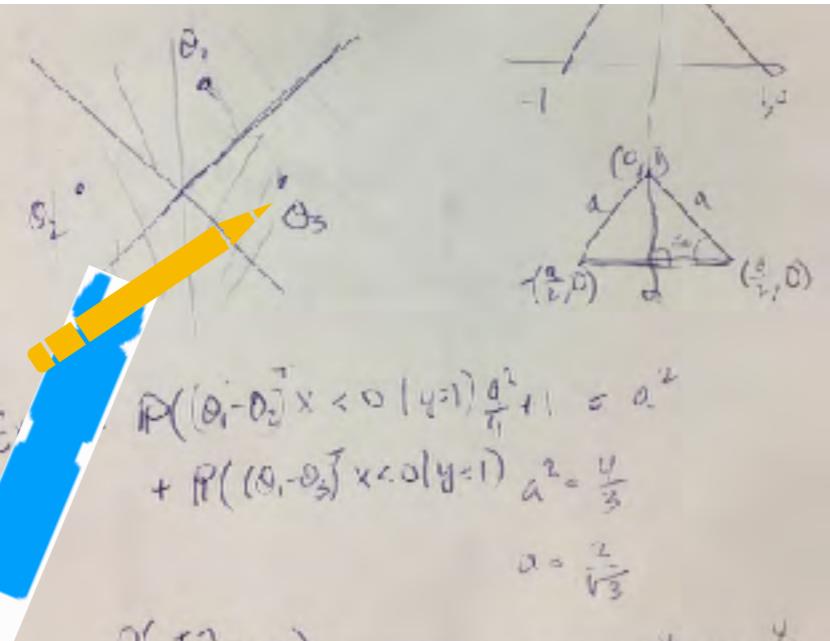
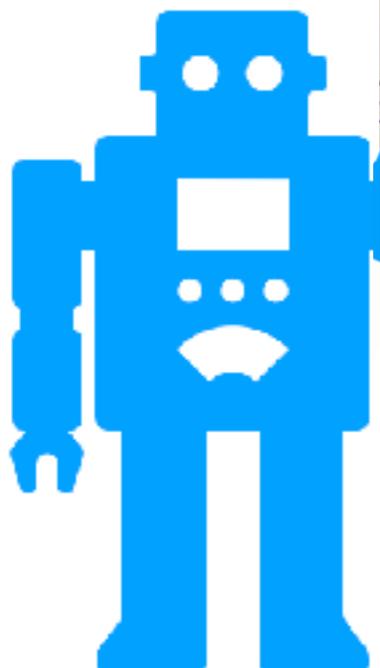


Active Learning from Theory to Practice



Steve Hanneke
Toyota Technological
Institute at Chicago
steve.hanneke@gmail.com

Robert Nowak
UW-Madison
rdnowak@wisc.edu

ICML | 2019

Thirty-sixth International Conference on
Machine Learning

Tutorial Outline



Part 1: Introduction to Active Learning (Rob)

第1部分:主动学习简介(Rob)
第二部分:主动学习理论(Steve)
第3部分:
高级主题和开放问题(Steve)
Part
4:非参数主动学习(Rob)

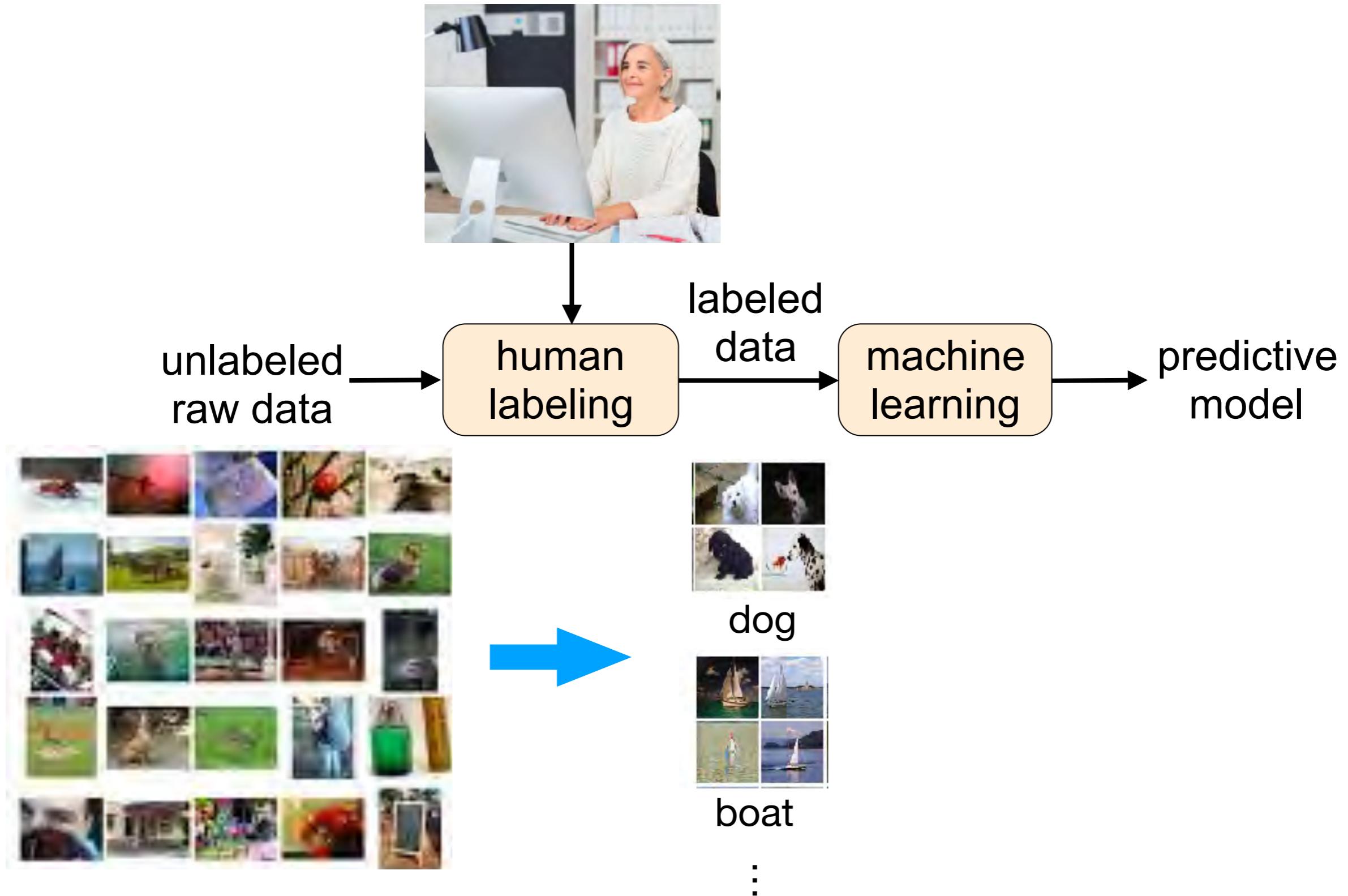
Part 2: Theory of Active Learning (Steve)

Part 3: Advanced Topics and Open Problems (Steve)

Part 4: Nonparametric Active Learning (Rob)

slides: <http://nowak.ece.wisc.edu/ActiveML.html>

Conventional (Passive) Machine Learning



ALL SYSTEMS GO

?

the guardian

Computers now better than humans at
recognising and sorting images

millions of labeled images
1000's of human hours

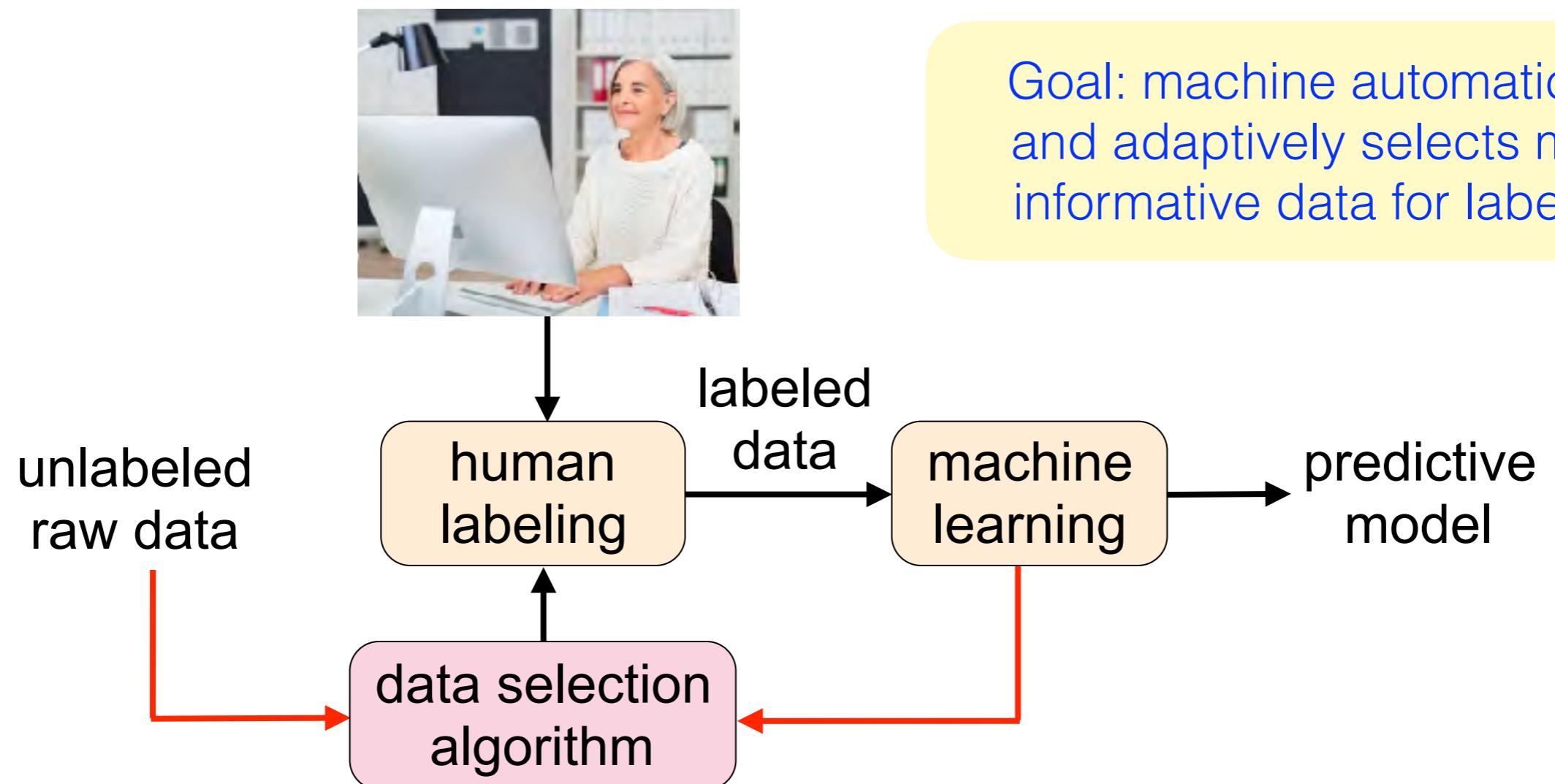
QUARTZ

**Google says its new AI-powered
translation tool scores nearly identically to
human translators**

trained on more texts than a
human could read in a lifetime

Can we train machines with less labeled
data and less human supervision?

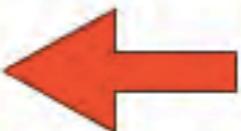
Active Machine Learning



Motivating Application



unlabeled electronic
health records (EHRs)



prediction rule
that can be applied
to unlabeled EHRs



human experts



cataracts

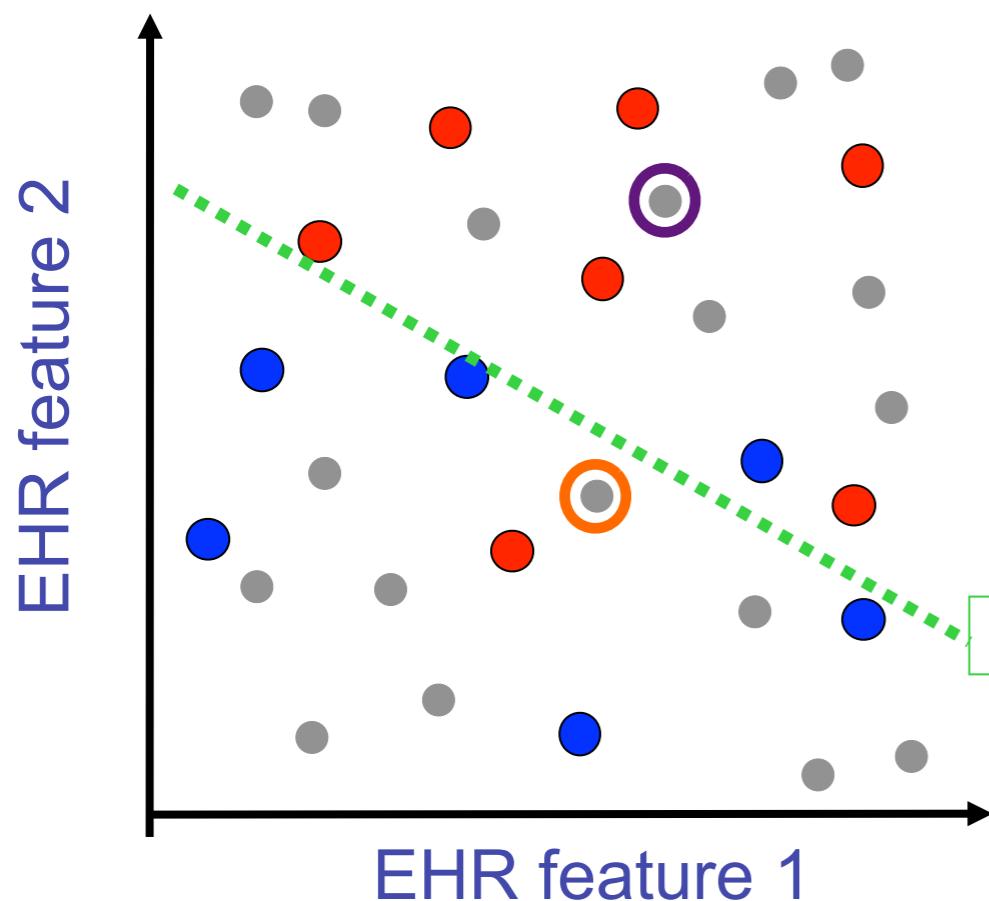


healthy



provides labels to machine learner
(several minutes / EHR)

Active Learning

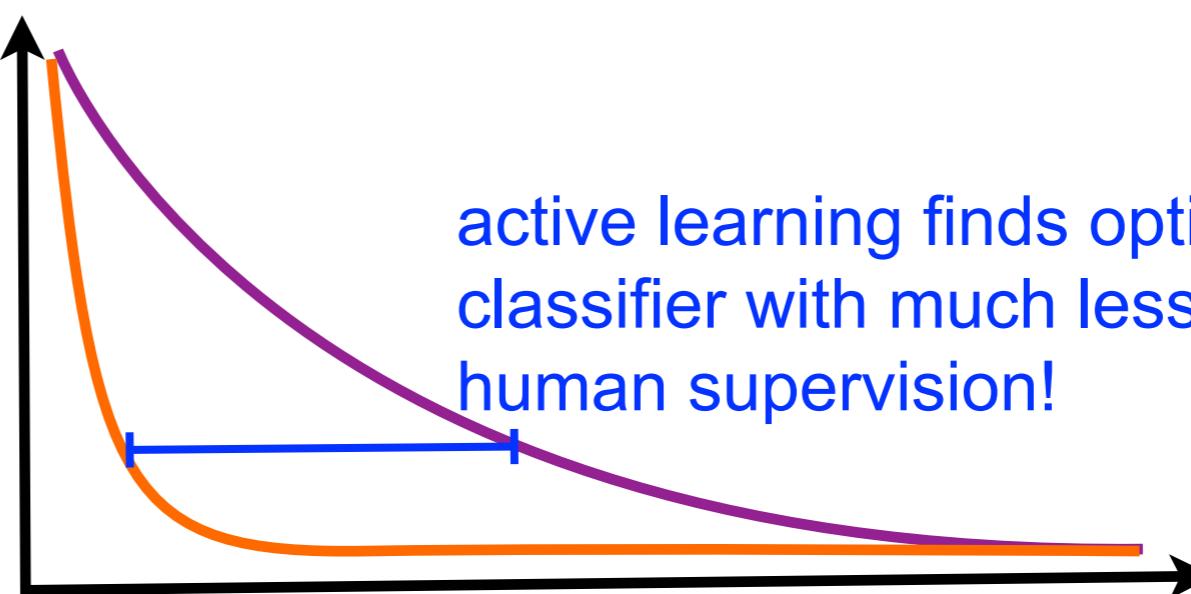


Non-adaptive strategy: Label a random sample

Active strategy: Label a sample near best decision boundary based on labels seen so far

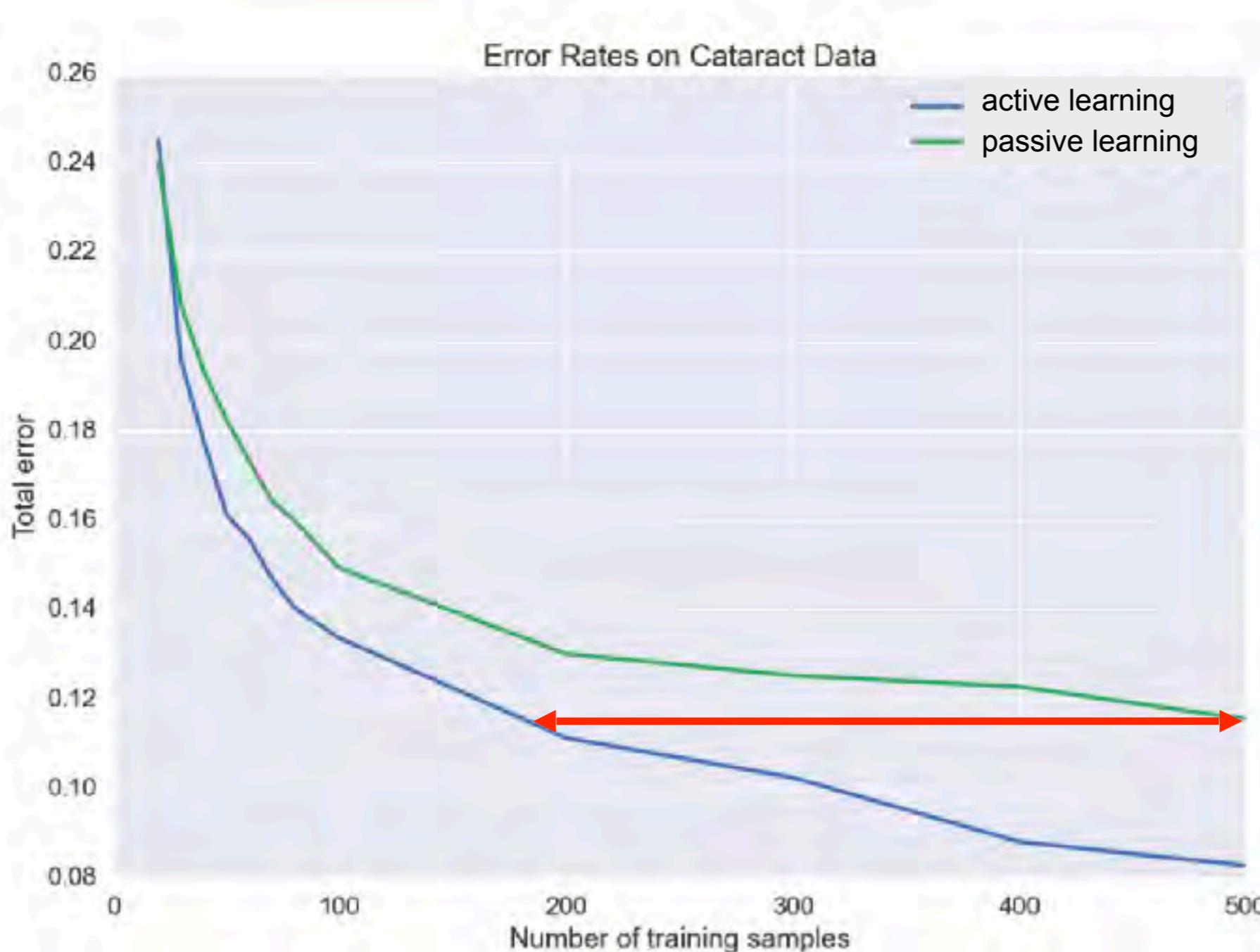
best linear classifier

error rate ϵ



active learning finds optimal classifier with much less human supervision!

Active Logistic Regression



11000 patient records
8000 positive
3000 negative

6182 Numerical Features
icd9 codes
lab tests
patient data

Classification task:
cataracts or healthy

less than half as many labeled examples needed by active learning

NEXT
ASK BETTER QUESTIONS.
GET BETTER RESULTS.
FASTER. AUTOMATED.



GitHub



Paper



Docs



Blog



Team



Data

Active learning to optimize crowdsourcing and rating in New Yorker Cartoon Caption Contest



digg

BY DOING THE EXACT OPPOSITE

How New Yorker Cartoons Could Teach Computers To Be Funny

3 diggs CNET Technology



With the help of computer scientists from the University of Wisconsin at Madison, The New Yorker for the first time is using crowdsourcing algorithms to uncover the best captions.



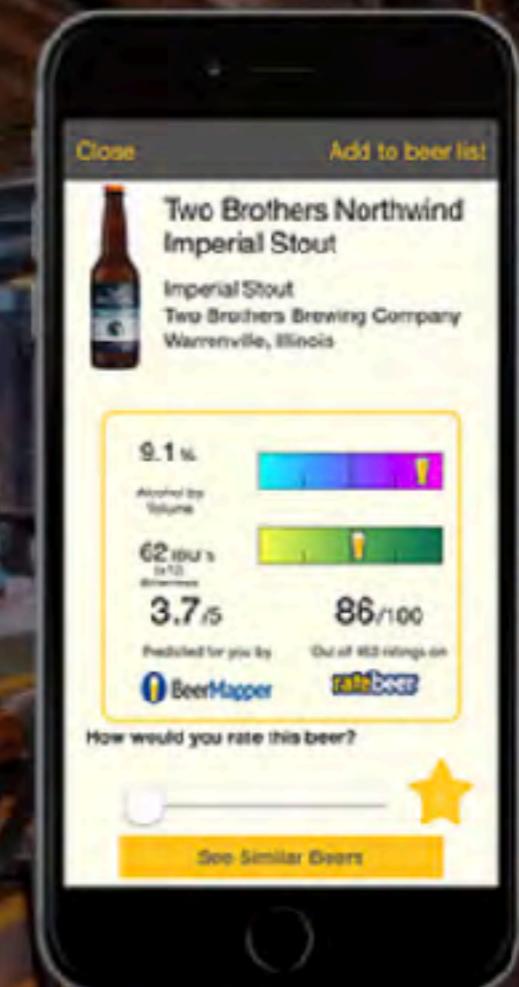
Actively learning user's beer preferences



BeerMapperSM

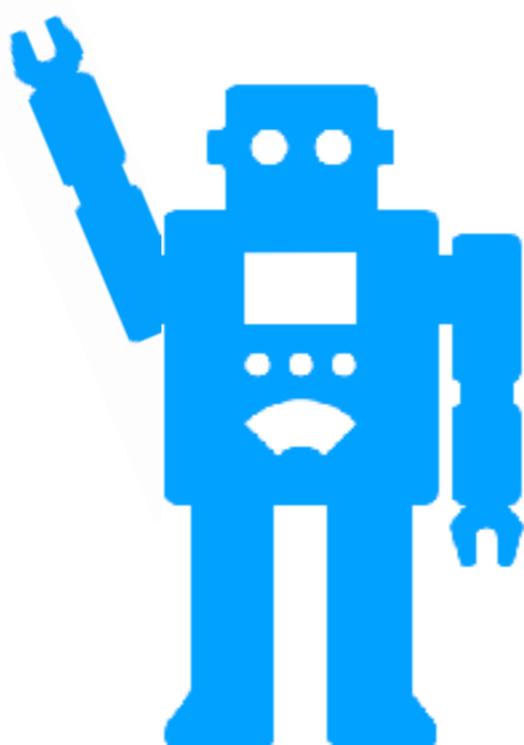
[Home](#)[Contact](#)[About](#)[FAQs](#)

Discover better beer.



The most powerful beer app on the planet.

