \[ \theta(2\nu^* + \epsilon) \]

Passive Learning: 
\[ \frac{d\nu^* \log(1/\epsilon)}{\epsilon^2} \]
Linear Classification on Logconcave Distribution on Sphere

Our Algorithm: (Realizable Case)

\[ d \log(1/\epsilon) + \log(1/\epsilon) \log \log(1/\epsilon) \]

(matches Margin-based Active Learning [BL13])

Disagreement-based Active Learning:

\[ d^{3/2} \log^2(1/\epsilon)(\log d + \log \log(1/\epsilon)) \]
Conclusions

Novel reduction from active learning to CRP
  - Exploits non-zero error guarantee for better label reqmt

Novel active learning algorithm:
  - Better asymptotic label complexity than disagreement based
  - More general than margin based
Talk Outline

What is Active Learning?

Active Learning from Single Annotator [NIPS 14]
  — Previous Work
  — Confidence Based Active Learning
  — Performance Guarantees

Active Learning from Multiple Annotators [NIPS 15]
What if we have auxiliary information?
..as an extra oracle
Oracle and Weak Labeler

Oracle:
expensive but correct

Weak labeler:
cheap, sometimes wrong
The Model

Given: \( (x_i, y_i) \)

Interactive Label Queries
The Model

Given: \((x_i, y_i)\)

Interactive Label Queries to Oracle O or Weak Labeler W
The Model

Given: \((x_i, y_i)\)

Interactive Label Queries to Oracle O or Weak Labeler W

Find: Prediction rule to predict y from x using few label queries to O
Formal Model

Given: Concept class $C$ (best $c$ has error $\nu^*$ wrt $O$)

$(x_i, y_i)$ drawn from $D$
Formal Model

Given: Concept class $C$ (best $c$ has error $\nu^*$ wrt $O$)

$(x_i, y_i)$ drawn from $D$

Oracle $O$ and Weak labeler $W$
Given:  Concept class $C$ (best $c$ has error $\nu^*$ wrt $O$) 

$\langle x_i, y_i \rangle$ drawn from $D$ 

Oracle $O$ or Weak labeler $W$

Find:  $c$ in $C$ with error $\leq \nu^* + \epsilon$ wrt $O$ 

using minimum label queries to $O$
Formal Model Implications

Weak labeler $W$ may be biased

$y_O = -1$

$y_O = 1$

$y_W = -1$

$y_W = 1$

Labels by $O$

Labels by $W$
Previous Work

[UBS12] Explicit assumptions on where W and O differ, provides algorithms in the non-parametric case.

[MCR14] No explicit assumptions, but applies to online selective linear classification and robust regression.

This talk: General learning strategy from W and O with no explicit assumptions.
Talk Outline

What is Active Learning?

Active Learning from Single Annotator [NIPS 14]

Active Learning from Multiple Annotators [NIPS 15]
  — The Model
  — Algorithm
How to learn in this model?

Main Ideas:

Learn a difference classifier $h$ to predict when $O$ and $W$ differ
How to learn in this model?

Main Ideas:

Learn a **difference classifier** $h$ to predict when $O$ and $W$ differ

Use $h$ with standard active learning to decide if we should query $O$ or $W$
Algorithm Outline

1. Draw $x_1,..,x_m$. For each $x_i$, query $O$ and $W$. Set:

$$y_{i,D} = 1 \text{ if } y_{i,O} \neq y_{i,W}$$
Algorithm Outline

1. Draw $x_1, \ldots, x_m$. For each $x_i$, query $O$ and $W$. Set:
   $$y_{i,D} = 1 \quad \text{if} \quad y_{i,O} \neq y_{i,W}$$

2. Train difference classifier $h$ in $H$ on $\{(x_i, y_{i,D})\}$
Algorithm Outline

1. Draw $x_1, \ldots, x_m$. For each $x_i$, query $O$ and $W$. Set:
   $$y_{i,D} = 1 \text{ if } y_{i,O} \neq y_{i,W}$$

2. Train **difference classifier** $h$ in $H$ on $\{(x_i, y_{i,D})\}$

3. Run standard disagreement based active learning algorithm $A$. If $A$ queries the label of $x$ then:
   - if $h(x) = 1$, query $O$, else query $W$
Algorithm Outline