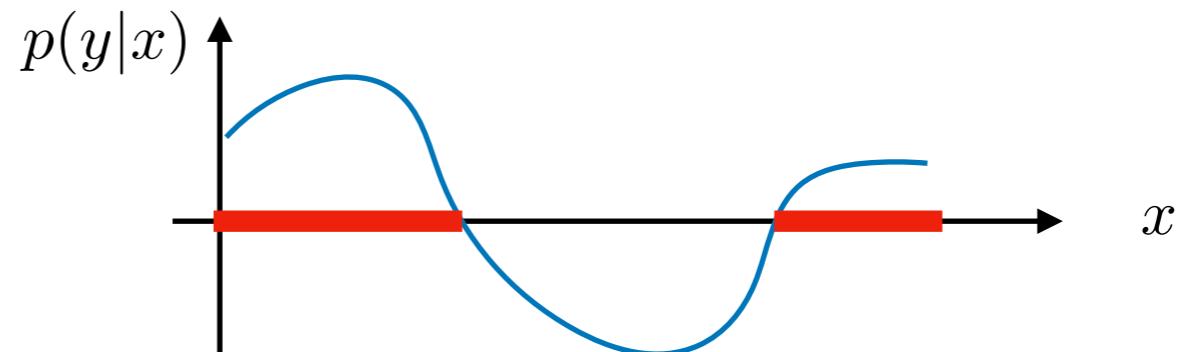
Principles of Active Learning



What and Where Information

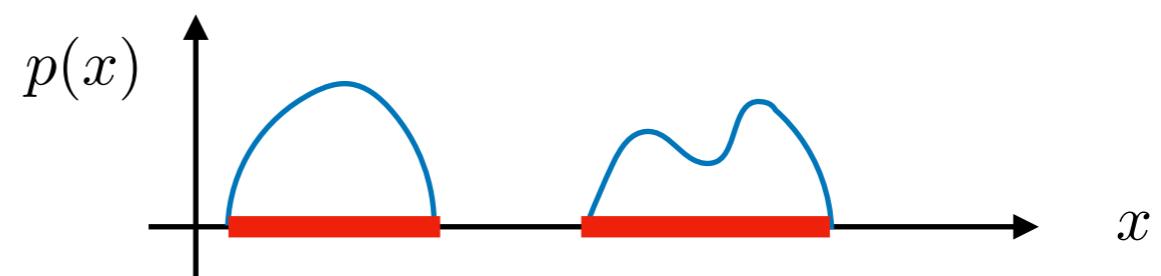
Density estimation: What is $p(y|x)$?

Classification: Where is $p(y|x) > 0$?



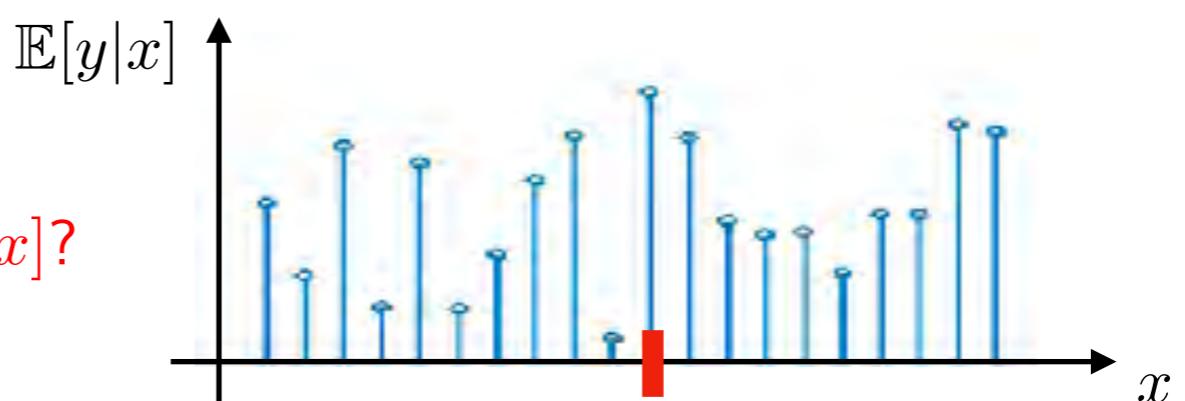
Density estimation: What is $p(x)$?

Clustering: Where is $p(x) > \epsilon$?



Function estimation: What is $\mathbb{E}[y|x]$?

Bandit optimization: Where is $\max_x \mathbb{E}[y|x]$?



Active learning is more efficient than passive learning for localized “where” information

Meta-Algorithm for Active Learning

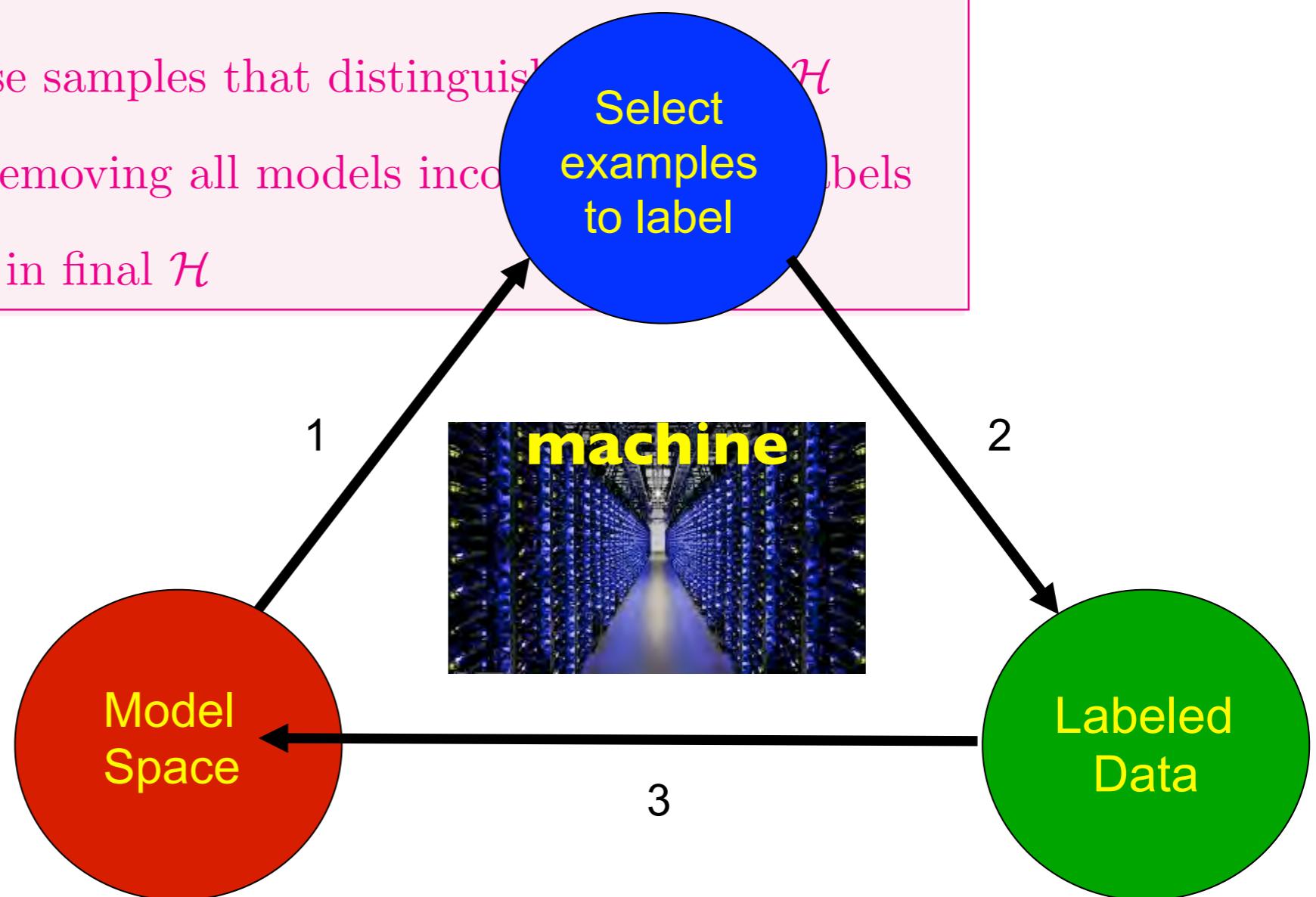
Version-Space (VS) Active Learning

initialize VS: \mathcal{H} = all models/hypotheses

while (*stopping-criterion*) not met

1. sample at random from available dataset
2. label only those samples that distinguish \mathcal{H}
3. reduce \mathcal{H} by removing all models inconsistent with labels

output: best model in final \mathcal{H}



Learning a 1-D Classifier



binary search quickly finds **decision boundary**

passive : err $\sim n^{-1}$

active : err $\sim 2^{-n}$

Vapnik-Chervonenkis (VC) Theory

Given training data $\{(x_j, y_j)\}_{j=1}^n$, learn a function f to predict y from x

Consider a possibly infinite set of hypotheses \mathcal{F} with *finite VC dimension d* and for each $f \in \mathcal{F}$ define the risk (error rate):

$$R(f) := \mathbb{P}(f(x) \neq y)$$

error rate on
training data:

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(f(x_i) \neq y_i) \quad \begin{array}{l} \text{“empirical risk”} \\ \text{“经验风险”} \end{array}$$

VC bound: $\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}(f)| \leq 6\sqrt{\frac{d \log(n/\delta)}{n}}$

w.p. $\geq 1 - \delta$

Empirical Risk Minimization (ERM)

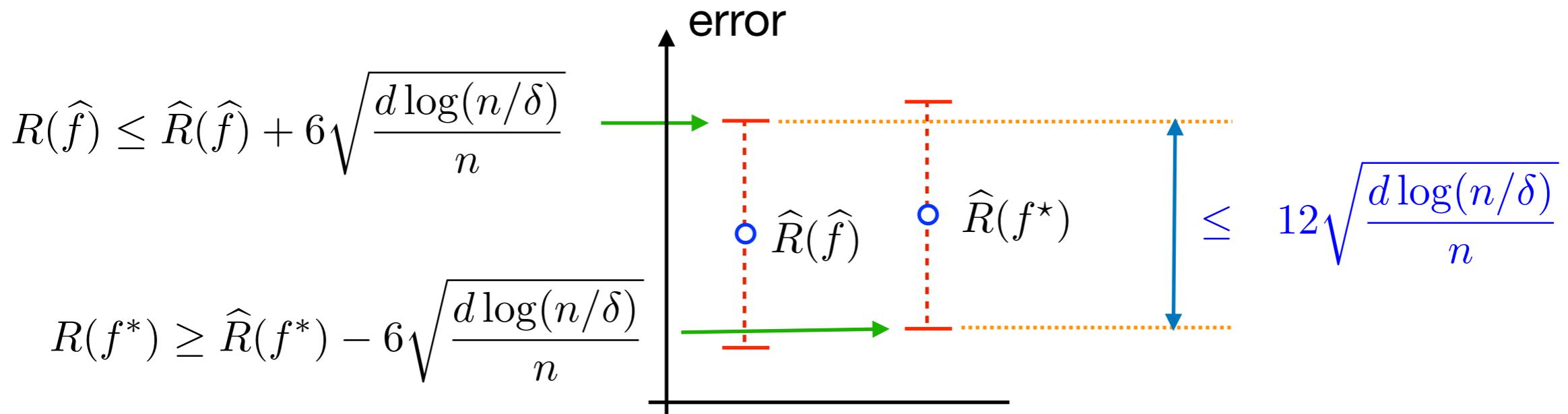
Goal: select hypothesis with true error rate within $\epsilon > 0$ of $\min_{f \in \mathcal{F}} R(f)$

$$f^* = \arg \min_{f \in \mathcal{F}} R(f) \quad \text{true risk minimizer}$$

\hat{f} minimizes empirical risk:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \hat{R}(f) \quad \text{empirical risk minimizer}$$

$$\hat{R}(\hat{f}) \leq \hat{R}(f^*)$$

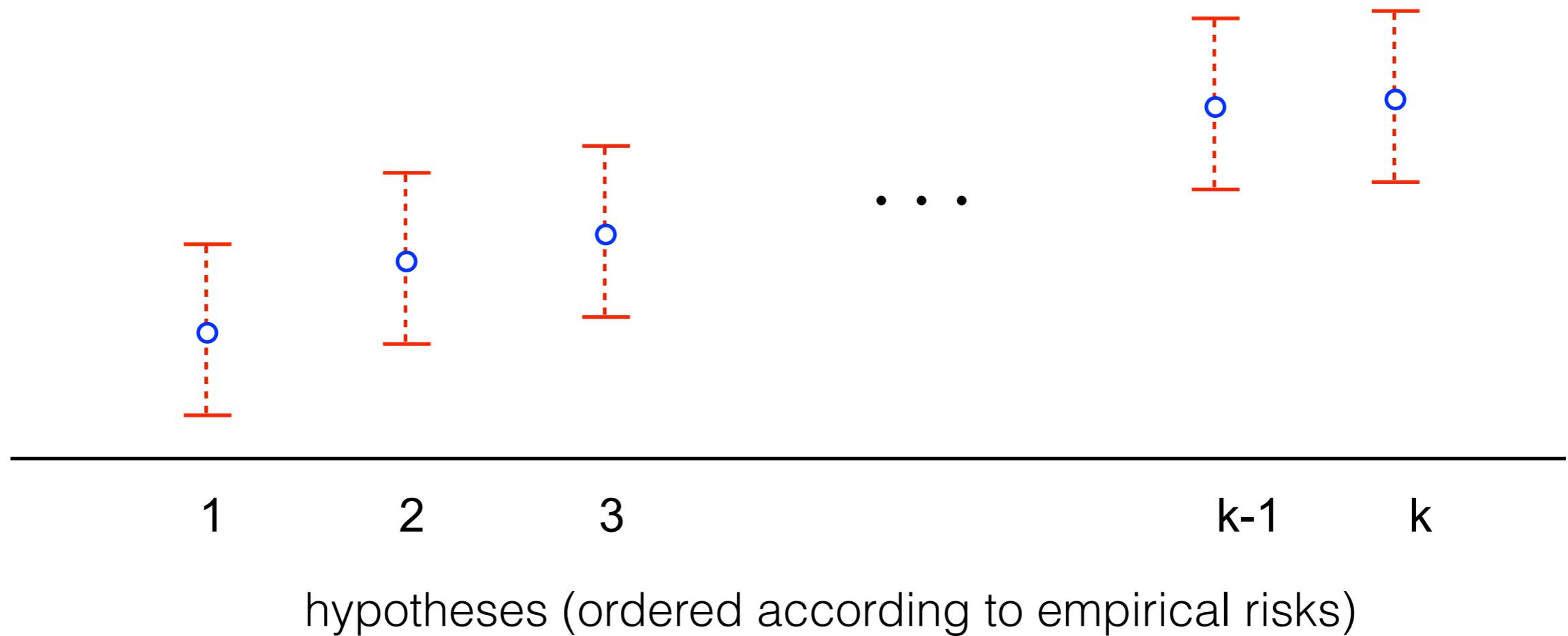


sufficient number
of training examples:

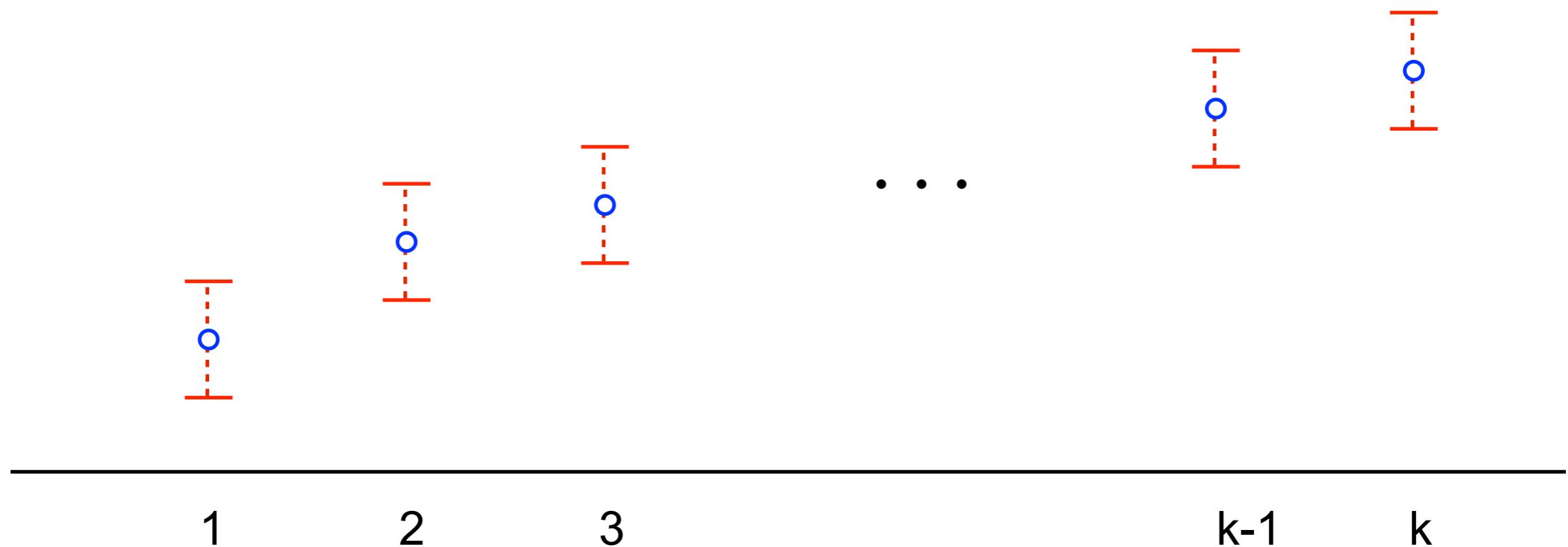
$$12\sqrt{\frac{d \log(n/\delta)}{n}} \leq \epsilon \quad \rightarrow$$

$$n = \tilde{O}\left(\frac{d \log(1/\delta)}{\epsilon^2}\right)$$

Empirical Risks and Confidence Intervals



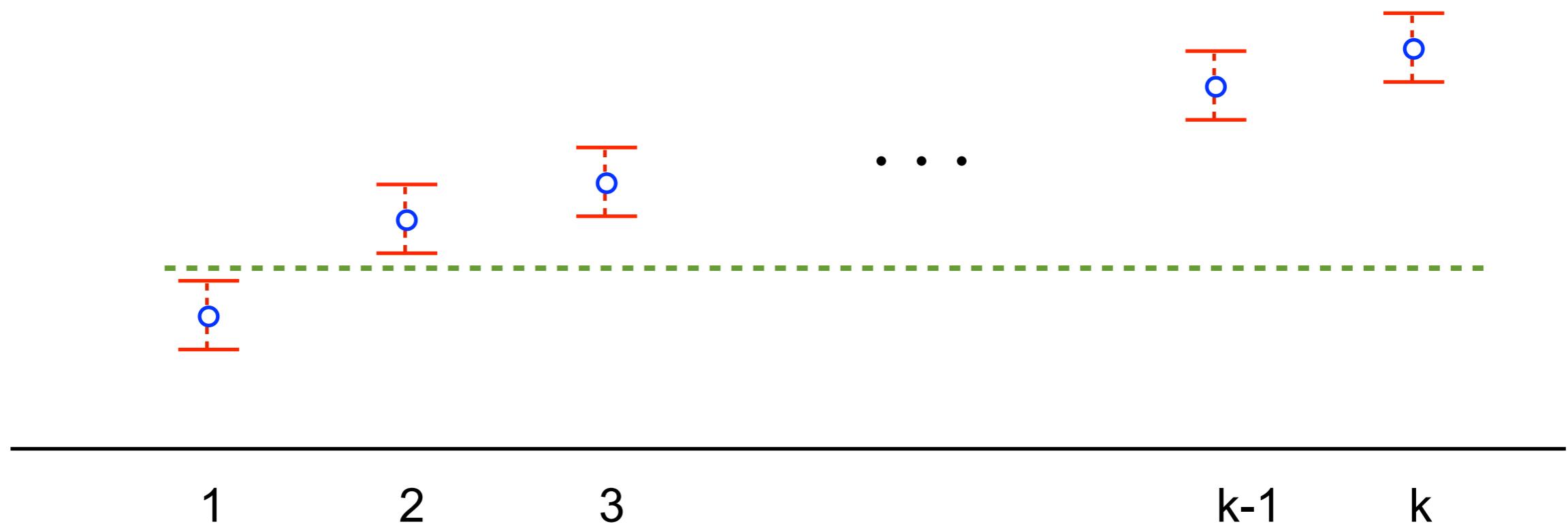
Empirical Risks and Confidence Intervals



hypotheses (ordered according to empirical risks)

more training data \Rightarrow smaller confidence intervals

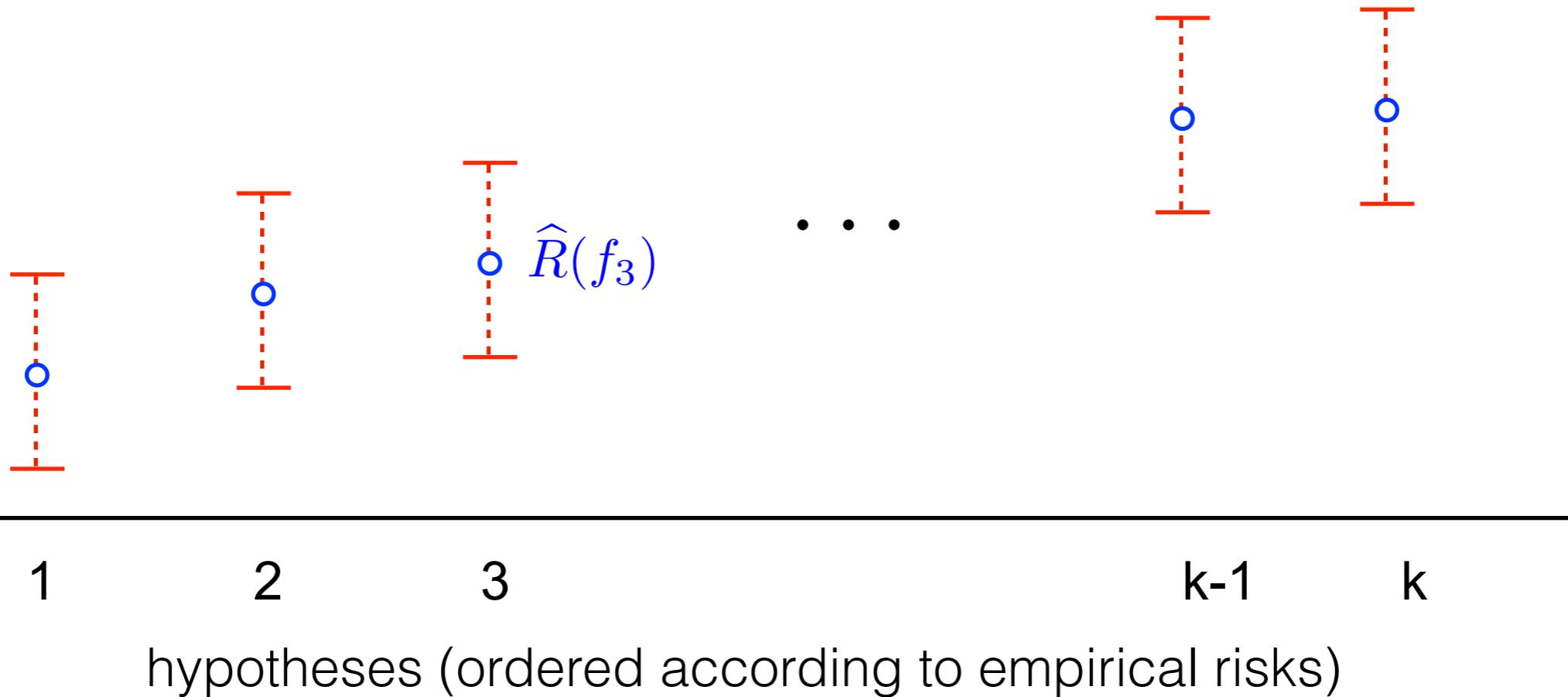
Empirical Risks and Confidence Intervals



hypotheses (ordered according to empirical risks)

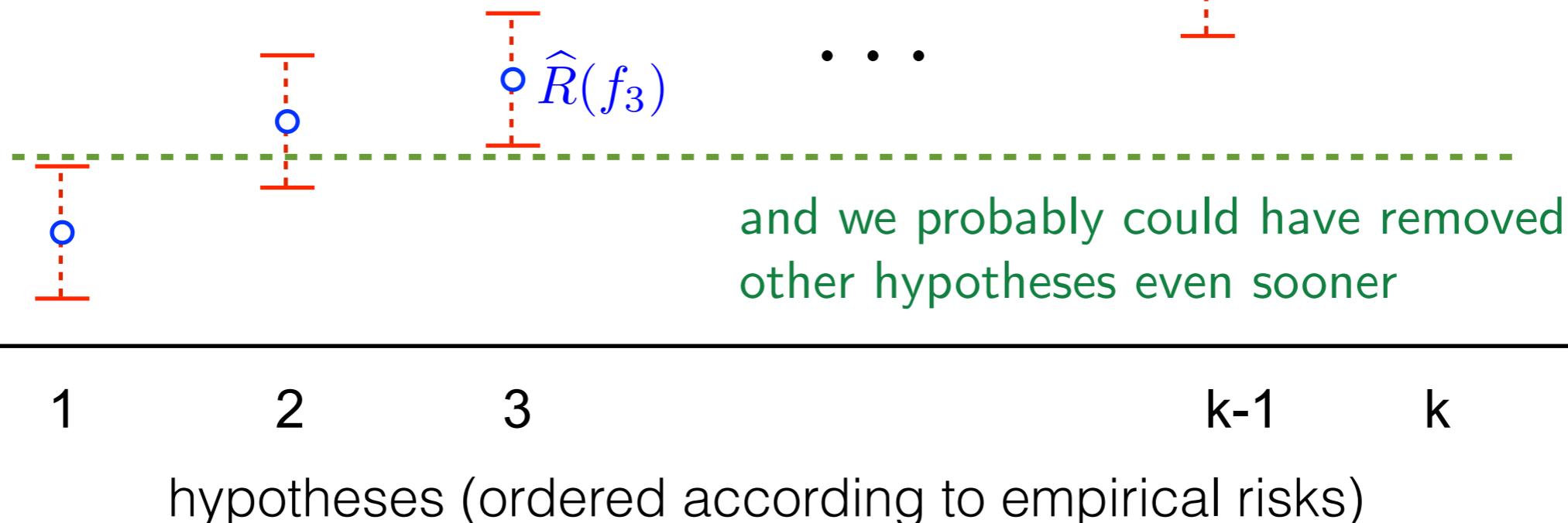
more training data \Rightarrow smaller confidence intervals

ERM is Wasting Labeled Examples

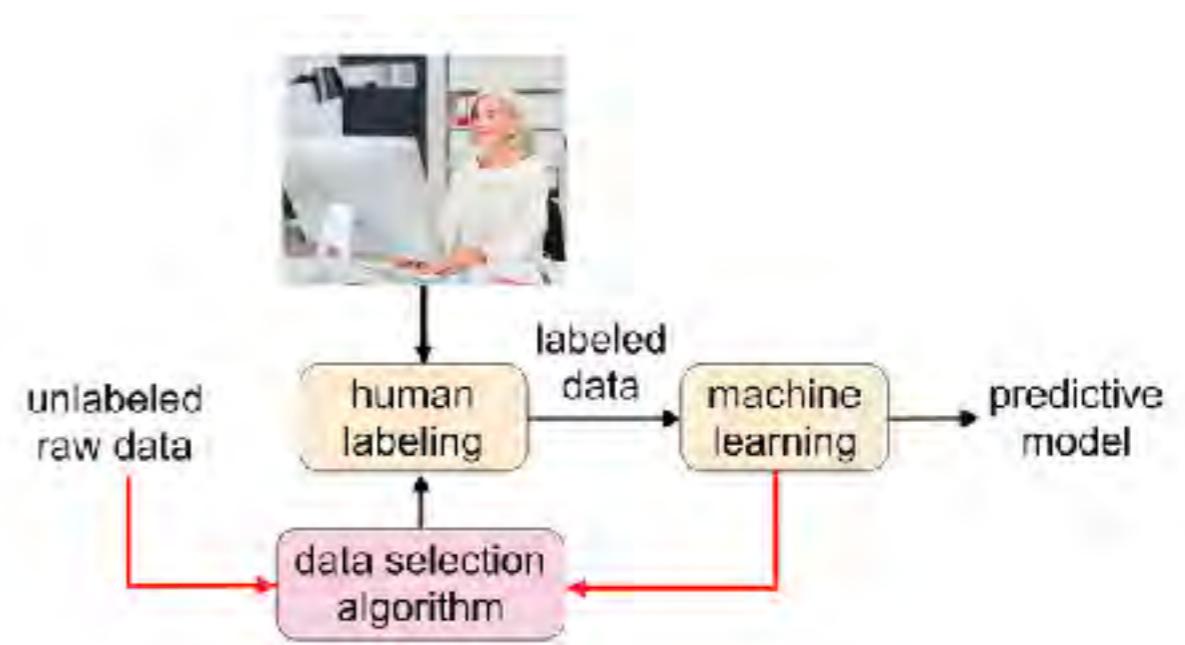


ERM is Wasting Labeled Examples

at this point we can safely remove
 f_3 from further consideration

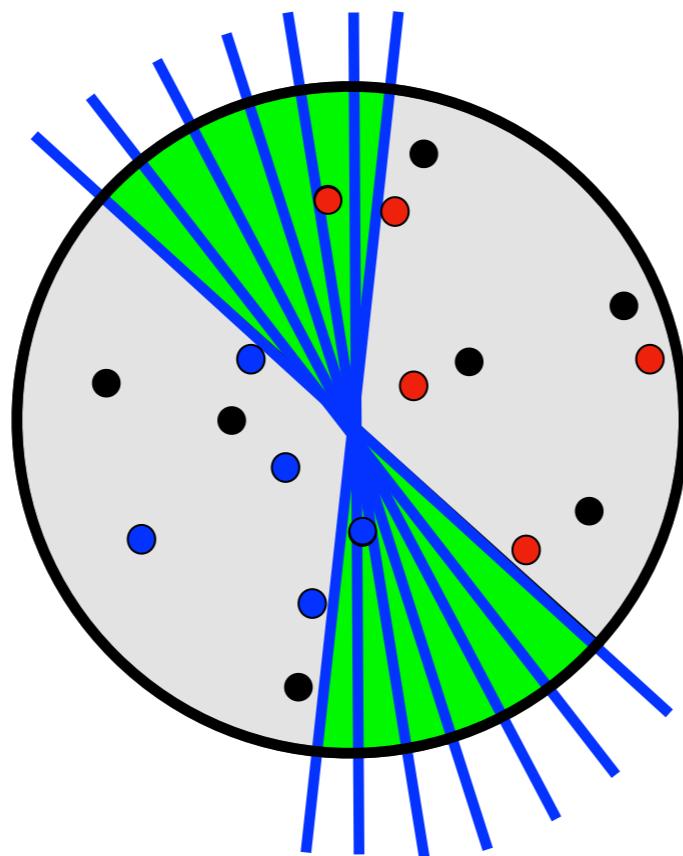


only require labels for examples that hypotheses 1 and 2 label differently
(i.e., examples where they *disagree*)



Disagreement-Based Active Learning

consider points uniform on unit ball and linear classifiers passing through origin



only label points in the
region of disagreement \mathcal{D}

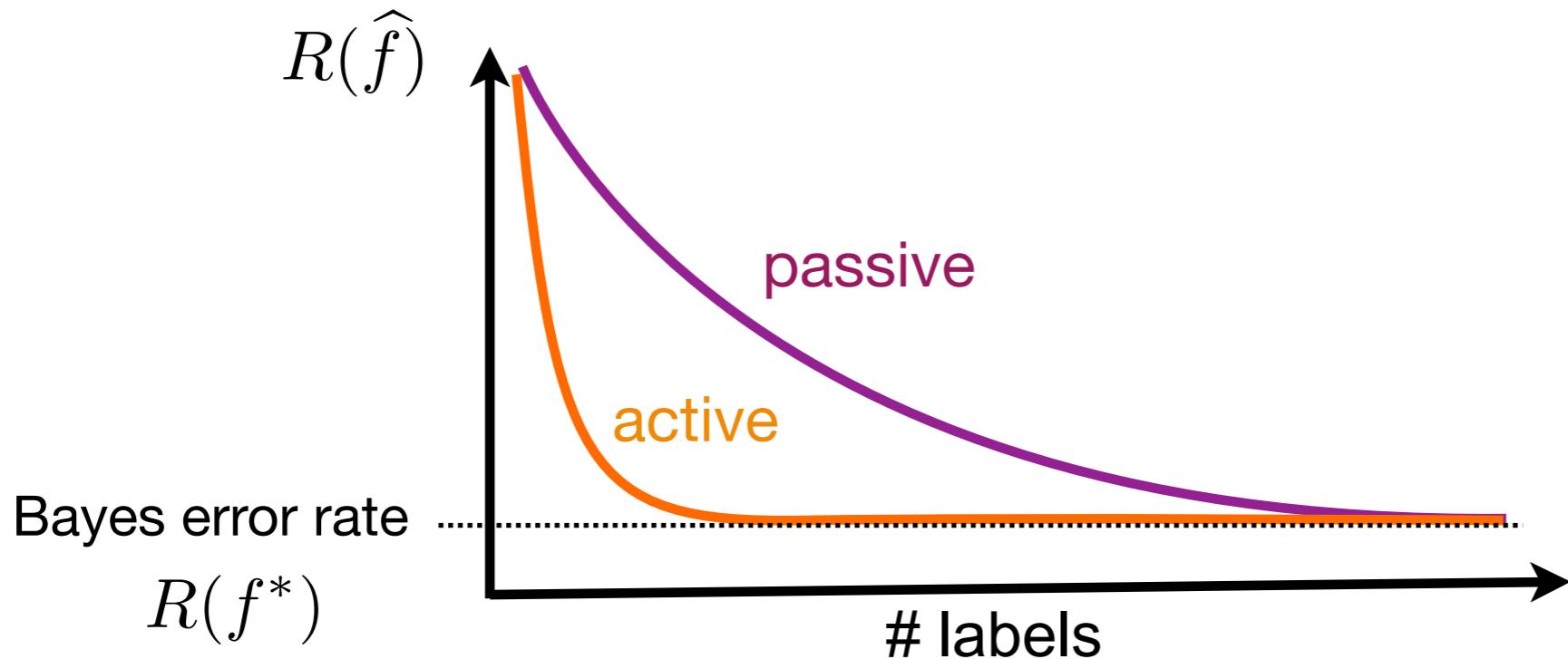
Active Binary Classification

Assuming optimal Bayes classifier f^* in VC class with dimension d and “nice” distributions (e.g., bounded label noise)

$$\epsilon = R(\hat{f}) - R(f^*)$$

passive $\epsilon \sim \frac{d}{n}$ parametric rate

active $\epsilon \sim \exp\left(-c \frac{n}{d}\right)$ exponential speed-up



Tutorial Outline

Part 1: Introduction to Active Learning (Rob)

Part 2: Theory of Active Learning (Steve)

Part 3: Advanced Topics and Open Problems (Steve)

Part 4: Nonparametric Active Learning (Rob)

slides: <http://nowak.ece.wisc.edu/ActiveML.html>

Recommended Reading (Foundations of Active Learning)

Settles, Burr. "Active learning." *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6.1 (2012): 1-114.

Dasgupta, Sanjoy. "Two faces of active learning." *Theoretical computer science* 412.19 (2011): 1767-1781.

Cohn, David, Les Atlas, and Richard Ladner. "Improving generalization with active learning." *Machine learning* 15.2 (1994): 201-221.

Castro, Rui M., and Robert D. Nowak. "Minimax bounds for active learning." *IEEE Transactions on Information Theory* 54, no. 5 (2008): 2339-2353.

Zhu, Xiaojin, John Lafferty, and Zoubin Ghahramani. "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions." *ICML 2003 workshop*. Vol. 3. 2003.

Dasgupta, Sanjoy, Daniel J. Hsu, and Claire Monteleoni. "A general agnostic active learning algorithm." *Advances in neural information processing systems*. 2008.

Balcan, Maria-Florina, Alina Beygelzimer, and John Langford. "Agnostic active learning." *Journal of Computer and System Sciences* 75.1 (2009): 78-89.

Nowak, Robert D. "The geometry of generalized binary search." *IEEE Transactions on Information Theory* 57, no. 12 (2011): 7893-7906.

Hanneke, Steve. "Theory of active learning." *Foundations and Trends in Machine Learning* 7, no. 2-3 (2014).

Part 2: Theory of Active Learning General Case

- Disagreement-Based Agnostic Active Learning
- Disagreement Coefficient
- Sample Complexity Bounds

**Tutorial on Active Learning:
Theory to Practice**

Steve Hanneke

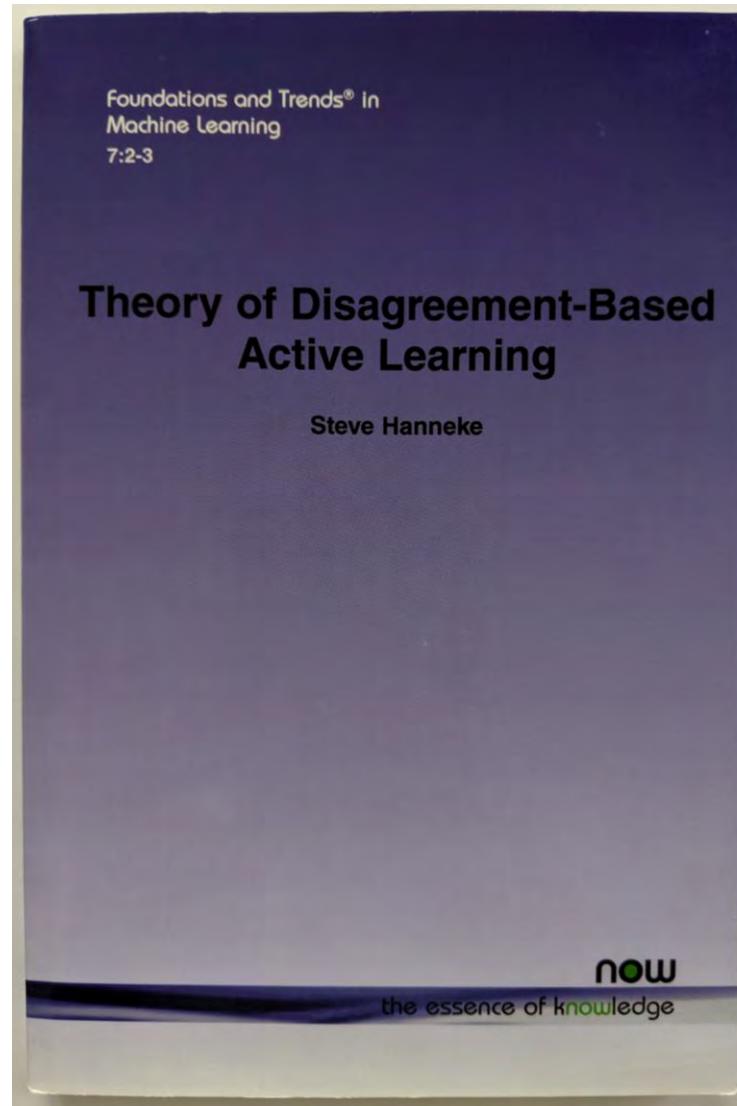
Toyota Technological Institute at Chicago
steve.hanneke@gmail.com

Robert Nowak

University of Wisconsin - Madison
rdnowak@wisc.edu

Agnostic Active Learning

不可知论者



Uniform Bernstein Inequality

Bernstein's inequality:

For m iid samples

$\forall f, f'$, w.p. $1 - \delta$,

$$R(f) - R(f') \leq \hat{R}(f) - \hat{R}(f') + c\sqrt{\hat{P}(f \neq f') \frac{\log(1/\delta)}{m}} + \frac{\log(1/\delta)}{m}$$

Uniform Bernstein inequality:

w.p. $1 - \delta$, $\forall f, f' \in \mathcal{H}$,

$$R(f) - R(f') \leq \hat{R}(f) - \hat{R}(f') + c\sqrt{\hat{P}(f \neq f') \frac{d \log(m/\delta)}{m}} + \frac{d \log(m/\delta)}{m}$$

VC dimension

Roughly:

$\forall f, f' \in \mathcal{H}$,

$$R(f) - R(f') \leq \hat{R}(f) - \hat{R}(f') + \sqrt{\hat{P}(f \neq f') \frac{d}{m}}$$

Agnostic Active Learning

Balcan, Beygelzimer, & Langford (2006)

Region of disagreement:

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. **reduce** \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}

Agnostic Active Learning

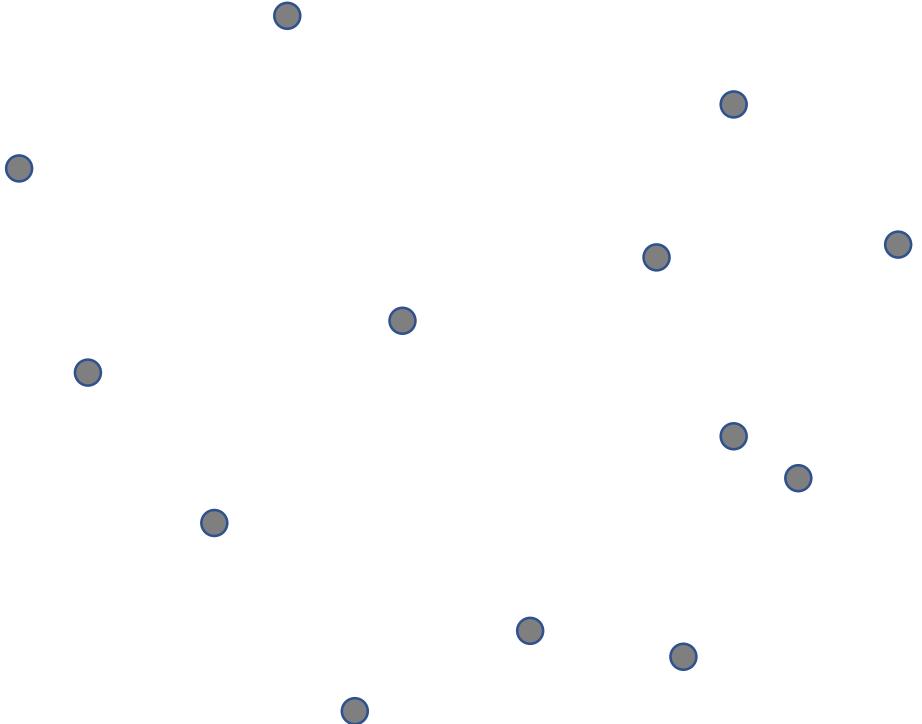
$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

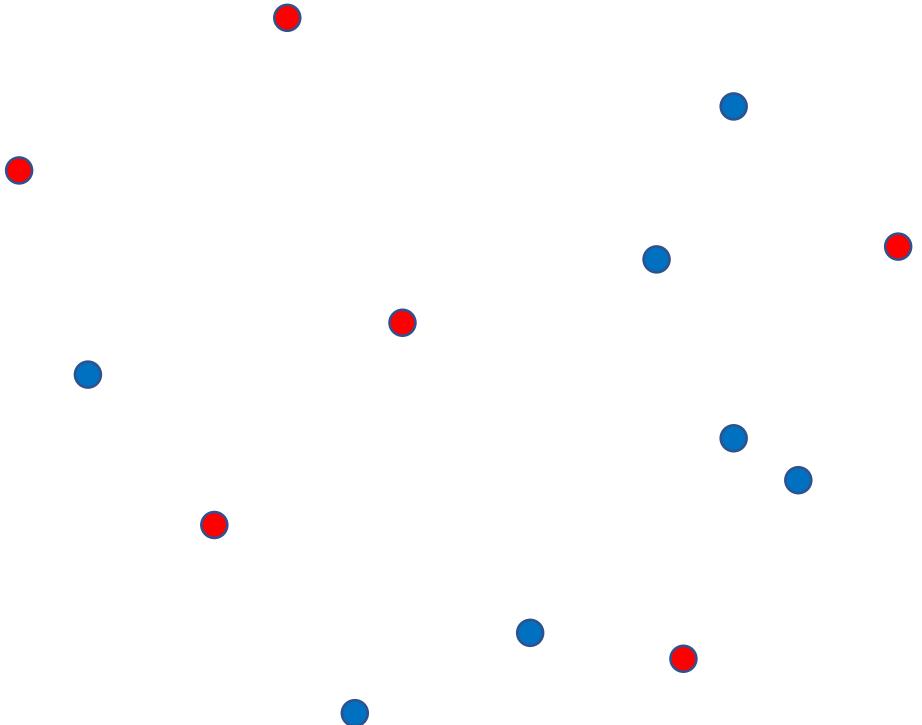
$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

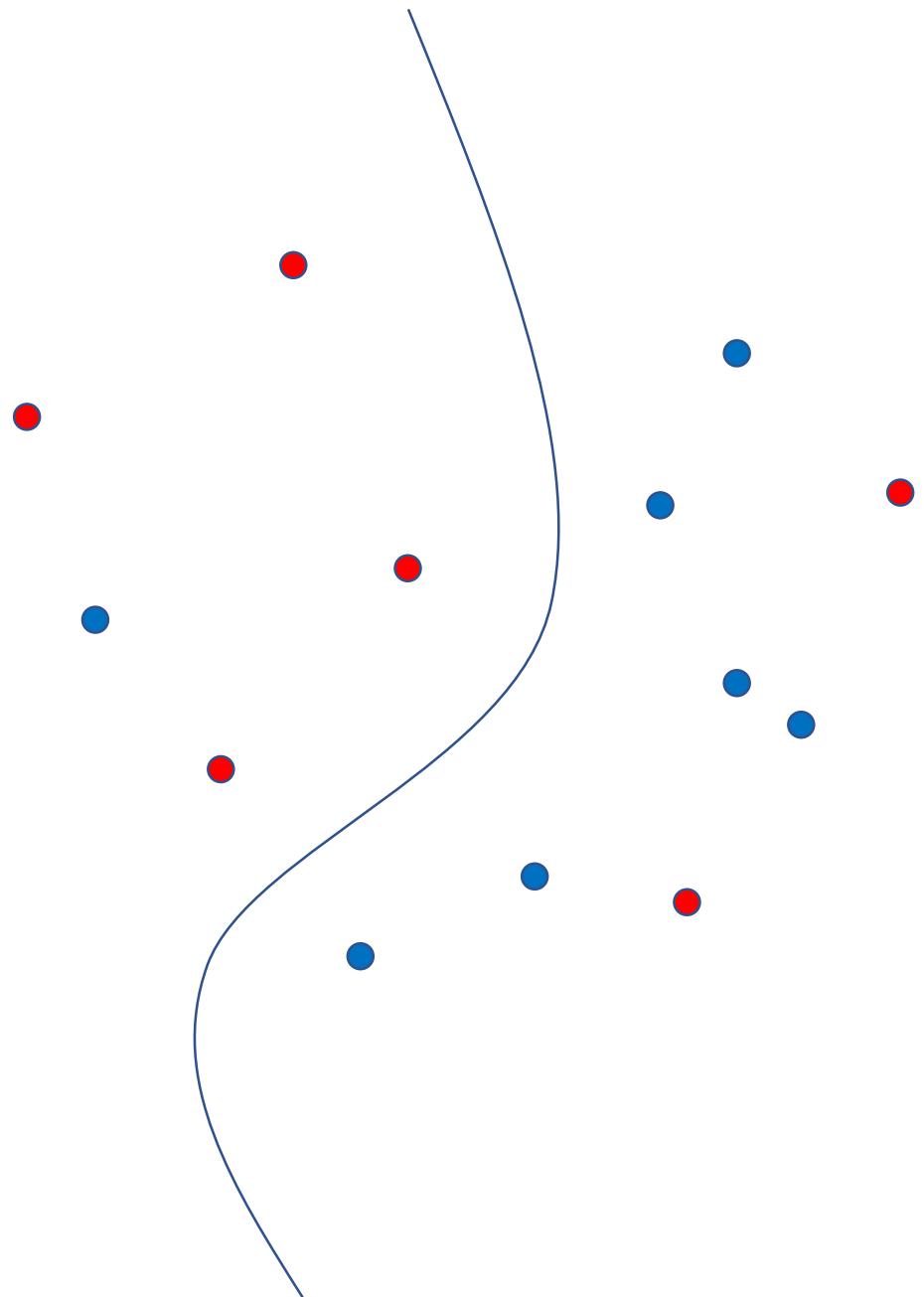
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