1. Draw $x_1, \ldots, x_m$. For each $x_i$, query $O$ and $W$. Set:
   $$ y_{i,D} = 1 \quad \text{if} \quad y_{i,O} \neq y_{i,W} $$

2. Train difference classifier $h$ in $\mathcal{H}$ on $\{(x_i, y_{i,D})\}$

3. Run standard disagreement based active learning algorithm $A$. If $A$ queries the label of $x$ then:
   $$ \text{if } h(x) = 1, \text{ query } O, \text{ else query } W $$

Is this statistically consistent?
Key Observation 1

Directly learning difference classifier may lead to inconsistent annotation on target task

\( y_0 = 1 \)  

Actual Labels
Key Observation 1

Directly learning difference classifier may lead to inconsistent annotation on target task

\[ y_w = -1 \]

\[ y_o = 1 \]

Actual Labels
Key Observation 1

Directly learning difference classifier may lead to inconsistent annotation on target task

\[ y_w = -1 \]

\[ h^* \]

\[ y_0 = 1 \]

Actual Labels
Key Observation 1

Directly learning difference classifier may lead to inconsistent annotation on target task

Actual Labels

$y_w = -1$

$y_o = 1$

Annotation using $h^*$ as difference classifier

Query $W$

Query $O$

$h^*$
Key Observation 1

Directly learning difference classifier may lead to inconsistent annotation on target task

\[ y_w = -1 \]
\[ y = 1 \]

Actual Labels

Annotation using \( h^* \) as difference classifier
Solution

Train a **cost-sensitive** difference classifier
Constrain False Negative (FN) rate as very low

\[
y_w = -1
\]

\[
y_o = 1
\]

Actual Labels
Solution

Train a cost-sensitive difference classifier
Constrain False Negative (FN) rate as very low

Actual Labels

\[ y_w = -1 \]
\[ y_o = 1 \]

\[ h_{FN}^* \]

Annotation using \( h_{FN}^* \) as difference classifier

Query W

Query O
Solution

Train a **cost-sensitive** difference classifier

Constrain False Negative (FN) rate as very low

\[
y_w = -1
\]

\[
y_\odot = 1
\]

Actual Labels

\[
h^*_\text{FN}
\]

Annotation using \(h^*_\text{FN}\) as difference classifier
Algorithm Outline

1. Draw $x_1, \ldots, x_m$. For each $x_i$, query $O$ and $W$. Set:
   
   $y_{i,D} = 1$ if $y_{i,O} \neq y_{i,W}$

2. Train difference classifier $h$ in $H$ on $\{(x_i, y_{i,D})\}$
   with false negative (FN) rate $\leq \epsilon$

3. Run standard disagreement based active learning algorithm $A$. If $A$ queries the label of $x$ then:
   
   if $h(x) = 1$, query $O$, else query $W$

Theorem: This is statistically consistent
What about label complexity?

Label complexity = \#label queries to O
#labels to train difference classifier $\approx \tilde{O}\left(\frac{d'}{\epsilon}\right)$

$(d' = \text{VCdim}(H), \quad \epsilon = \text{target excess error})$

Can we do better?
Key Observation 2

\[ R = \text{disagreement region of current confidence set} \]
Key Observation 2

$R = \text{disagreement region of current confidence set}$

Need to learn difference classifier with FN rate

$\leq \epsilon / \Pr(R) \text{ over } R$
Key Observation 2

\( R = \) disagreement region of current confidence set

Need to learn difference classifier with FN rate

\[ \leq \epsilon / \Pr(R) \text{ over } R \]

Need \( \approx \tilde{O}\left(\frac{d' \Pr(R)}{\epsilon}\right) \) labels

\( R \) is a region in the input space.
Key Observation 2

\( R = \) disagreement region of current confidence set

Need to learn difference classifier with FN rate
\[ \leq \frac{\epsilon}{\Pr(R)} \] over \( R \)

Need \( \approx \tilde{O} \left( \frac{d' \Pr(R)}{\epsilon} \right) \) labels

Problem: \( R \) keeps changing…. 
Full Algorithm

\( H = \) difference concept class, \( d' = \text{VCdim}(H) \), \( \theta = \) disagreement coeff

For epochs 1, 2, 3, ....