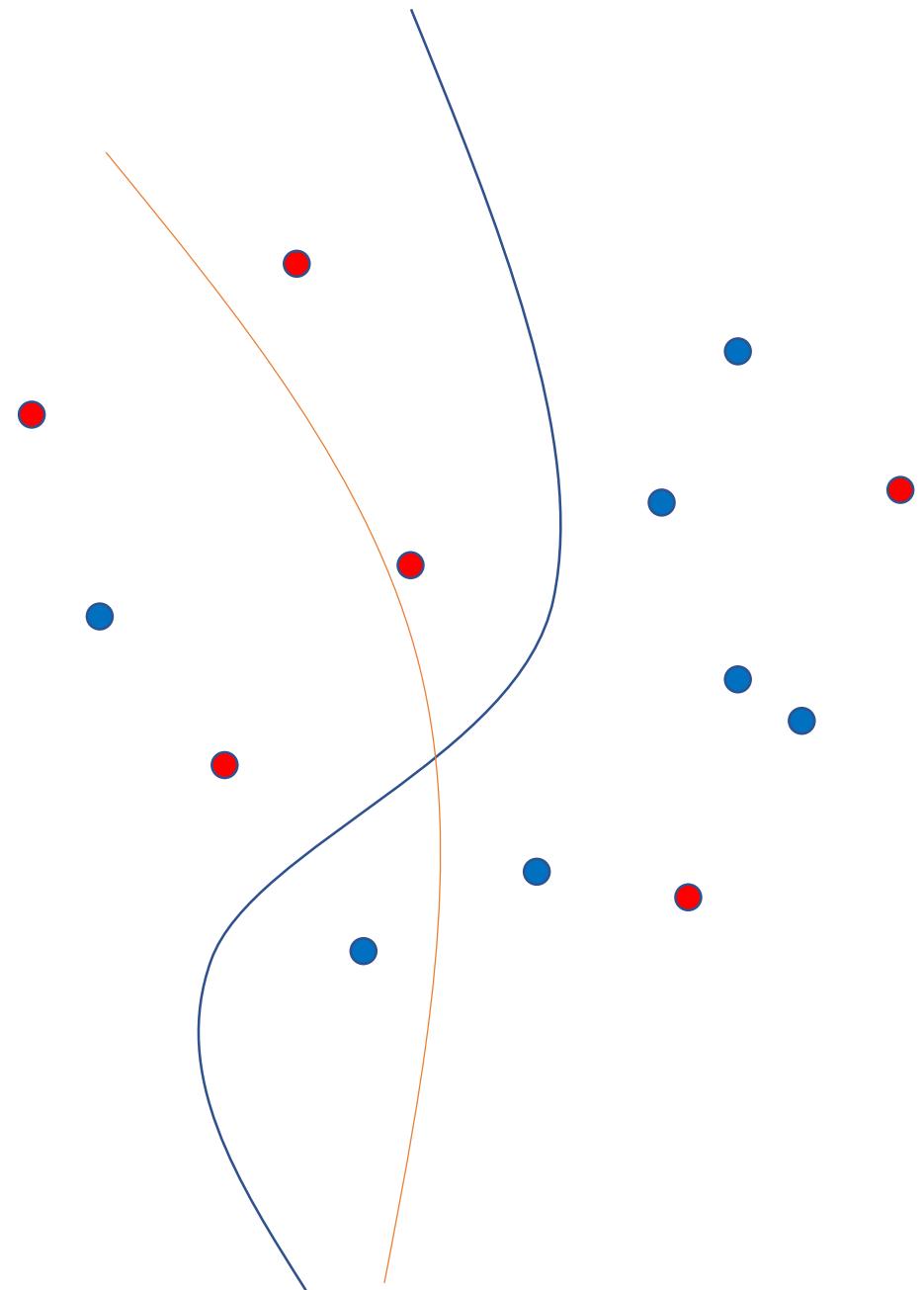
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

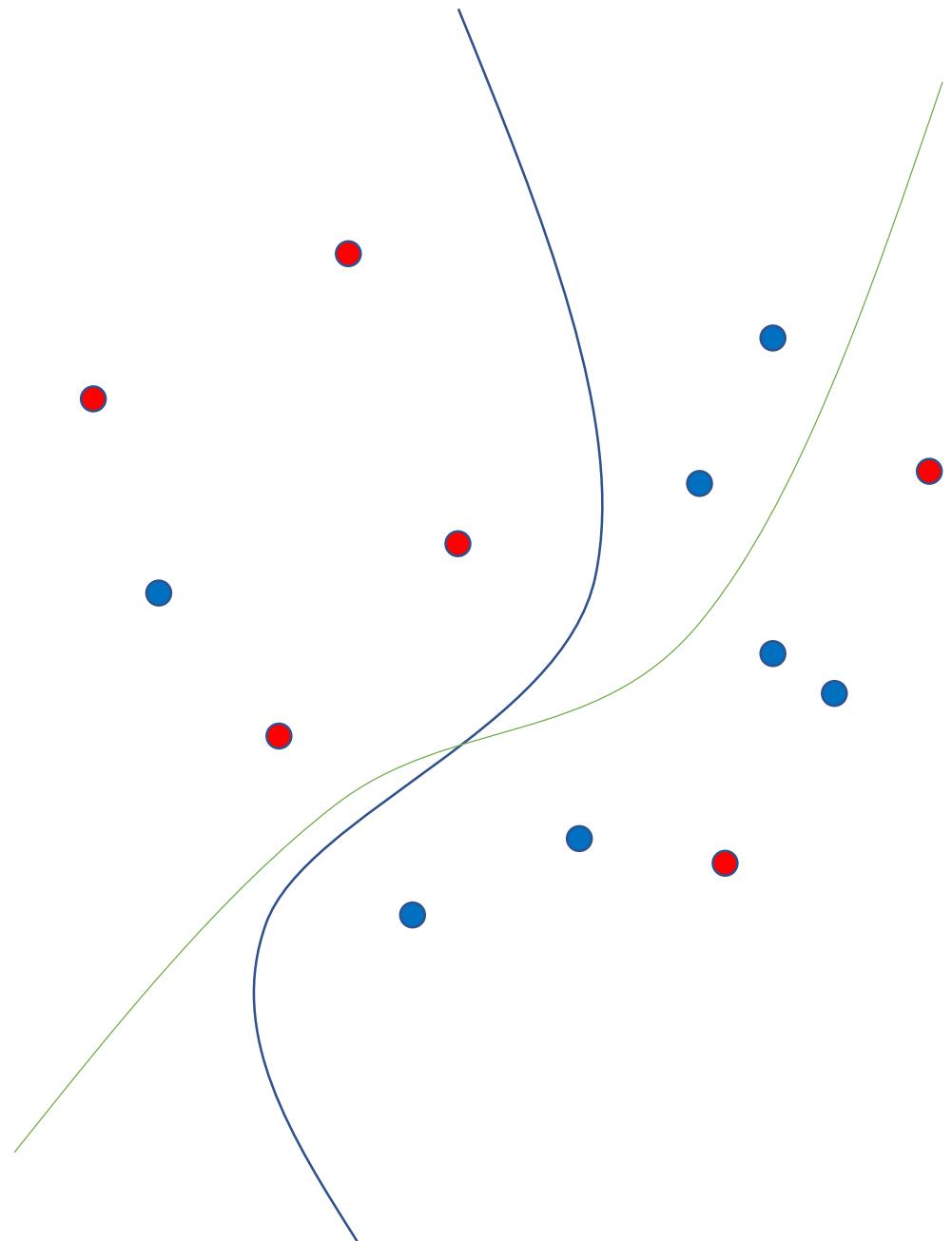
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

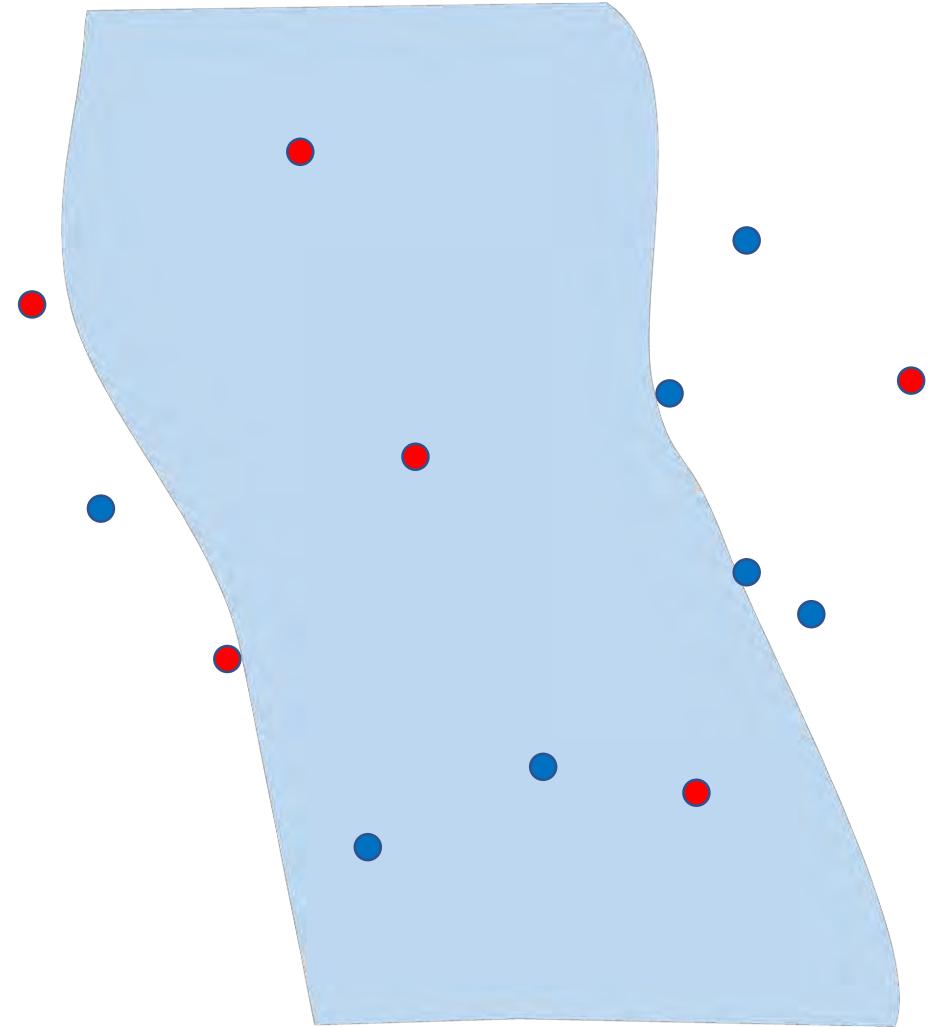
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

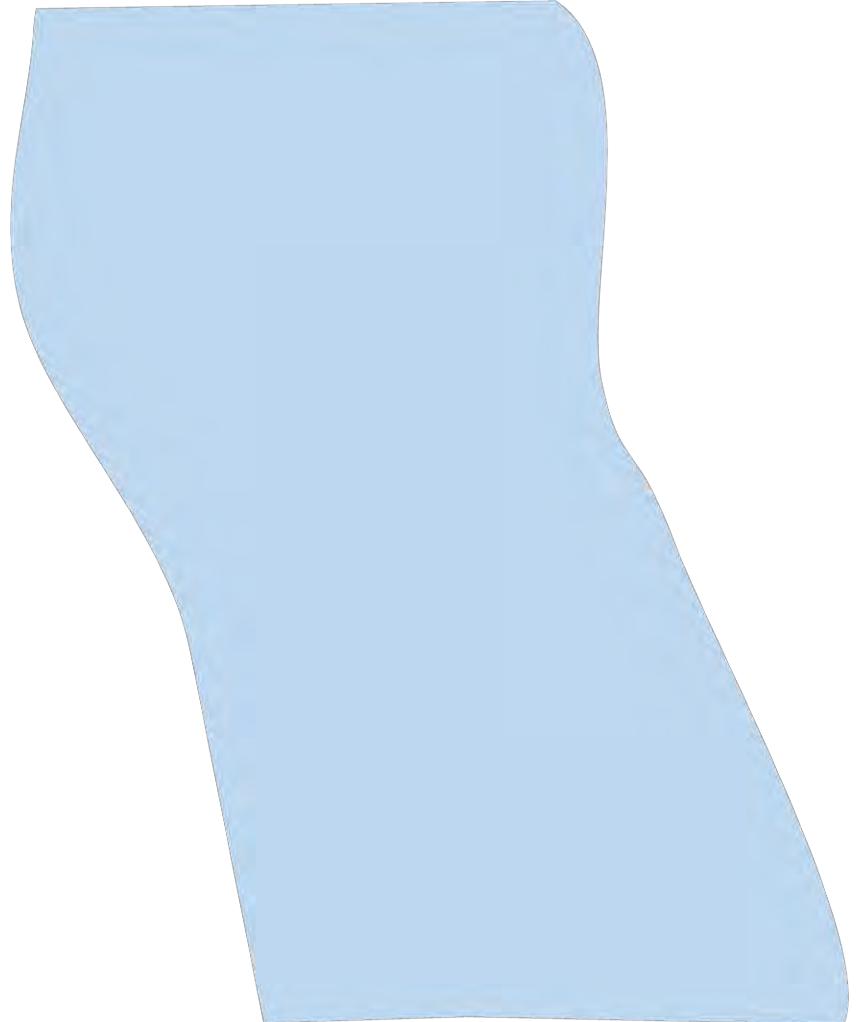
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. **reduce** \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

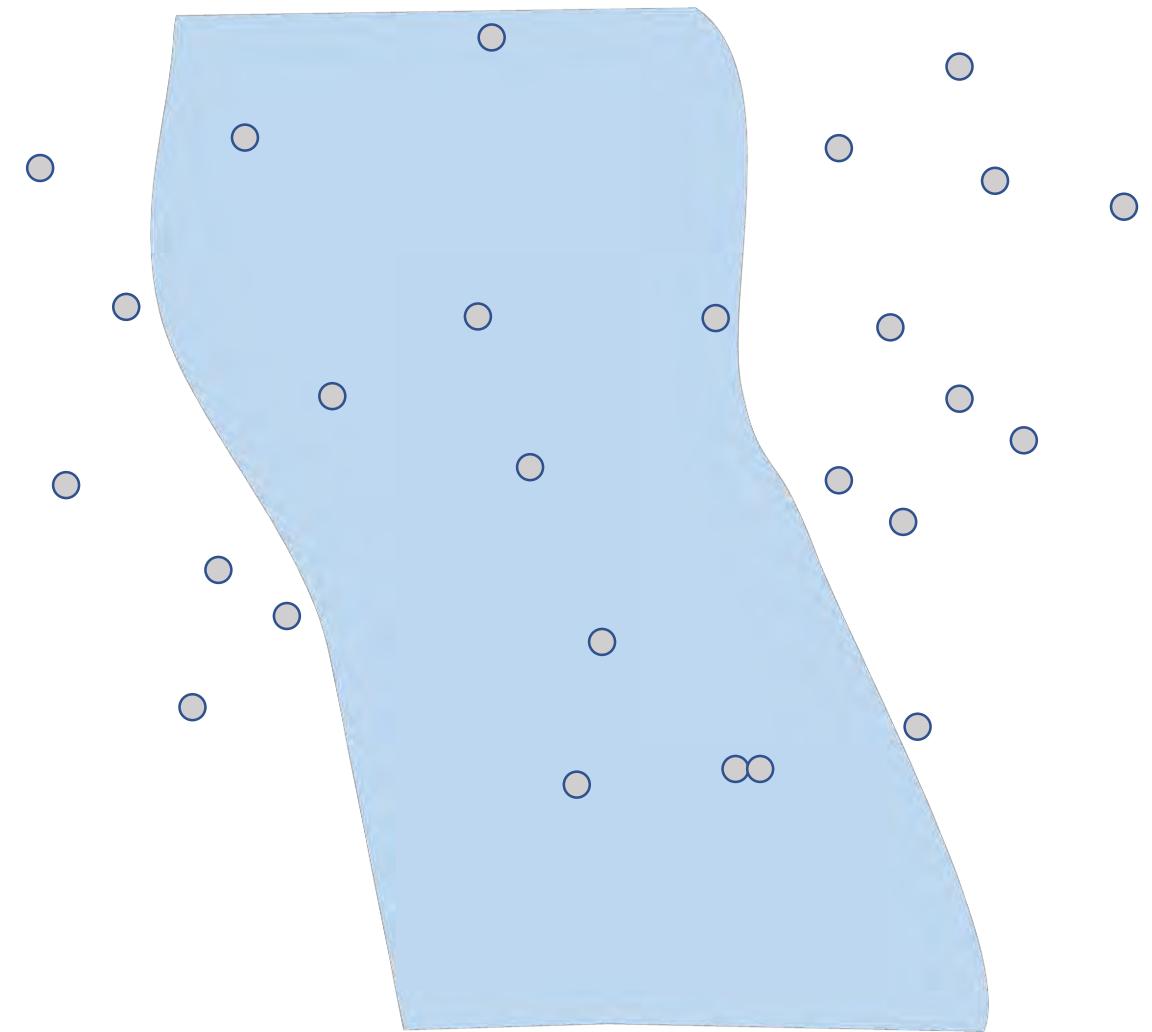
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

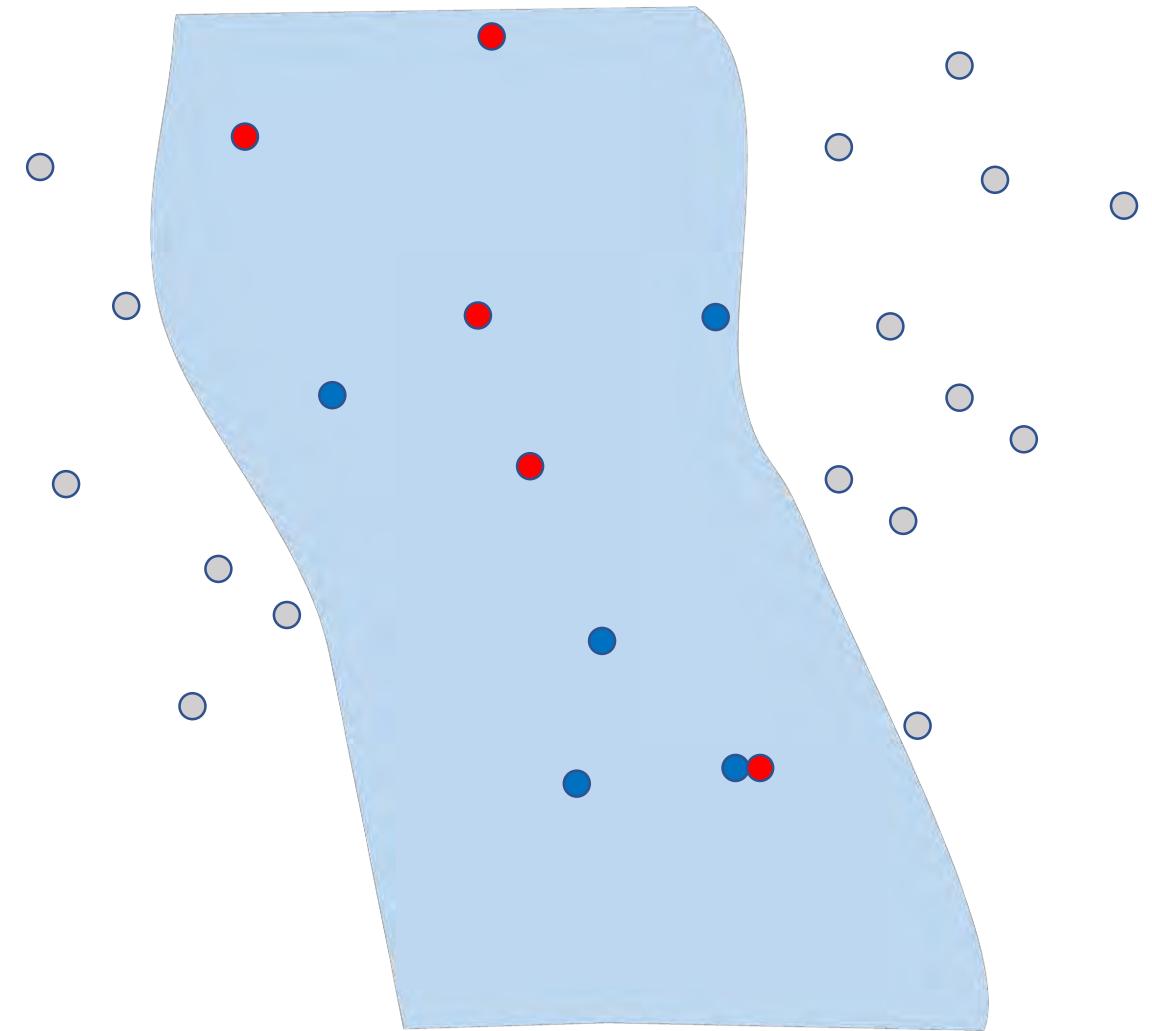
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

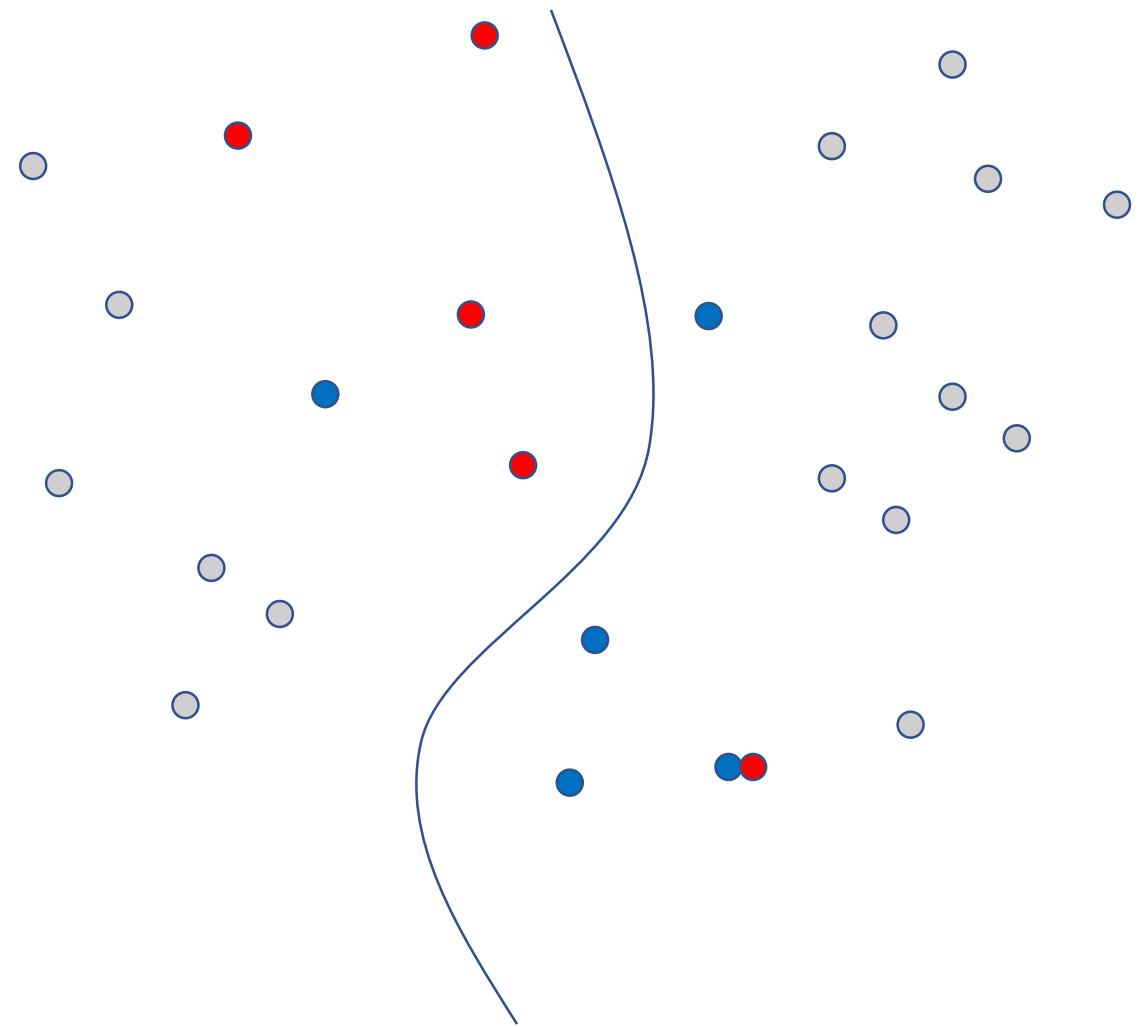
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

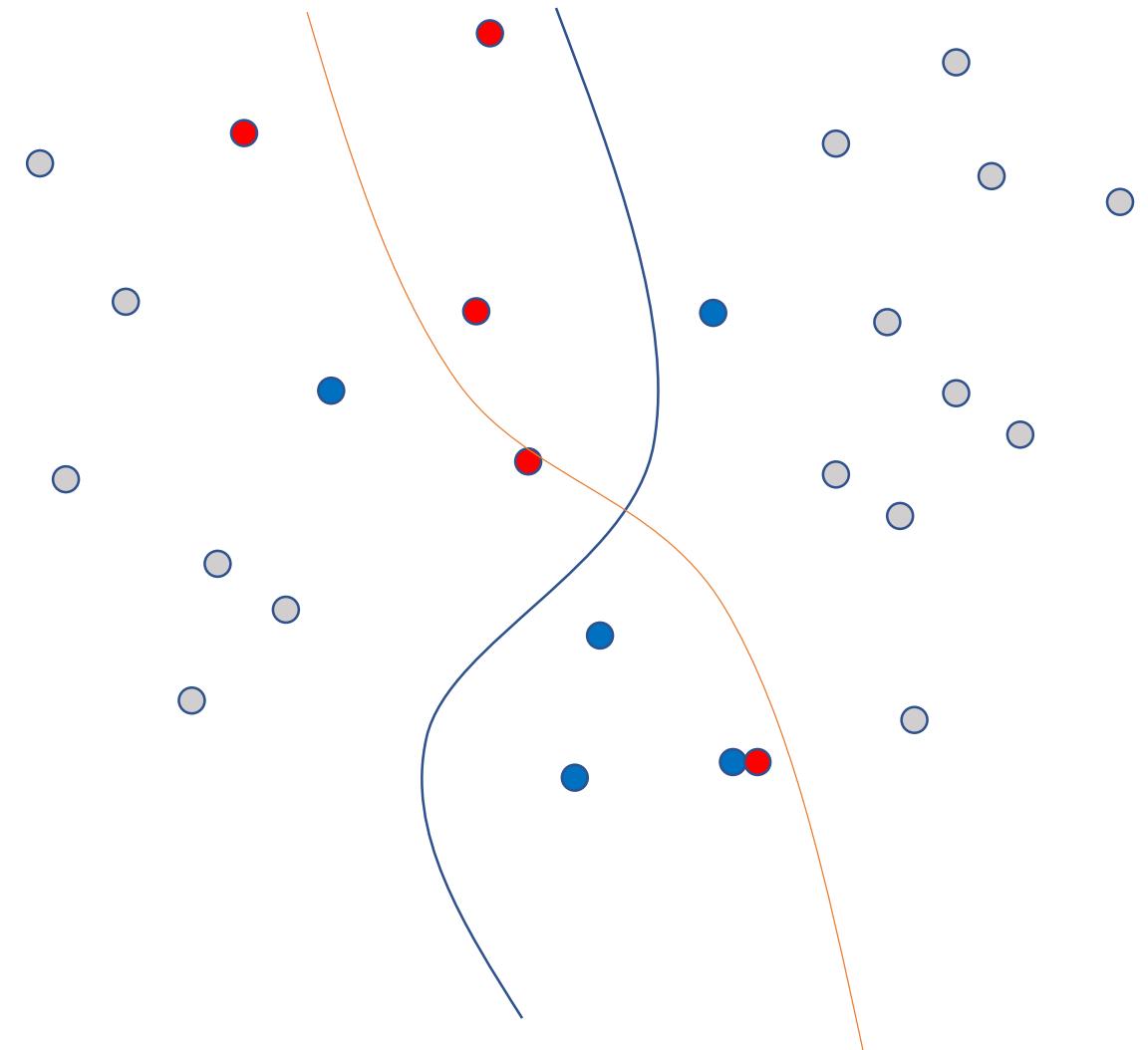
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

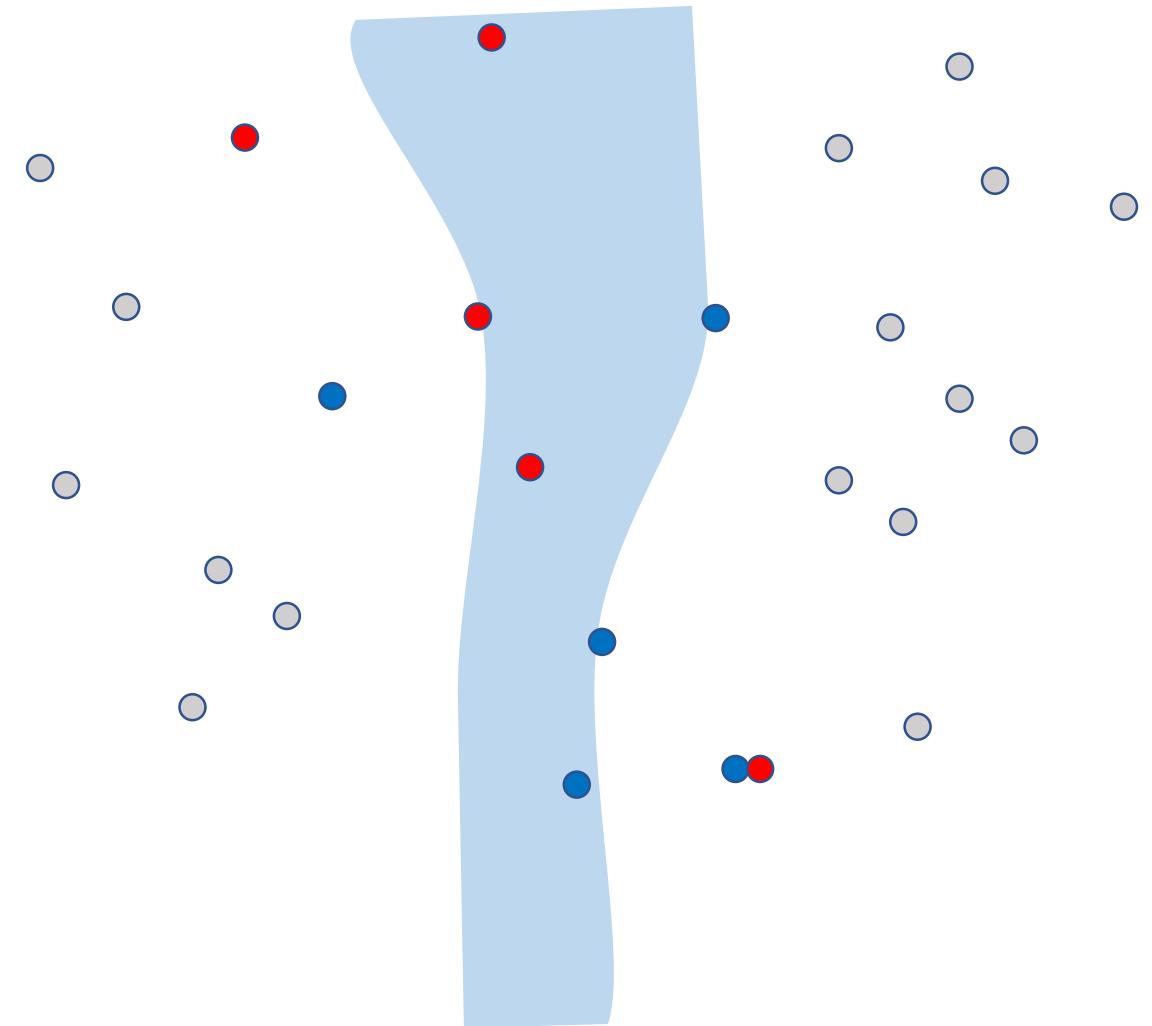
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

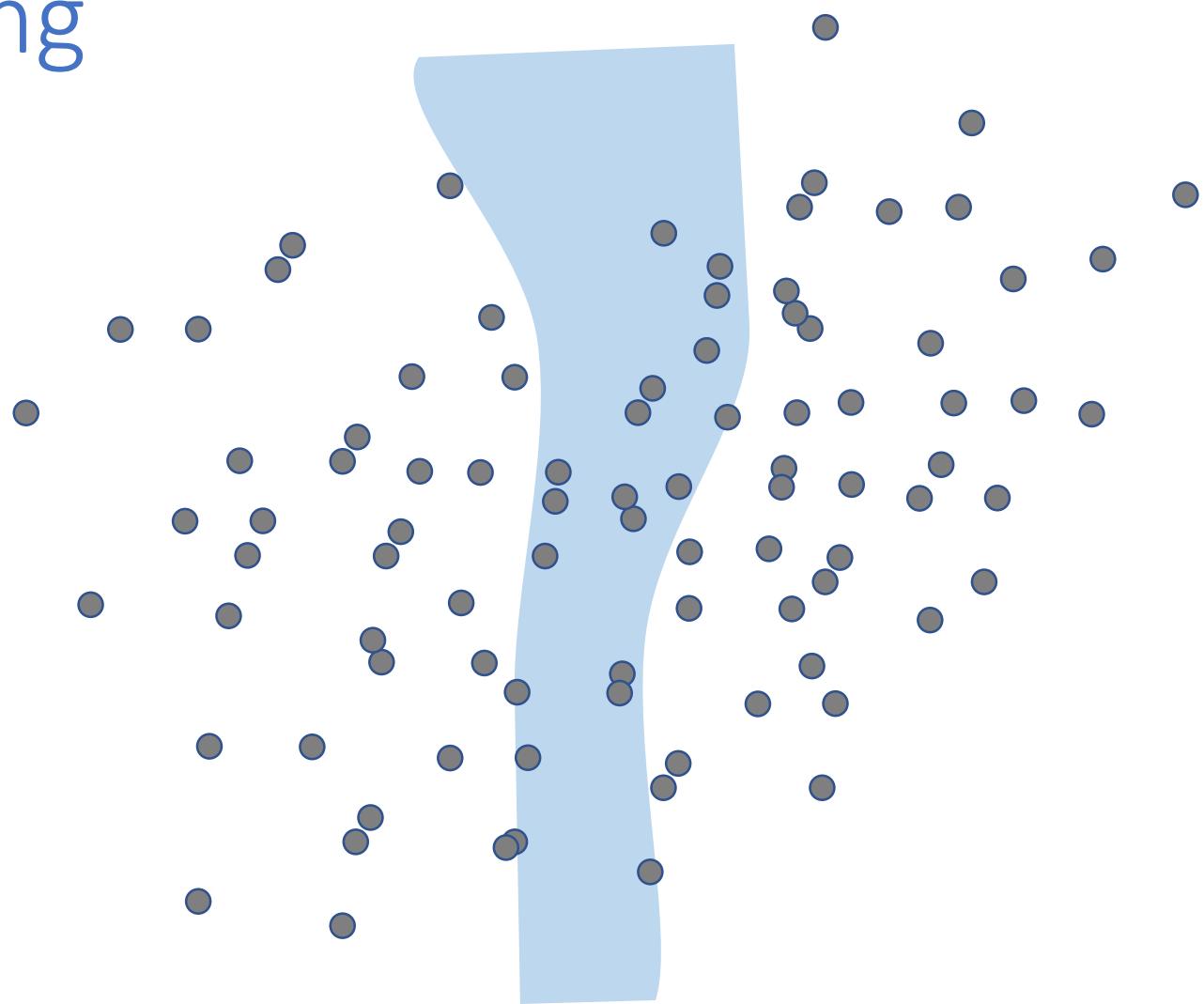
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

The point:

Any t with $f^* \in \mathcal{H}$ still,
 $R(f^*|\text{DIS}(\mathcal{H}))$ still **minimal** in \mathcal{H}

$$\begin{aligned} &\Rightarrow \hat{R}_Q(f^*) - \hat{R}_Q(\hat{f}) \\ &\leq R(f^*|\text{DIS}(\mathcal{H})) - R(\hat{f}|\text{DIS}(\mathcal{H})) + \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}} \\ &\leq \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}} \\ &\Rightarrow \underline{f^* \text{ never removed.}} \end{aligned}$$

Agnostic Active Learning

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

The point:

Any t with $f^* \in \mathcal{H}$ still,
 $R(f^*|\text{DIS}(\mathcal{H}))$ still **minimal** in \mathcal{H}

$$\begin{aligned} &\Rightarrow \hat{R}_Q(f^*) - \hat{R}_Q(\hat{f}) \\ &\leq R(f^*|\text{DIS}(\mathcal{H})) - R(\hat{f}|\text{DIS}(\mathcal{H})) + \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}} \\ &\leq \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}} \\ &\Rightarrow \underline{f^* \text{ never removed.}} \end{aligned}$$

Next: **How many labels does it use?**

Sample Complexity Analysis

Hanneke (2007,...)

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Sample Complexity Analysis

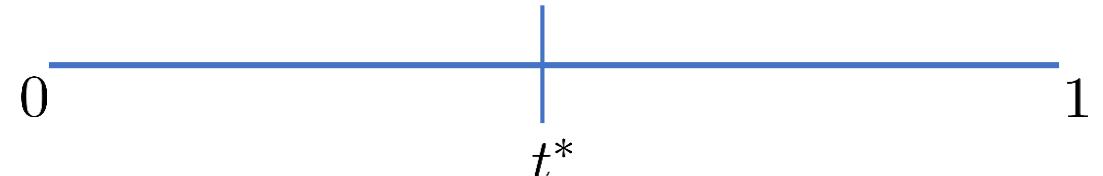
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Thresholds**, P_X Uniform(0, 1)
 $f(x) = \mathbb{I}[x \geq t]$



Sample Complexity Analysis

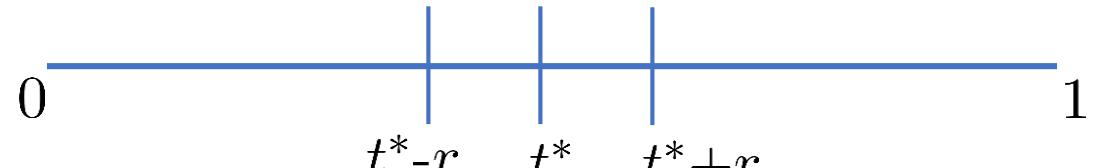
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Thresholds**, P_X Uniform(0, 1)
 $f(x) = \mathbb{I}[x \geq t]$



$$\text{DIS}(B(f^*, r)) = [t^* - r, t^* + r)$$

$$P_X(\text{DIS}(B(f^*, r))) = 2r$$

$$\theta = 2$$

Sample Complexity Analysis

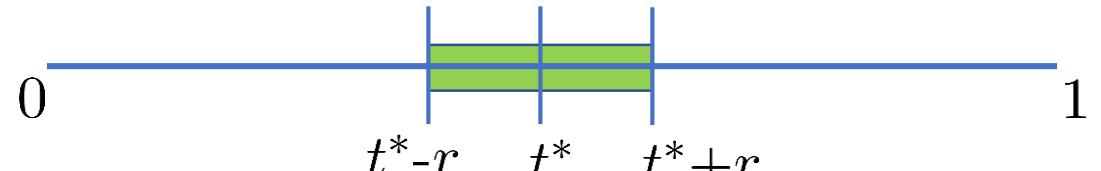
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Thresholds**, P_X Uniform(0, 1)
 $f(x) = \mathbb{I}[x \geq t]$



$$\text{DIS}(B(f^*, r)) = [t^* - r, t^* + r)$$

$$P_X(\text{DIS}(B(f^*, r))) = 2r$$

$$\Rightarrow \theta = 2$$

Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

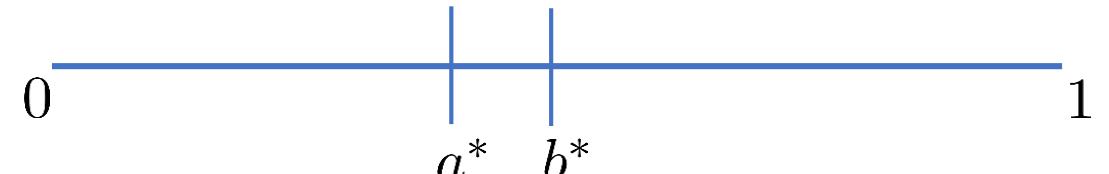
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

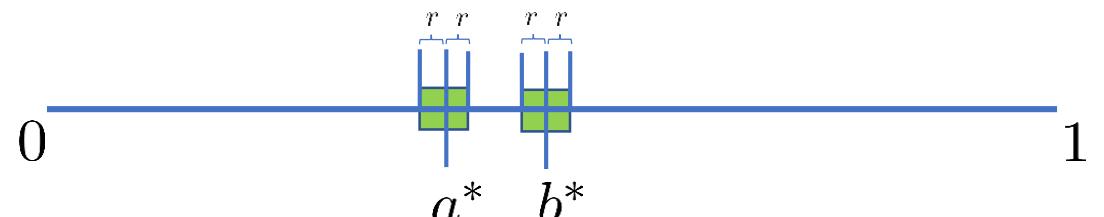
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



$$w^* := b^* - a^*$$

If $r < w^*$,

$$\text{DIS}(B(f^*, r)) = [a^* - r, a^* + r] \cup (b^* - r, b^* + r]$$

$$P_X(\text{DIS}(B(f^*, r))) = 4r$$

Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

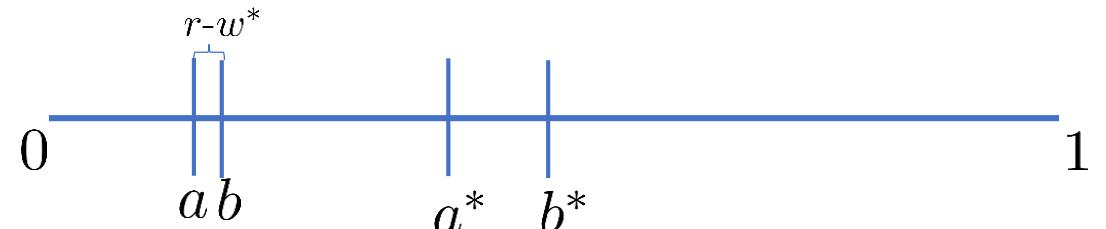
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



$$w^* := b^* - a^*$$

If $r > w^*$,

$$\text{DIS}(B(f^*, r)) = \mathcal{X}$$

$$P_X(\text{DIS}(B(f^*, r))) = 1$$

Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

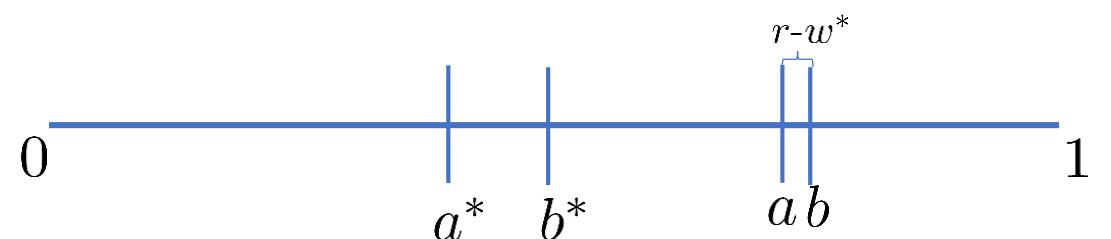
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



$$w^* := b^* - a^*$$

If $r > w^*$,

$$\text{DIS}(B(f^*, r)) = \mathcal{X}$$

$$P_X(\text{DIS}(B(f^*, r))) = 1$$

Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

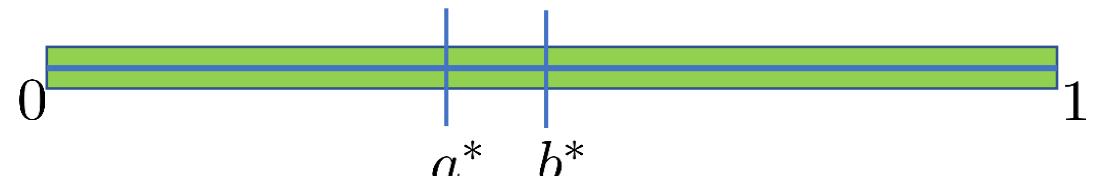
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



$$w^* := b^* - a^*$$

If $r > w^*$,

$$\text{DIS}(B(f^*, r)) = \mathcal{X}$$

$$P_X(\text{DIS}(B(f^*, r))) = 1$$

Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

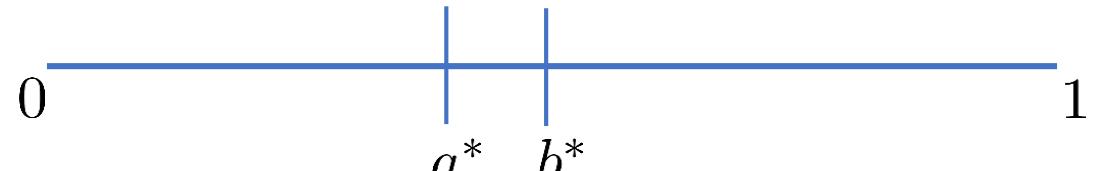
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



$$w^* := b^* - a^*$$

If $r < w^*$, $P_X(\text{DIS}(B(f^*, r))) = 4r$

If $r > w^*$, $P_X(\text{DIS}(B(f^*, r))) = 1$

$$\Rightarrow \theta \leq \max\left\{4, \frac{1}{w^*}\right\}$$

Sample Complexity Analysis

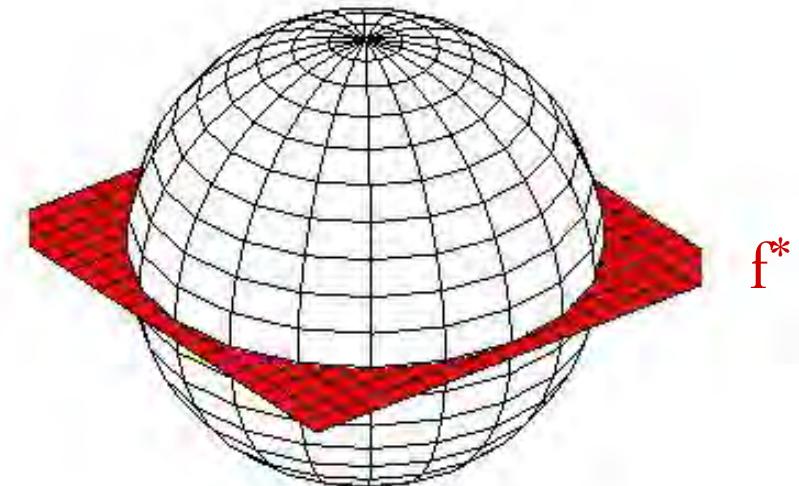
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0),
 n dimensions, uniform P_X on sphere.



Sample Complexity Analysis

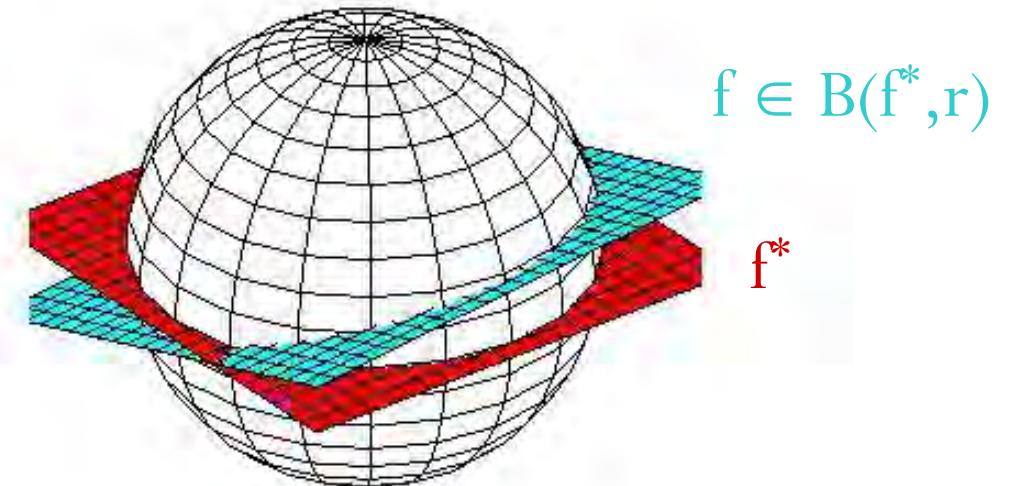
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0),
 n dimensions, uniform P_X on sphere.