Target excess error in epoch \( k \) \( \epsilon_k \approx \frac{1}{2^k} \)
Full Algorithm

H = difference concept class, d’ = VCdim(H), θ = disagreement coeff

For epochs 1, 2, 3, ….  

Target excess error in epoch k  \( \epsilon_k \approx 1/2^k \)

Draw  \( \tilde{O}(d' \theta (\nu^* + \epsilon_k) / \epsilon_k) \) samples \( x_1, \ldots, x_m \). For each \( x_i \), query O and W. Set:  \( y_{i,D} = 1 \) if  \( y_{i,O} \neq y_{i,W} \)
H = difference concept class, \( d' = \text{VCdim}(H) \), \( \theta = \text{disagreement coeff} \)

For epochs 1, 2, 3, ....

Target excess error in epoch \( k \) \( \epsilon_k \approx 1/2^k \)

Draw \( \tilde{O}(d' \theta(\nu^* + \epsilon_k) / \epsilon_k) \) samples \( x_1, .., x_m \). For each \( x_i \), query \( O \) and \( W \). Set: \( y_{i,D} = 1 \) if \( y_{i,O} \neq y_{i,W} \)

Train difference classifier \( h \) on \( \{(x_i, y_{i,D})\} \) with FN rate

\( \approx \epsilon_k / \theta(\nu^* + \epsilon_k) \) over the current disagreement region
Full Algorithm

H = difference concept class, \( d' = \text{VCdim}(H) \), \( \theta = \text{disagreement coeff} \)

For epochs 1, 2, 3, ....

Target excess error in epoch \( k \) \( \epsilon_k \approx 1/2^k \)

Draw \( \tilde{O}(d' \theta (\nu^* + \epsilon_k)/\epsilon_k) \) samples \( x_1, \ldots, x_m \). For each \( x_i \), query O and W. Set: \( y_{i,D} = 1 \) if \( y_{i,O} \neq y_{i,W} \)

Train difference classifier \( h \) on \( \{(x_i, y_{i,D})\} \) with FN rate

\( \approx \epsilon_k / \theta (\nu^* + \epsilon_k) \) over the current disagreement region

Run disagreement based active learning algorithm A to target excess error \( \epsilon_k \). If A queries the label of \( x \) then:

if \( h(x) = 1 \), query O, else query W
Label Complexity

Total #labels to train difference classifier \( \approx \tilde{O} \left( \frac{d' \theta (\nu^* + \epsilon)}{\epsilon} \right) \)

How many labels for the rest of active learning?
Label Complexity: Assumptions

For any $r, t$, there is a $h$ in $H$ such that:

$$\Pr(h(x) = -1, x \in DIS(B(c^*, r), y_O \neq y_W) \leq t$$

(Low FN over disagreement region)

$$\Pr(h(x) = 1, x \in DIS(B(c^*, r)) \leq \alpha(r, t)$$

(Low positives)

Note: $\alpha(r, t) \leq \Pr(DIS(B(c^*, r)))$
Label Complexity

#labels to train difference classifier \( \approx \tilde{O}\left(\frac{d'\theta (\nu^* + \epsilon)}{\epsilon}\right) \)

#labels for active learning \( \approx \tilde{O}\left(\frac{d\sigma (\nu^*)^2}{\epsilon^2}\right) \)

where: \( \sigma \approx \frac{\alpha(2\nu^* + \epsilon, O(\epsilon))}{2\nu^* + \epsilon} \leq \theta \)

Compare:

#labels for disagreement based active learning: \( \approx \tilde{O}\left(\frac{d\theta (\nu^*)^2}{\epsilon^2}\right) \)
Conclusion

Auxiliary oracle may help if there is a good difference classifier in the difference concept class.

Open Question:
What if there is no gold standard?
Thank You!