Sample Complexity Analysis

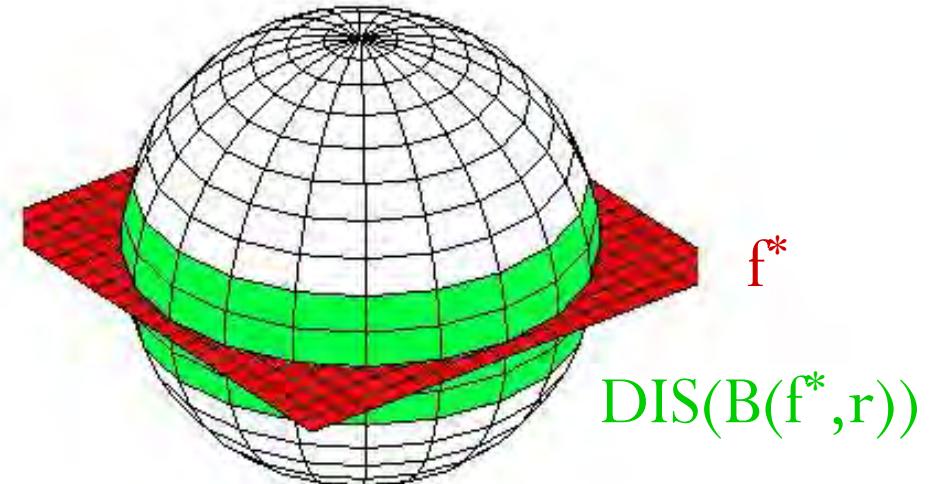
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0),
 n dimensions, uniform P_X on sphere.



Sample Complexity Analysis

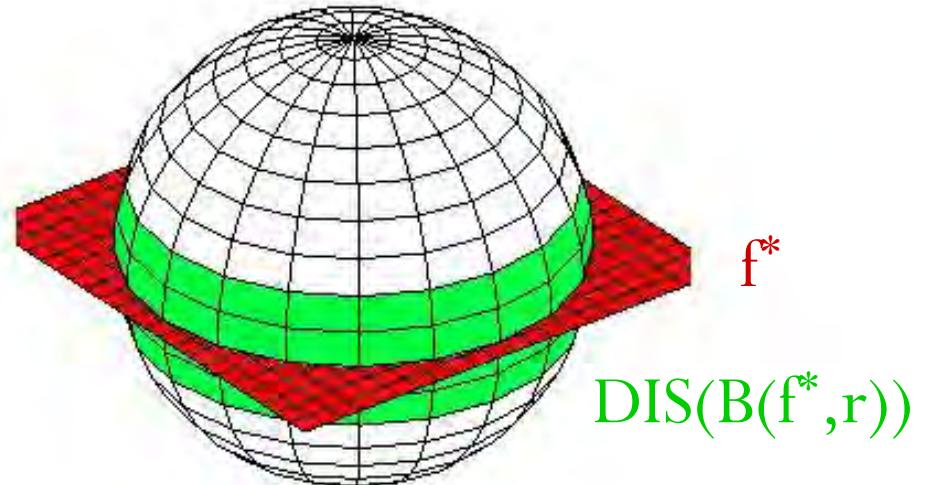
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0),
 n dimensions, uniform P_X on sphere.



Some geometry \Rightarrow for small r ,
 $P_X(\text{DIS}(B(f^*, r))) \propto \sqrt{nr}$.
 $\Rightarrow \theta \propto \sqrt{n}$.

Sample Complexity Analysis

Bounded Noise assumption: (aka Massart noise)

$$\exists \beta < 1/2 \text{ s.t. } P(Y \neq f^*(X)|X) \leq \beta \text{ everywhere}$$

	Sample Complexity: $R(\hat{f}) \leq R(f^*) + \epsilon$	Excess Error: n labels
Passive	$\frac{d}{\epsilon}$	$\frac{d}{n}$
Active	$d\theta \log(\frac{1}{\epsilon})$	$e^{-n/d\theta}$

Sample Complexity Analysis

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Theorem: $P(Y \neq f^*(X)|X) \leq \beta$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\# \text{ labels} \approx d\theta \log\left(\frac{1}{\epsilon}\right).$$

Proof Sketch:

Round t , all $f \in \mathcal{H}$ agree on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with

$$R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*) \frac{d}{2^t}}$$

\Rightarrow surviving f after round t have $R(f) - R(f^*) \lesssim \frac{d}{2^t}$

$\Rightarrow t \gtrsim \log\left(\frac{d}{\epsilon}\right)$ suffices

Also \Rightarrow after round $t-1$, $\mathcal{H} \subseteq B(f^*, d/2^{t-1})$

$$\Rightarrow |Q| \lesssim P_X(\text{DIS}(B(f^*, d/2^{t-1})))|S| \leq \theta \frac{d}{2^{t-1}} |S| = \theta d 2$$

$$\sum_{t=1}^{\log(d/\epsilon)} \theta d = \theta d \log\left(\frac{d}{\epsilon}\right)$$



Sample Complexity Analysis

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Bounded noise:

$$\begin{aligned} R(f) - R(f^*) &= \int_{f \neq f^*} (P(Y = f^*(X)|X) - P(Y \neq f^*(X)|X)) dP_X \\ &\geq (1 - 2\beta) P_X(f \neq f^*) \end{aligned}$$

Theorem: $P(Y \neq f^*(X)|X) \leq \beta$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\# \text{ labels} \approx d\theta \log\left(\frac{1}{\epsilon}\right).$$

Proof Sketch:

Round t , all $f \in \mathcal{H}$ agree on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with
 $R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*) \frac{d}{2^t}}$

\Rightarrow surviving f after round t have $R(f) - R(f^*) \lesssim \frac{d}{2^t}$
 $\Rightarrow t \gtrsim \log\left(\frac{d}{\epsilon}\right)$ suffices

Also \Rightarrow after round $t - 1$, $\mathcal{H} \subseteq B(f^*, d/2^{t-1})$

$$\Rightarrow |Q| \lesssim P_X(\text{DIS}(B(f^*, d/2^{t-1}))) |S| \leq \theta \frac{d}{2^{t-1}} |S| = \theta d 2$$

$$\sum_{t=1}^{\log(d/\epsilon)} \theta d = \theta d \log\left(\frac{d}{\epsilon}\right)$$



Sample Complexity Analysis

Agnostic Learning: (no assumptions)

Denote $\beta = R(f^*)$

	Sample Complexity: $R(\hat{f}) \leq R(f^*) + \epsilon$	Excess Error: n labels
Passive	$d \frac{\beta}{\epsilon^2}$	$\sqrt{\frac{d\beta}{n}}$
Active	$d\theta \frac{\beta^2}{\epsilon^2}$	$\sqrt{\frac{d\beta^2\theta}{n}}$

Sample Complexity Analysis

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Theorem: $\beta = R(f^*)$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\# \text{ labels} \approx d\theta \frac{\beta^2}{\epsilon^2}.$$

Proof Sketch:

Round t , all $f \in \mathcal{H}$ agree on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with

$$R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*) \frac{d}{2^t}}$$

\Rightarrow surviving f after round t have $R(f) - R(f^*) \lesssim \sqrt{\beta \frac{d}{2^t}} + \frac{d}{2^t}$

(Roughly) $\sqrt{\beta \frac{d}{2^t}}$

$\Rightarrow t \gtrsim \log(d \frac{\beta}{\epsilon^2})$ suffices

Also \Rightarrow after round $t-1$, $\mathcal{H} \subseteq B\left(f^*, 2\beta + \sqrt{\beta \frac{d}{2^{t-1}}}\right) \subseteq B(f^*, 3\beta)$ (for large t)

$\Rightarrow |Q| \lesssim P_X(\text{DIS}(B(f^*, 3\beta)))|S| \lesssim \theta\beta|S| = \theta\beta 2^t$

$$\sum_{t=1}^{\log(d\beta/\epsilon^2)} \theta\beta 2^t \sim \theta d \frac{\beta^2}{\epsilon^2}$$



Sample Complexity Analysis

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
 3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

$$P_X(f \neq f^*) \leq R(f) + R(f^*) = 2\beta + R(f) - R(f^*)$$

Theorem: $\beta = R(f^*)$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\# \text{ labels} \approx d\theta \frac{\beta^2}{\epsilon^2}.$$

Proof Sketch:

Round t , all $f \in \mathcal{H}$ agree on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with

$$R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*) \frac{d}{2^t}}$$

\Rightarrow surviving f after round t have $R(f) - R(f^*) \lesssim \sqrt{\beta \frac{d}{2^t}} + \frac{d}{2^t}$

(Roughly) $\sqrt{\beta \frac{d}{2^t}}$

$\Rightarrow t \gtrsim \log(d \frac{\beta}{\epsilon^2})$ suffices

Also \Rightarrow after round $t-1$, $\mathcal{H} \subseteq B\left(f^*, 2\beta + \sqrt{\beta \frac{d}{2^{t-1}}}\right) \subseteq B(f^*, 3\beta)$ (for large t)

$$\Rightarrow |Q| \lesssim P_X(\text{DIS}(B(f^*, 3\beta)))|S| \lesssim \theta\beta|S| = \theta\beta 2^t$$

$$\sum_{t=1}^{\log(d\beta/\epsilon^2)} \theta\beta 2^t \sim \theta d \frac{\beta^2}{\epsilon^2}$$



Sample Complexity Analysis

When is θ small?

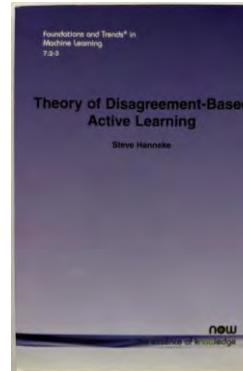
- Linear separators, P_X has a density,
 f^* boundary intersects interior of support
 $\Rightarrow \theta$ bounded
- Linear separators, P_X has a density
 $\Rightarrow \theta \ll \frac{1}{\epsilon}$
- \mathcal{H} smoothly-parametrized model,
 P_X “regular” density w/ compact support,
other technical conditions on \mathcal{H}
 $\Rightarrow \theta \propto \# \text{ parameters for } \mathcal{H}$
- ...

Sample Complexity Analysis

When is θ small?

- Linear separators, P_X has a density,
 f^* boundary intersects interior of support
 $\Rightarrow \theta$ bounded
- Linear separators, P_X has a density
 $\Rightarrow \theta \ll \frac{1}{\epsilon}$
- \mathcal{H} smoothly-parametrized model,
 P_X “regular” density w/ compact support,
other technical conditions on \mathcal{H}
 $\Rightarrow \theta \propto \# \text{ parameters for } \mathcal{H}$
- ...

Lots more 



Stopping Criterion

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Stopping criteria:

- Any-time
- Label budget
- Run out of unlabeled data
- Check $\max_{f \in \mathcal{H}} \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}} < \epsilon$

Simpler Agnostic Active Learning

Hsu (2010,...)

$Q \leftarrow \{\}$

for $m = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** a random point x

2. **optimize** $\forall y, \hat{f}_y \leftarrow \operatorname{argmin}_{f \in \mathcal{H}: f(x)=y} \hat{R}_Q(f)$

3. if $|\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_- \neq \hat{f}_+) \frac{d}{|Q|}}$

then **label** x , add it to Q

output $\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$

- Roughly same sample complexity as A^2 .
- Can implement as a **reduction** to ERM.
- In practice, replace ERM with any passive learner.

Surrogate Loss

Hanneke & Yang (2012)

$$Q \leftarrow \{\}$$

for $m = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** a random point x

2. **optimize** $\forall y, \hat{f}_y \leftarrow \operatorname{argmin}_{f \in \mathcal{H}: f(x)=y} \hat{R}_Q^\ell(f)$

3. if $|\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_- \neq \hat{f}_+) \frac{d}{|Q|}}$

then **label** x , add it to Q

output $\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$

- Roughly same sample complexity as A^2 .
- Can implement as a **reduction** to ERM.
- In practice, replace ERM with any passive learner.

Consider learner that minimizes a **surrogate loss**
 $\ell : \mathbb{R} \times \{-1, +1\} \rightarrow \mathbb{R}_+$
(e.g., hinge loss, squared loss, exponential loss, ...)

Now \mathcal{H} is **real-valued** functions

$$\hat{R}_Q^\ell(f) = \frac{1}{|Q|} \sum_{(x,y) \in Q} \ell(f(x), y)$$

Theorem: Bounded noise, plus strong assumptions on \mathcal{H}, ℓ, P
still get $R(\hat{f}) \leq R(f^*) + \epsilon$ with # labels

$$\approx \theta d \log\left(\frac{1}{\epsilon}\right)$$

Importance-Weighted Active Learning

Beygelzimer, Dasgupta,
Langford (2009)

$$Q \leftarrow \{\}$$

for $m = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** a random point x
2. **set** sampling probability p_x
3. **flip** coin with prob p_x of heads
4. if heads, **label** x , add to Q with weight $1/p_x$

output $\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$ (weighted loss)

Use importance weights to stay **unbiased**:
 $\mathbb{E}[\hat{R}_Q(f)] = R(f)$

Now Q set of triples (x, y, w)

$$\hat{R}_Q(f) = \frac{1}{|Q|} \sum_{(x,y,w) \in Q} w \mathbb{I}[f(x) \neq y]$$

- **Any** choice of Step 2 (setting p_x) is fine (just p_x not too small, else high variance)
- Can set p_x in a way to recover A^2 sample complexity
$$p_x = \mathbb{I}\left[|\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_+ \neq \hat{f}_-) \frac{d}{|Q|}} \right]$$

Importance-Weighted Active Learning

Beygelzimer, Dasgupta,
Langford (2009)

$Q \leftarrow \{\}$

for $m = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** a random point x
2. **set** sampling probability p_x
3. **flip** coin with prob p_x of heads
4. if heads, **label** x , add to Q with weight $1/p_x$

output $\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$ (weighted loss)

Use importance weights to stay **unbiased**:
 $\mathbb{E}[\hat{R}_Q(f)] = R(f)$

Now Q set of triples (x, y, w)

$$\hat{R}_Q(f) = \frac{1}{|Q|} \sum_{(x,y,w) \in Q} w \mathbb{I}[f(x) \neq y]$$

- **Any** choice of Step 2 (setting p_x) is fine (just p_x not too small, else high variance)
- Can set p_x in a way to recover A^2 sample complexity
$$p_x = \mathbb{I}\left[|\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_+ \neq \hat{f}_-) \frac{d}{|Q|}} \right]$$
- In practice, replace ERM with any passive learner (e.g., ERM with a surrogate loss)
- (approx) implementation in **Vowpal Wabbit** library

Questions?

Further reading:

- D. Cohn, L. Atlas, R. Ladner. Improving generalization with active learning. *Machine Learning*, 1994
- M. F. Balcan, A. Beygelzimer, J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 2009.
- S. Hanneke. A bound on the label complexity of agnostic active learning. ICML 2007.
- S. Dasgupta, D. Hsu, C. Monteleoni. A general agnostic active learning algorithm. NeurIPS 2007.
- S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 2011.
- A. Beygelzimer, S. Dasgupta, J. Langford. Importance weighted active learning. ICML 2009.
- A. Beygelzimer, D. Hsu, J. Langford, T. Zhang. Agnostic active learning without constraints. NeurIPS 2010.
- S. Hanneke. Theoretical Foundations of Active Learning. PhD Thesis, CMU, 2009.
- D. Hsu. Algorithms for Active Learning. PhD Thesis, UCSD, 2010.
- Y. Wiener, S. Hanneke, R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *Journal of Machine Learning Research*, 2015.
- S. Hanneke. Refined error bounds for several learning algorithms. *Journal of Machine Learning Research*, 2016.
- E. Friedman. Active learning for smooth problems. COLT 2009.
- S. Mahalanabis. Subset and Sample Selection for Graphical Models: Gaussian Processes, Ising Models and Gaussian Mixture Models. PhD Thesis, University of Rochester, 2012.
- S. Hanneke. Theory of Disagreement-Based Active Learning. *Foundations and Trends in Machine Learning*, 2014.
- S. Hanneke, L. Yang. Surrogate losses in passive and active learning. arXiv:1207.3772.

Part 3: Beyond Disagreement-Based Active Learning – Current Directions

- Subregion-Based Active Learning
- Margin-Based Active Learning: Linear Separators
- Shattering-Based Active Learning
- Distribution-Free Analysis, Optimality
- TicToc: Adapting to Heterogeneous Noise
- Tsybakov Noise

**Tutorial on Active Learning:
Theory to Practice**

Steve Hanneke

Toyota Technological Institute at Chicago
steve.hanneke@gmail.com

Robert Nowak

University of Wisconsin - Madison
rdnowak@wisc.edu

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. **reduce** \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

Subregion-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \mathcal{R}_{\epsilon'_t}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Instead, pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.
 $\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

Pick ϵ' carefully each round,
 $R(\hat{f}) - R(f^*) \leq \epsilon$ at end

e.g., Bounded noise: $\epsilon' \propto d2^{-t}$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

Subregion-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \mathcal{R}_{\epsilon'_t}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

$$\varphi_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using $\#$ labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

Subregion-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \mathcal{R}_{\epsilon'_t}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

$$\textbf{Agnostic case: } \varphi'_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, 2\beta + r)))}{2\beta + r}$$

Theorem:

$$\begin{aligned} R(\hat{f}) &\leq R(f^*) + \epsilon \text{ using # labels} \\ &\approx \varphi'_c d \frac{\beta^2}{\epsilon^2} \end{aligned}$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- $\mathcal{R}_{\epsilon'}(\mathcal{H}) = \text{DIS}(\mathcal{H})$ works
- Empirically (Zhang & Chaudhuri, 2014)
- Nice structure: e.g., **Linear separators**

Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r>\epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- $\mathcal{R}_{\epsilon'}(\mathcal{H}) = \text{DIS}(\mathcal{H})$ works
- Empirically (Zhang & Chaudhuri, 2014)
- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r>\epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r>\epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

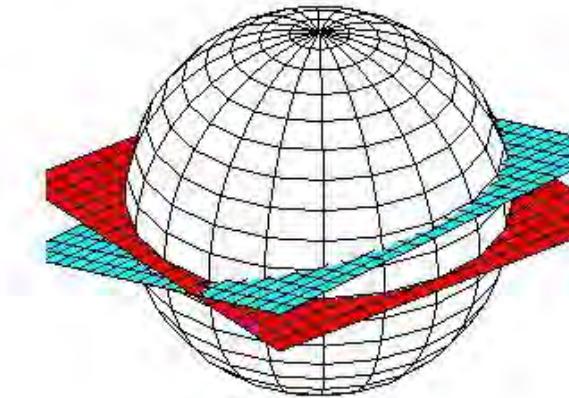
(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

Uniform P_X on d -dim sphere

For $w \in B(w^*, r)$, **project** to $\text{Span}(w, w^*)$



Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

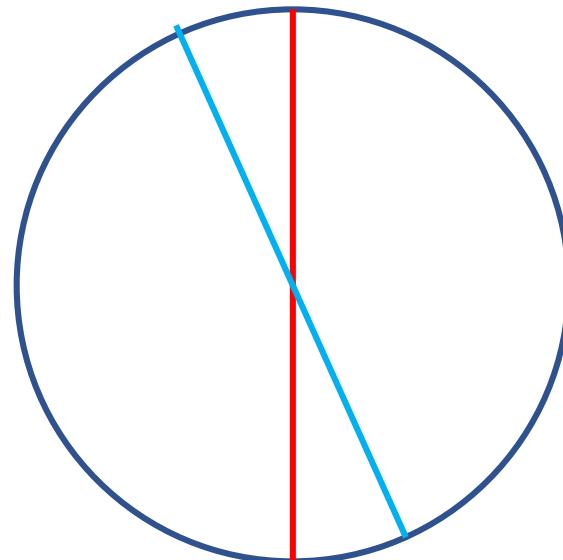
Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

Uniform P_X on d -dim sphere

For $w \in B(w^*, r)$, **project** to $\text{Span}(w, w^*)$

Most projected prob mass toward middle



Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

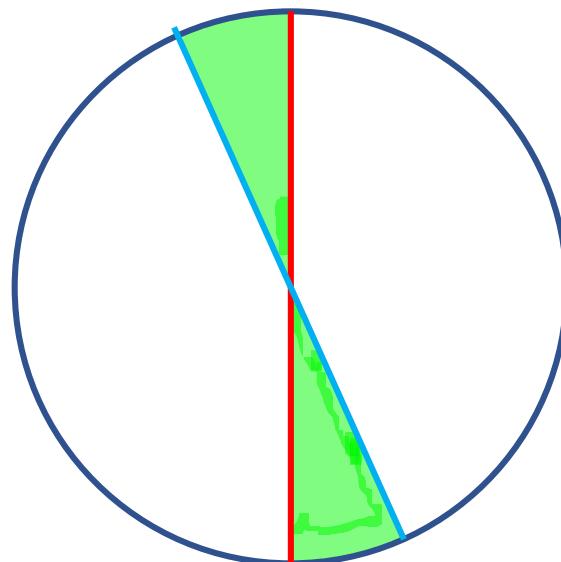
Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

Uniform P_X on d -dim sphere

For $w \in B(w^*, r)$, **project** to $\text{Span}(w, w^*)$

Most projected prob mass toward middle



Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

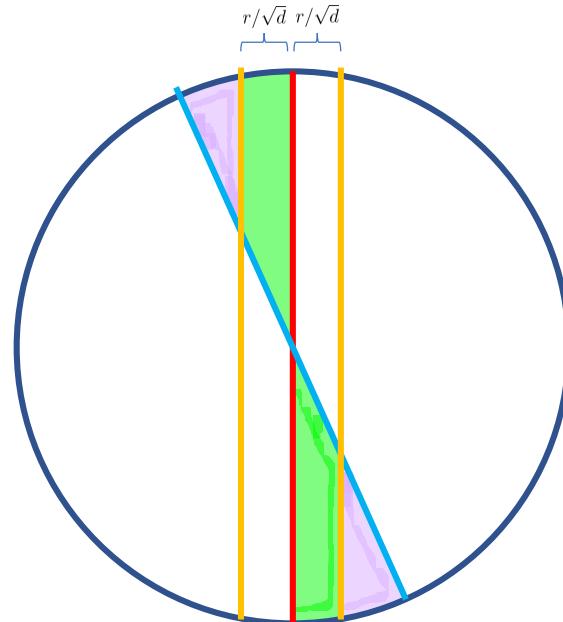
Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

Uniform P_X on d -dim sphere

For $w \in B(w^*, r)$, **project** to $\text{Span}(w, w^*)$

Most projected prob mass toward middle



$\text{DIS}(\{w, w^*\})$ in
slab of width $\approx r$

Most of its prob in
slab of width $\approx r/\sqrt{d}$

Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

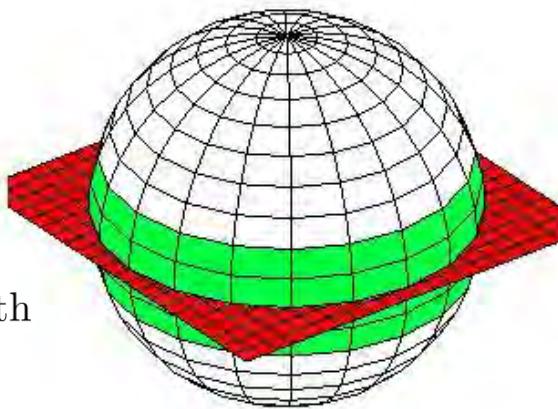
(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

$\text{DIS}(\mathcal{B}(f^*, r)) =$
slab of width $\approx r$

Take $\mathcal{R}_{r/c}(\mathcal{B}(f^*, r)) =$
slab of width $\approx r/\sqrt{d}$

Prob in slab $\approx \sqrt{d} \times \text{width}$

$\Rightarrow \varphi_c \leq \text{constant}$



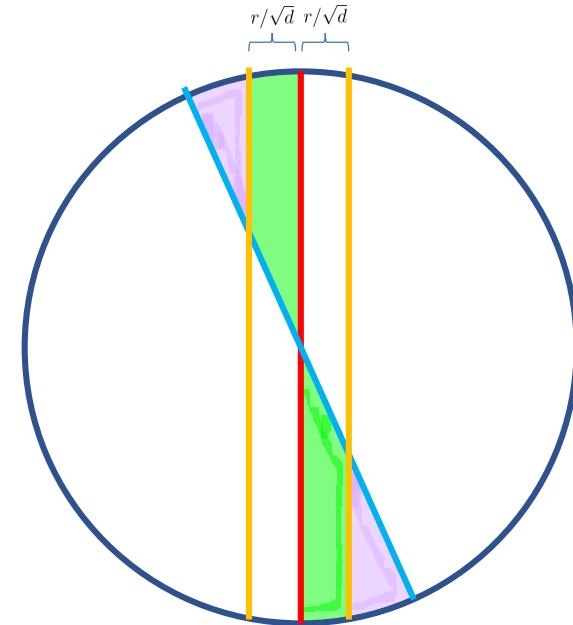
Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

Uniform P_X on d -dim sphere

For $w \in \mathcal{B}(w^*, r)$, **project** to $\text{Span}(w, w^*)$

Most projected prob mass toward middle



$\text{DIS}(\{w, w^*\})$ in
slab of width $\approx r$

Most of its prob in
slab of width $\approx r/\sqrt{d}$

Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

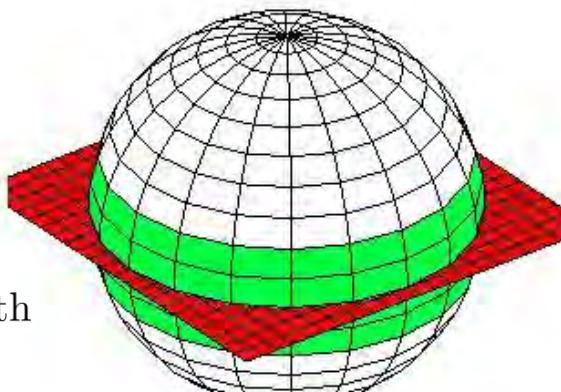
(Dasgupta, Kalai, Monteleoni, 2005;
Balcan - - - - -)

$\text{DIS}(\mathcal{B}(f^*, r)) =$
slab of width $\approx r$

Take $\mathcal{R}_{r/c}(\mathcal{B}(f^*, r)) =$
slab of width $\approx r/\sqrt{d}$

Prob in slab $\approx \sqrt{d} \times \text{width}$

$\Rightarrow \varphi_c \leq \text{constant}$



Pick **region $\mathcal{R}_{\epsilon'}(\mathcal{H})$** s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(\mathcal{B}(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels
 $\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$

\Rightarrow # labels $\approx d \log\left(\frac{1}{\epsilon}\right)$ suffice

Comparison:

Recall $\theta \approx \sqrt{d}$

$\Rightarrow A^2$ # labels $\approx d^{3/2} \log\left(\frac{1}{\epsilon}\right)$

Recall:
Passive $\approx \frac{d}{\epsilon}$

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

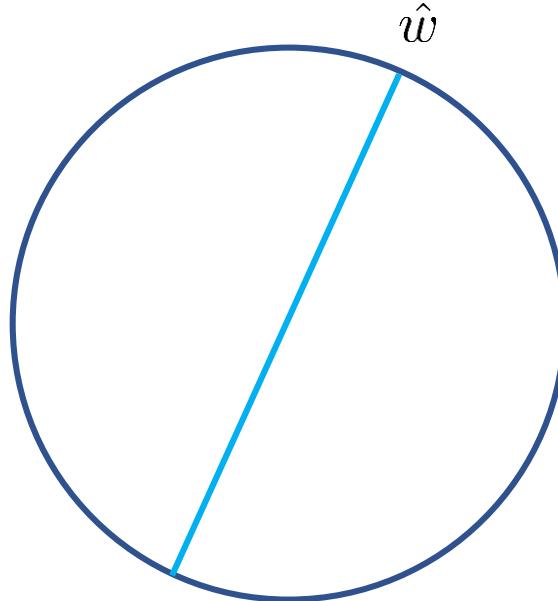
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

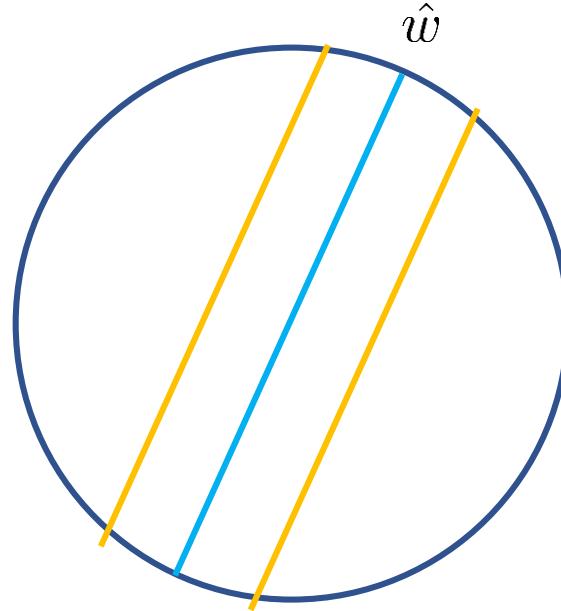
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

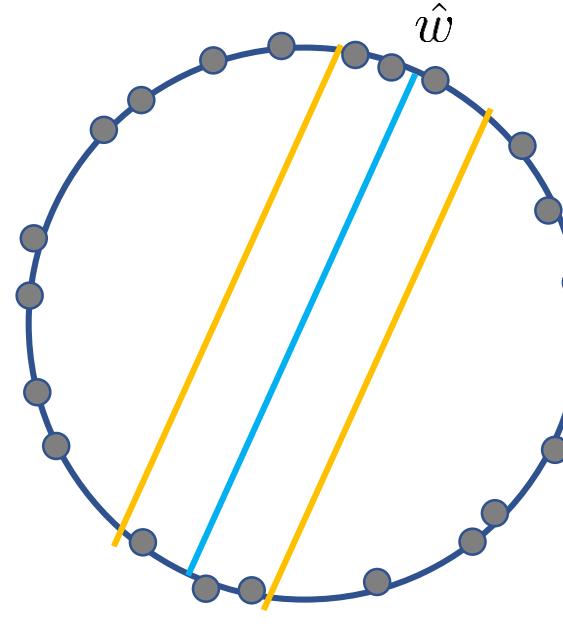
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

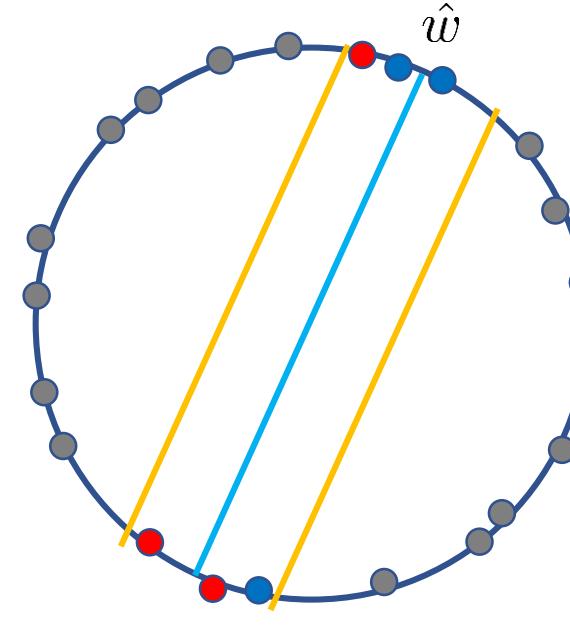
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

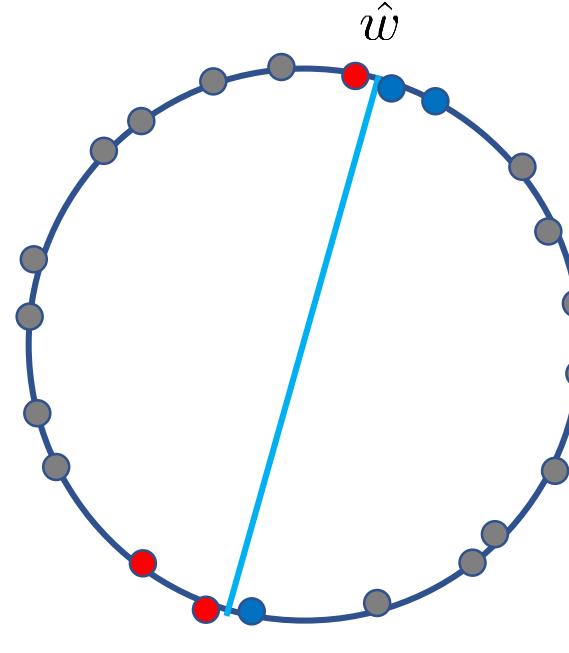
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

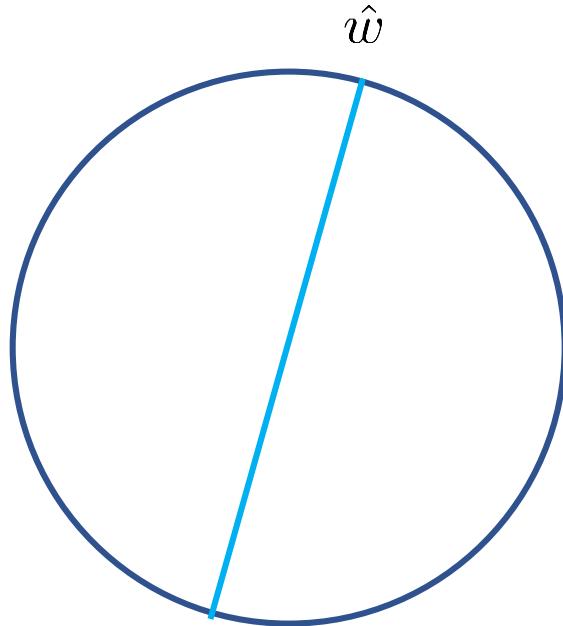
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels
 $\approx d \log\left(\frac{1}{\epsilon}\right)$

(also works for isotropic log-concave distributions)

Computational Efficiency

(Awasthi, Balcan, Long, 2014,...)

Uniform P_X on d -dim sphere

Efficient Alg

Initialize \hat{w}

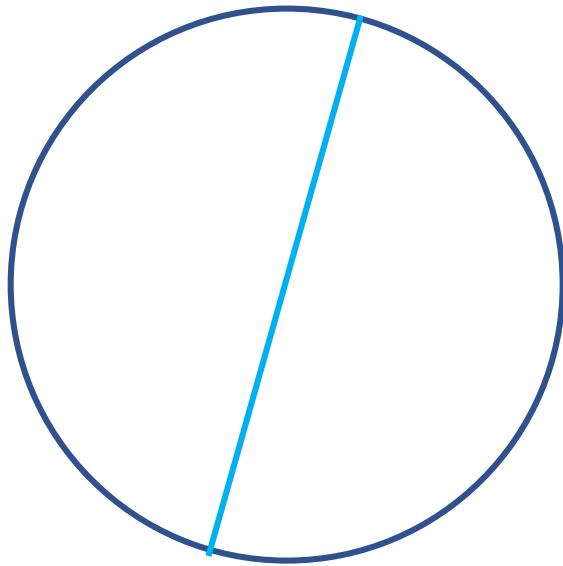
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q^{\ell_t}(w)$

output final \hat{w}



Surrogate loss

$$\ell_t(\langle w, x \rangle, y) \approx \max\{1 - 2^t \sqrt{d}(y \langle w, x \rangle), 0\}$$

Hinge loss slope changes each round

Computational Efficiency

(Awasthi, Balcan, Long, 2014,...)

Uniform P_X on d -dim sphere

Efficient Alg

Initialize \hat{w}

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q^{\ell_t}(w)$

output final \hat{w}

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels
 $\approx d \log\left(\frac{1}{\epsilon}\right)$
and running in polynomial time

Surrogate loss

$$\ell_t(\langle w, x \rangle, y) \approx \max\{1 - 2^t \sqrt{d}(y \langle w, x \rangle), 0\}$$

Hinge loss slope changes each round

Computational Efficiency

(Awasthi, Balcan, Long, 2014,...)

Uniform P_X on d -dim sphere

Efficient Alg

Initialize \hat{w}

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q^{\ell_t}(w)$

output final \hat{w}

Surrogate loss

$$\ell_t(\langle w, x \rangle, y) \approx \max\{1 - 2^t \sqrt{d}(y \langle w, x \rangle), 0\}$$

Hinge loss slope changes each round

Theorem: with **Bounded noise**,

$$R(\hat{f}) \leq R(f^*) + \epsilon \text{ using } \# \text{ labels} \\ \approx d \log\left(\frac{1}{\epsilon}\right)$$

and running in polynomial time

Theorem: with **Agnostic** case,

$$R(\hat{f}) \leq CR(f^*) \text{ in polynomial time}$$

(was first alg. known to achieve these; even passively)

(also works for isotropic log-concave distributions)

Up Next:
Shattering-Based Active Learning

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

$\text{DIS}(\mathcal{H})$ checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}

$\text{DIS}(\mathcal{H})$ checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}

$\text{DIS}(\mathcal{H})$ checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t.}$

$$P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$$

3. add the remaining points $x \in S$ to Q with label

$$\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$$

4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$

5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

$\text{DIS}(\mathcal{H})$ checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Denote $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

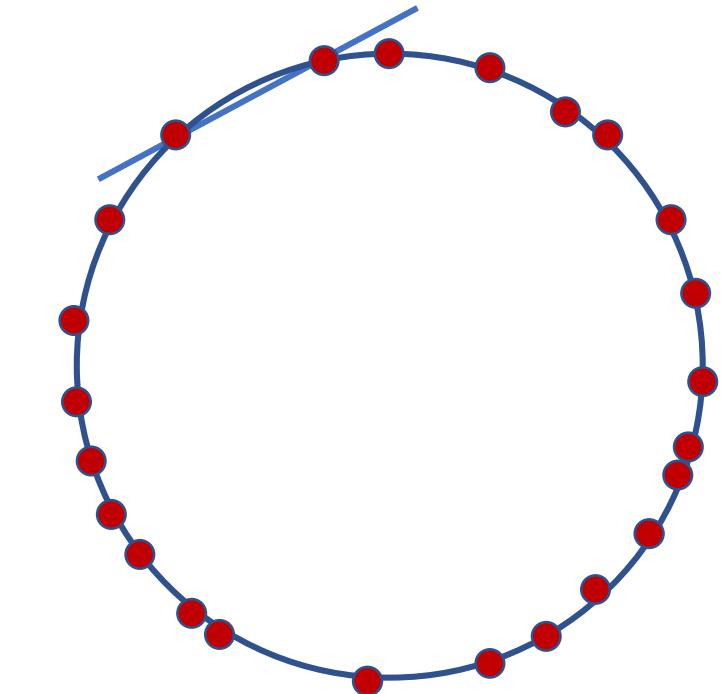
for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. add the remaining points $x \in S$ to Q with label $\hat{y}_x := \underset{y}{\operatorname{argmax}} P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. optimize $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all -1**

$\text{DIS}(\mathcal{H}) = \text{entire circle}$



Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. add the remaining points $x \in S$ to Q with label $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all -1**

$\text{DIS}(\mathcal{H}) = \text{entire circle}$

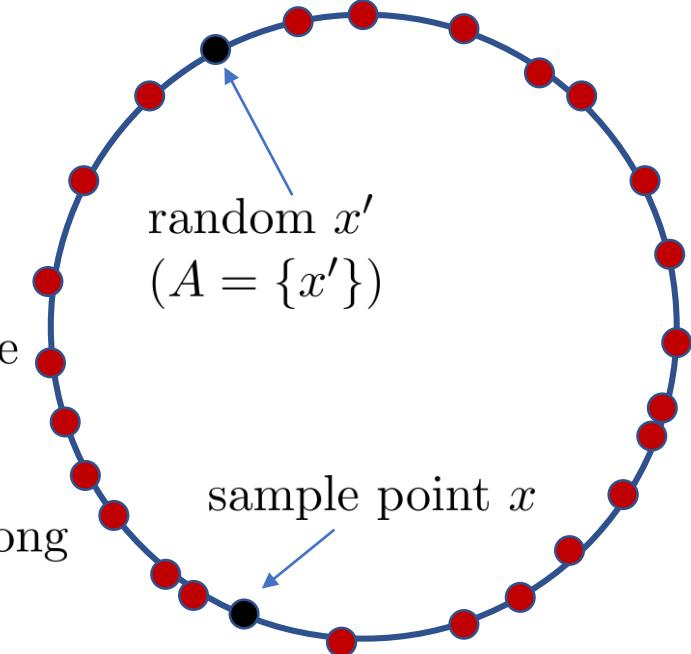
Try $k = 1$

Given sample x

Rand x' probably not close

Can't shatter $\{x, x'\}$
without a lot of points wrong

So won't query x



Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all -1**

$\text{DIS}(\mathcal{H}) = \text{entire circle}$

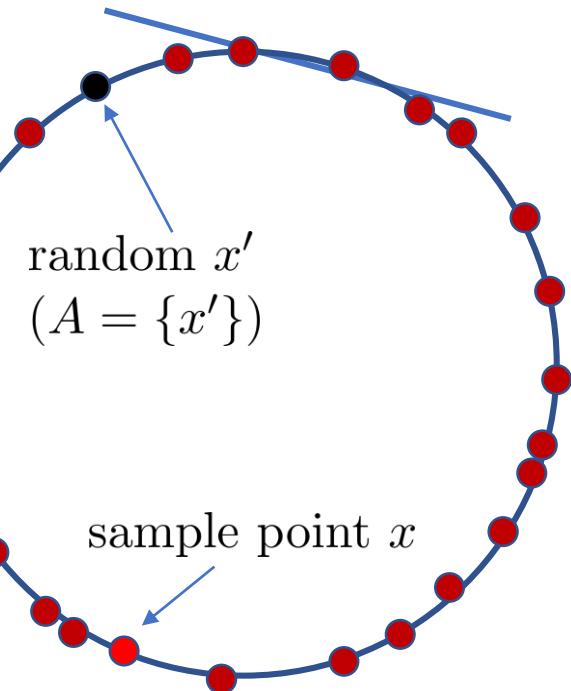
Try $k = 1$

Given sample x

Rand x' probably not close

Can't shatter $\{x, x'\}$
without a lot of points wrong

So won't query x



$\text{DIS}(\mathcal{H}_{x,-1})$ still entire circle (minus x)

$\text{DIS}(\mathcal{H}_{x,+1})$ **small** region

$$\Rightarrow \hat{y}_x = -1$$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all -1**

$\text{DIS}(\mathcal{H}) = \text{entire circle}$

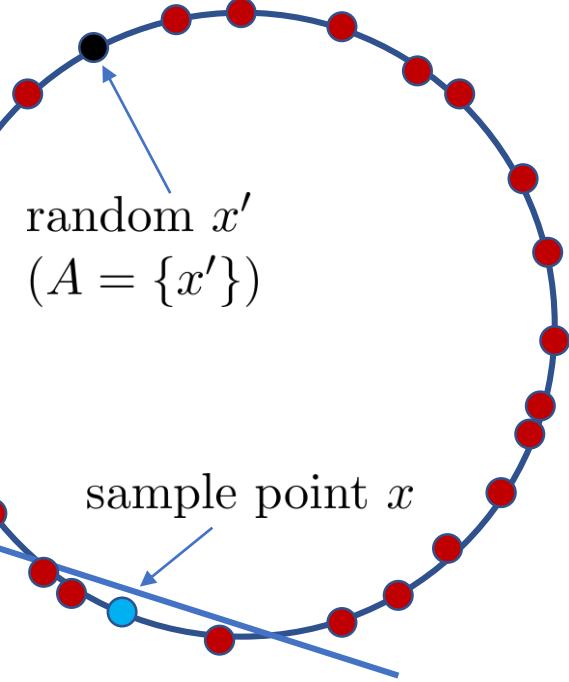
Try $k = 1$

Given sample x

Rand x' probably not close

Can't shatter $\{x, x'\}$
without a lot of points wrong

So won't query x



$\text{DIS}(\mathcal{H}_{x,-1})$ still entire circle (minus x)

$\text{DIS}(\mathcal{H}_{x,+1})$ **small** region

$$\Rightarrow \hat{y}_x = -1$$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all -1**

$\text{DIS}(\mathcal{H}) = \text{entire circle}$

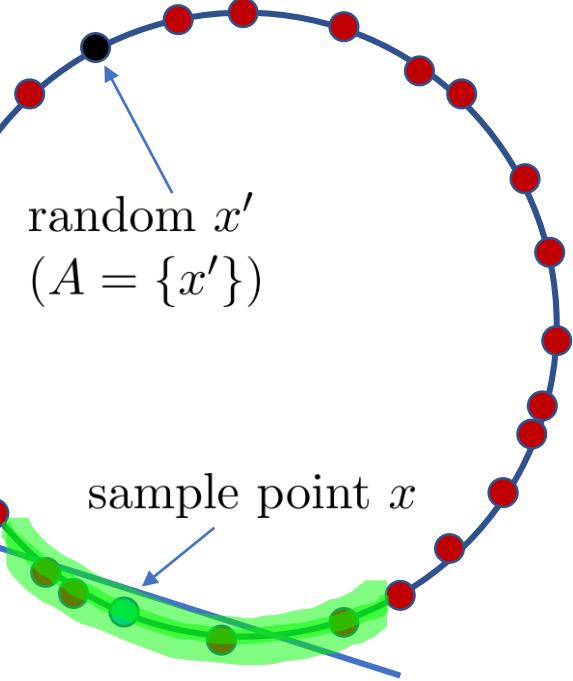
Try $k = 1$

Given sample x

Rand x' probably not close

Can't shatter $\{x, x'\}$
without a lot of points wrong

So won't query x



$\text{DIS}(\mathcal{H}_{x,-1})$ still entire circle (minus x)

$\text{DIS}(\mathcal{H}_{x,+1})$ **small** region

$$\Rightarrow \hat{y}_x = -1$$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S

2. **label** points in $Q = \text{all } x \in S \text{ s.t.}$

$$P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$$

3. **add** the remaining points $x \in S$ to Q with label

$$\hat{y}_x := \underset{y}{\operatorname{argmax}} P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$$

4. **optimize** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$

5. **reduce** \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Generally, need to try various k and pick one
(See the papers)

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label
 $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Generally, need to try various k and pick one
(See the papers)

$$\theta^{(k)} := \sup_{r > \epsilon} \frac{P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A)}{r}$$

$$\tilde{d} := \min \left\{ k : P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A) \xrightarrow[r \rightarrow 0]{} 0 \right\}$$

$$\tilde{\theta} := \theta^{(\tilde{d})}$$

Theorem: For Bounded noise, $R(\hat{f}) \leq R(f^*) + \epsilon$
with $\#$ labels

$$\approx C\tilde{\theta}d \log\left(\frac{1}{\epsilon}\right)$$

Note: $\tilde{\theta} \ll \frac{1}{\epsilon}$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label
 $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Generally, need to try various k and pick one
(See the papers)

$$\theta^{(k)} := \sup_{r > \epsilon} \frac{P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A)}{r}$$

$$\tilde{d} := \min \left\{ k : P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A) \xrightarrow[r \rightarrow 0]{} 0 \right\}$$

$$\tilde{\theta} := \theta^{(\tilde{d})}$$

Theorem: For Bounded noise, $R(\hat{f}) \leq R(f^*) + \epsilon$
with $\#$ labels

$$\approx C\tilde{\theta}d \log\left(\frac{1}{\epsilon}\right)$$

Note: $\tilde{\theta} \ll \frac{1}{\epsilon}$

In the example: $\tilde{\theta} = 2, \theta = \frac{1}{\epsilon}$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label
 $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Generally, need to try various k and pick one
(See the papers)

$$\theta^{(k)} := \sup_{r > \epsilon} \frac{P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A)}{r}$$

$$\tilde{d} := \min \left\{ k : P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A) \xrightarrow[r \rightarrow 0]{} 0 \right\}$$

$$\tilde{\theta} := \theta^{(\tilde{d})}$$

Theorem: For Bounded noise, $R(\hat{f}) \leq R(f^*) + \epsilon$
with $\#$ labels

$$\approx C\tilde{\theta}d \log\left(\frac{1}{\epsilon}\right)$$

Note: $\tilde{\theta} \ll \frac{1}{\epsilon}$ (may depend on f^* , P_X)

In the example: $\tilde{\theta} = 2$, $\theta = \frac{1}{\epsilon}$

Up Next:
Distribution-free Analysis

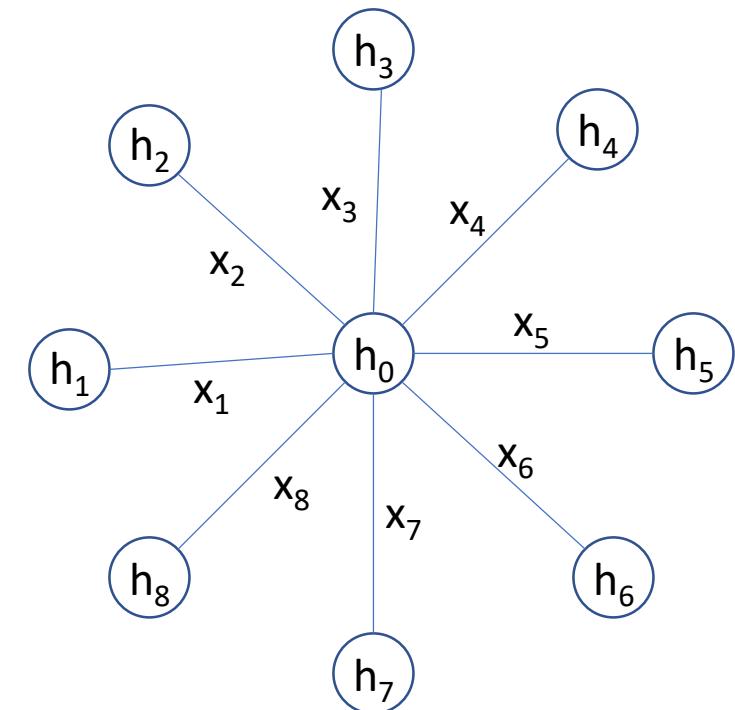
Distribution-Free Analysis

(Hanneke & Yang, 2015)

$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** s is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$,
 $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.



Distribution-Free Analysis

(Hanneke & Yang, 2015)

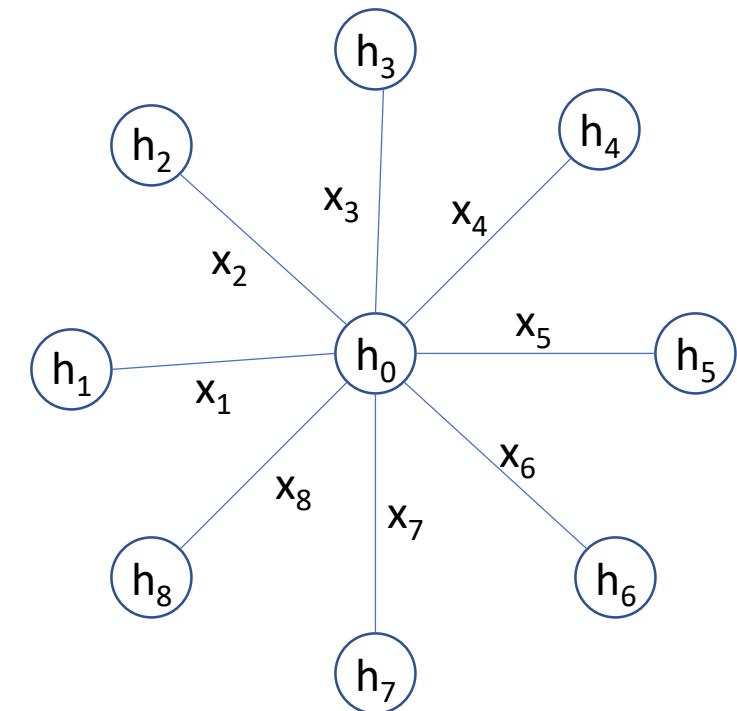
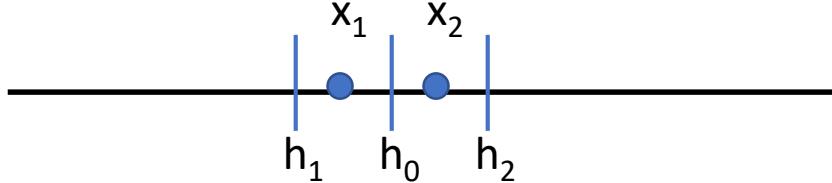
$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** s is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$, $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Example: Thresholds: $f(x) = \mathbb{I}[x \geq t]$.

$$s = 2.$$



Distribution-Free Analysis

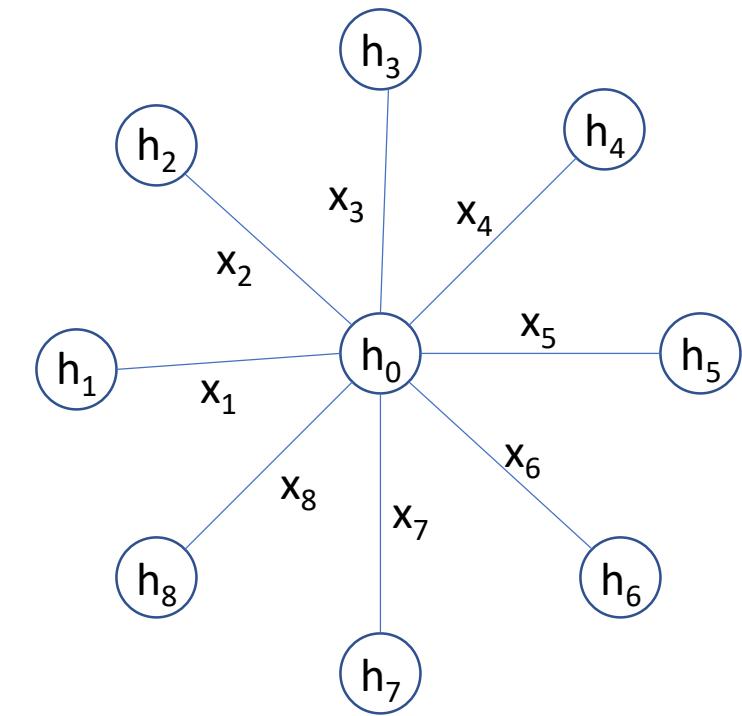
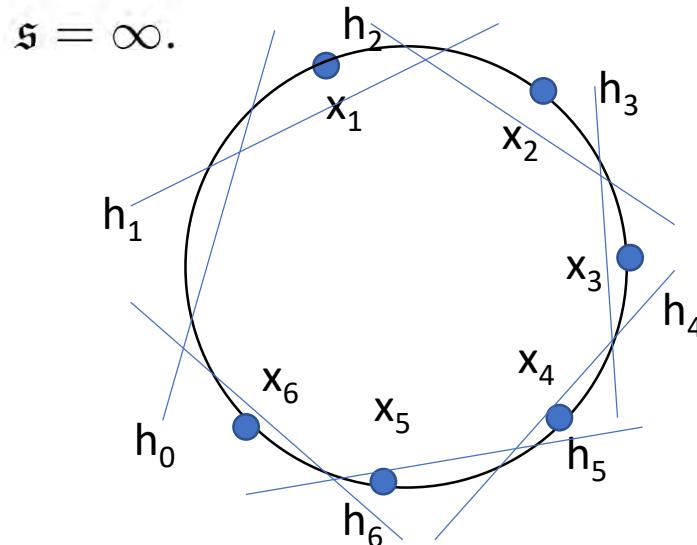
(Hanneke & Yang, 2015)

$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** s is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$, $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Example: Linear Separators in \mathbb{R}^n , $n \geq 2$:



Distribution-Free Analysis

(Hanneke & Yang, 2015)

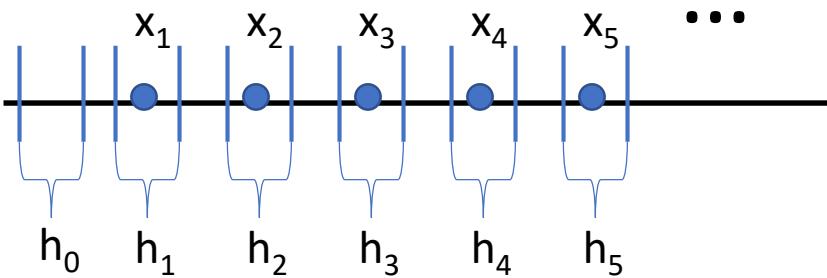
$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

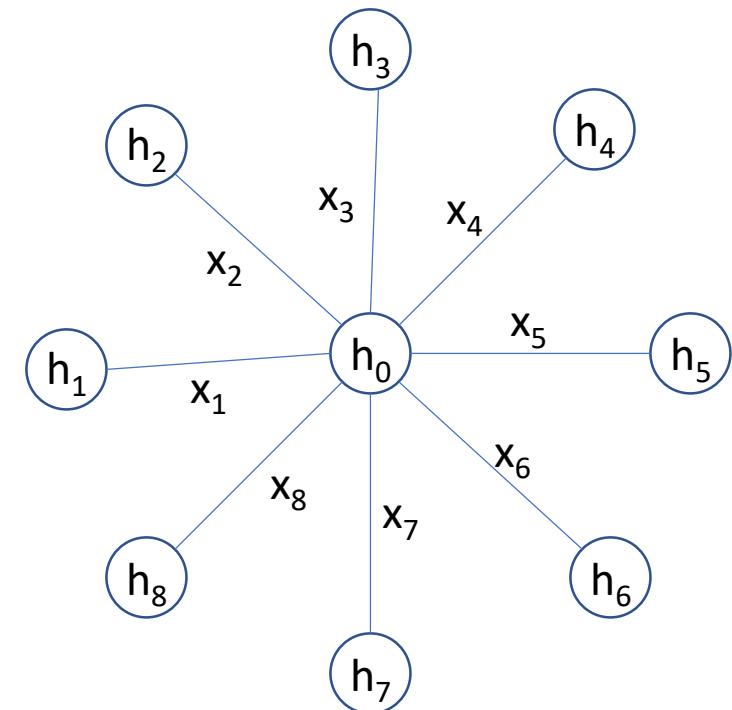
Definition: The **star number** \mathfrak{s} is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$, $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Example: Intervals: $x \mapsto \mathbb{I}[a \leq x \leq b]$

$$\mathfrak{s} = \infty.$$



Intervals of width w ($b - a = w > 0$) on $\mathcal{X} = [0, 1]$: $\mathfrak{s} \approx \lfloor \frac{1}{w} \rfloor$.



Distribution-Free Analysis

(Hanneke & Yang, 2015;
Hanneke, 2016)

$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** \mathfrak{s} is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$,
 $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

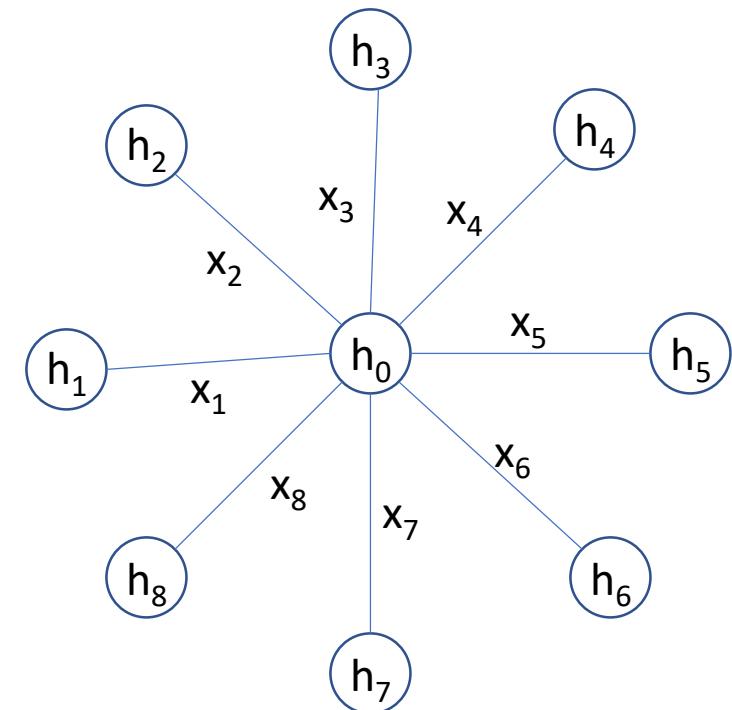
Theorem: $\sup_{P_X} \sup_{f^* \in \mathcal{H}} \theta = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \varphi_c = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \tilde{\theta} = \min\{\mathfrak{s}, \frac{1}{\epsilon}\} =: \mathfrak{s}_\epsilon$

Corollary:

Bounded noise # labels $\approx \mathfrak{s}_\epsilon d \log(\frac{1}{\epsilon})$

Agnostic ($\beta = R(f^*)$) # labels $\approx \mathfrak{s}_\beta d \frac{\beta^2}{\epsilon^2}$

Achieved by A^2



Distribution-Free Analysis

(Hanneke & Yang, 2015;
Hanneke, 2016)

$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** \mathfrak{s} is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$,
 $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Theorem: $\sup_{P_X} \sup_{f^* \in \mathcal{H}} \theta = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \varphi_c = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \tilde{\theta} = \min\{\mathfrak{s}, \frac{1}{\epsilon}\} =: \mathfrak{s}_\epsilon$

Corollary:

Bounded noise # labels $\approx \mathfrak{s}_\epsilon d \log(\frac{1}{\epsilon})$

Agnostic ($\beta = R(f^*)$) # labels $\approx \mathfrak{s}_\beta d \frac{\beta^2}{\epsilon^2}$

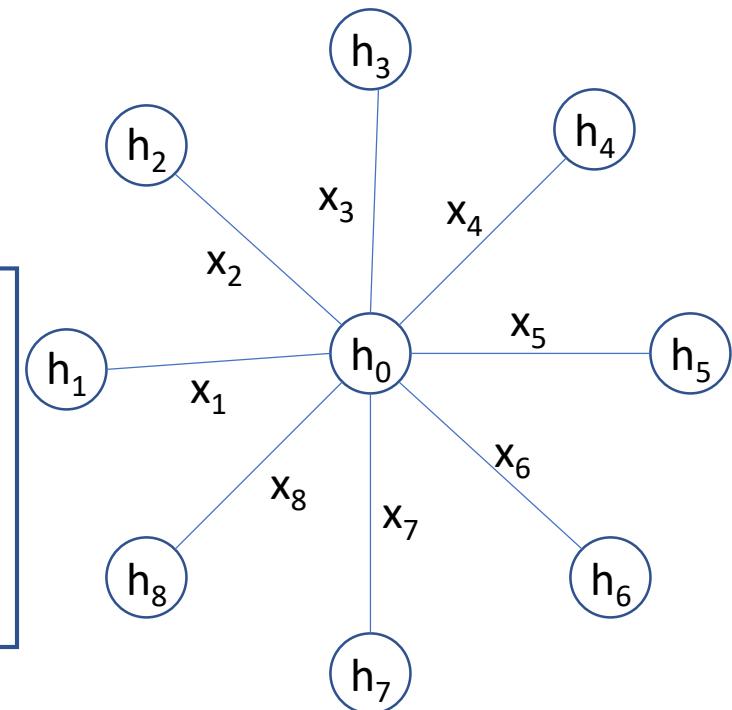
Achieved by A^2

Different alg., Bounded noise

labels $\approx \mathfrak{s}_{\epsilon/d} \log(\frac{1}{\epsilon})$

Near-matching **lower bound**:

$\mathfrak{s}_\epsilon + d \log(\frac{1}{\epsilon})$



Distribution-Free Analysis

(Hanneke & Yang, 2015;
Hanneke, 2016)

$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** \mathfrak{s} is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$,
 $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Theorem: $\sup_{P_X} \sup_{f^* \in \mathcal{H}} \theta = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \varphi_c = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \tilde{\theta} = \min\{\mathfrak{s}, \frac{1}{\epsilon}\} =: \mathfrak{s}_\epsilon$

Corollary:

Bounded noise # labels $\approx \mathfrak{s}_\epsilon d \log(\frac{1}{\epsilon})$

Agnostic ($\beta = R(f^*)$) # labels $\approx \mathfrak{s}_\beta d \frac{\beta^2}{\epsilon^2}$

Achieved by A^2

Different alg., Bounded noise

labels $\approx \mathfrak{s}_{\epsilon/d} \log(\frac{1}{\epsilon})$

Near-matching **lower bound**:

$\mathfrak{s}_\epsilon + d \log(\frac{1}{\epsilon})$

Open Question:

Agnostic ($\beta = R(f^*)$)

labels

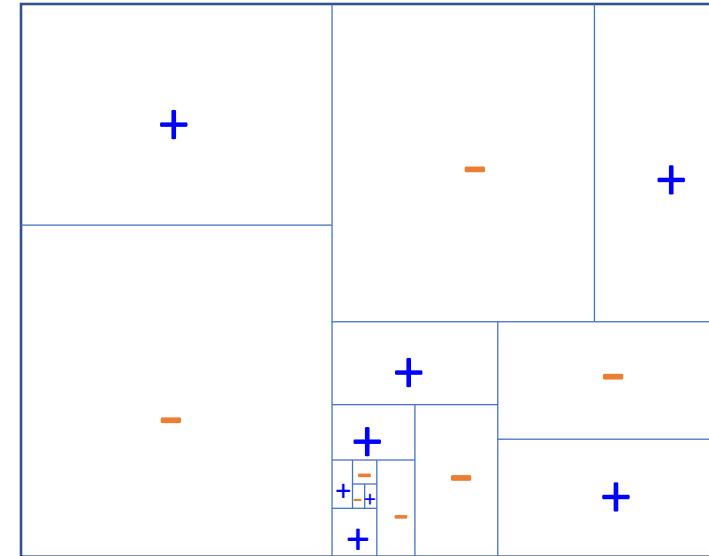
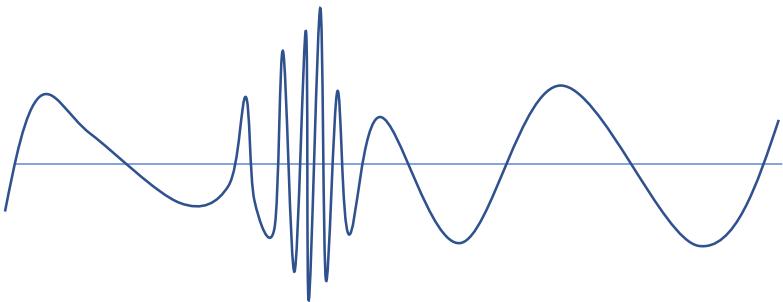
$\approx d \frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon/d} \log(\frac{1}{\epsilon})$?

lower bound:

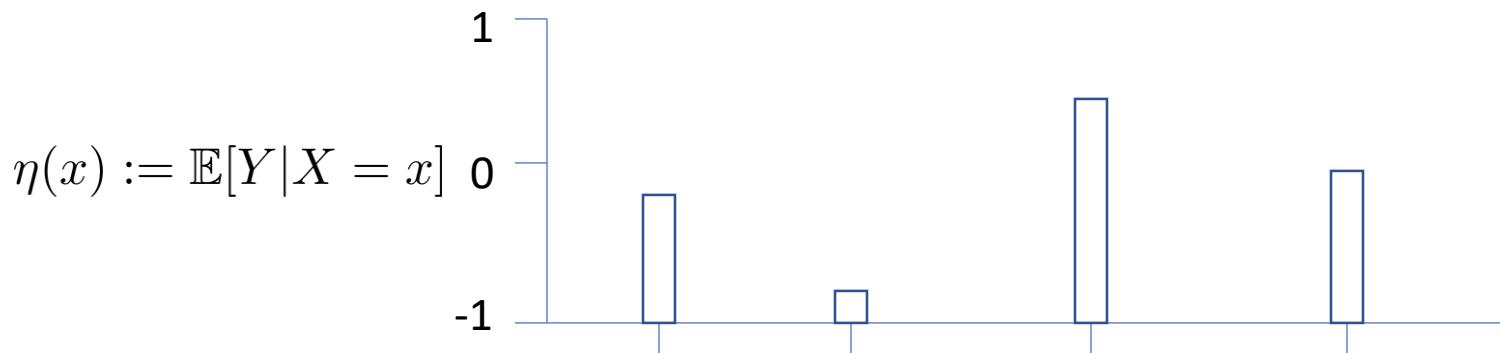
$d \frac{\beta^2}{\epsilon^2} + \mathfrak{s}_\epsilon + d \log(\frac{1}{\epsilon})$

Adapting to Heterogeneous Noise

So far: Active learning for spatial heterogeneity of **opt function**:



Also consider: Spatial heterogeneity of **noise**:



Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$

Input: Label budget n

Output: Classifier \hat{f}_n .

1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \dots$
3. $X_{s_m} \leftarrow \text{GETSEED}(\mathbb{L}, m)$
4. $\mathcal{L}_m \leftarrow \text{TICTOC}(X_{s_m}, m)$
5. if \mathcal{L}_m exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. If we've made n queries
7. Return $\hat{f}_n \leftarrow \text{LEARN}(\mathbb{L})$

An active learning alg.
(e.g. A^2)

Main new part

A passive learning alg.

Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$

Input: Label budget n

Output: Classifier \hat{f}_n .

1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \dots$
3. $X_{s_m} \leftarrow \text{GETSEED}(\mathbb{L}, m)$
4. $\mathcal{L}_m \leftarrow \text{TICTOC}(X_{s_m}, m)$
5. if \mathcal{L}_m exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. If we've made n queries
7. Return $\hat{f}_n \leftarrow \text{LEARN}(\mathbb{L})$

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Suppose f^* is the **global** optimal function: $f^*(x) = \text{sign}(\eta(x))$

TICTOC(X, m):

Query X (or nearby) to try to guess $f^*(X)$

If can figure it out, return that label

If can't figure it out by τ_m queries give up (don't return a label)

Focus queries on less-noisy points.

Double advantage:

- Focusing on the points we actually care about:

$$R(f|x) - R(f^*|x) = |\eta(x)|\mathbb{I}[f(x) \neq f^*(x)]$$

(small $|\eta(x)| \Rightarrow$ not much effect on $R(f|x)$ if $f(x) = f^*(x)$ or not).

- And those points require fewer queries to determine $f^*(X_i)$!

$\sim \frac{1}{\eta(X_i)^2}$ queries
to determine $f^*(X_i)$.

Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$

Input: Label budget n

Output: Classifier \hat{f}_n .

1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \dots$
3. $X_{s_m} \leftarrow \text{GETSEED}(\mathbb{L}, m)$
4. $\mathcal{L}_m \leftarrow \text{TICTOC}(X_{s_m}, m)$
5. if \mathcal{L}_m exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. If we've made n queries
7. Return $\hat{f}_n \leftarrow \text{LEARN}(\mathbb{L})$

Theorem: Bounded noise: # labels

$$\approx \mathfrak{s}_{\epsilon/d} \log\left(\frac{1}{\epsilon}\right)$$

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Suppose f^* is the **global** optimal function: $f^*(x) = \text{sign}(\eta(x))$

TICTOC(X, m):

Query X (or nearby) to try to guess $f^*(X)$

If can figure it out, return that label

If can't figure it out by τ_m queries give up (don't return a label)

Focus queries on less-noisy points.

Double advantage:

- Focusing on the points we actually care about:

$$R(f|x) - R(f^*|x) = |\eta(x)| \mathbb{I}[f(x) \neq f^*(x)]$$

(small $|\eta(x)| \Rightarrow$ not much effect on $R(f|x)$ if $f(x) = f^*(x)$ or not).

- And those points require fewer queries to determine $f^*(X_i)$!

$\sim \frac{1}{\eta(X_i)^2}$ queries
to determine $f^*(X_i)$.

Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$

Input: Label budget n

Output: Classifier \hat{f}_n .

1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \dots$
3. $X_{s_m} \leftarrow \text{GETSEED}(\mathbb{L}, m)$
4. $\mathcal{L}_m \leftarrow \text{TICTOC}(X_{s_m}, m)$
5. if \mathcal{L}_m exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. If we've made n queries
7. Return $\hat{f}_n \leftarrow \text{LEARN}(\mathbb{L})$

Theorem: Agnostic ($\beta = R(f^*)$)

and suppose f^* = global best:

labels

$$\approx d \frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon/d} \log\left(\frac{1}{\epsilon}\right)$$

Confirms agnostic sample complexity conjecture
but with extra assumption f^* = global opt.

Near-match lower bound: $d \frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon} + d \log\left(\frac{1}{\epsilon}\right)$

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Suppose f^* is the **global** optimal function: $f^*(x) = \text{sign}(\eta(x))$

TICTOC(X, m):

Query X (or nearby) to try to guess $f^*(X)$

If can figure it out, return that label

If can't figure it out by τ_m queries give up (don't return a label)

Focus queries on less-noisy points.

Double advantage:

- Focusing on the points we actually care about:

$$R(f|x) - R(f^*|x) = |\eta(x)| \mathbb{I}[f(x) \neq f^*(x)]$$

(small $|\eta(x)| \Rightarrow$ not much effect on $R(f|x)$ if $f(x) = f^*(x)$ or not).

- And those points require fewer queries to determine $f^*(X_i)$!

$\sim \frac{1}{\eta(X_i)^2}$ queries
to determine $f^*(X_i)$.

Principles of Active Learning

1. Query in dense regions where \hat{f} could disagree a lot with f^*
2. Query in regions with low noise

Tsybakov Noise

The alg. adapts to heterogeneity in the noise.

Let's try it with a model that explicitly describes heterogeneous noise:

Tsybakov Noise

Tsybakov Noise

(Tsybakov, 2004;
Mammen & Tsybakov 1999)

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Definition: (Tsybakov noise)

$f^*(x) = \text{sign}(\eta(x))$ and $\exists \alpha \in (0, 1)$ s.t. $\forall \tau > 0$,

$$P_X(x : |\eta(x)| \leq \tau) \lesssim \tau^{\frac{\alpha}{1-\alpha}}.$$

Tsybakov Noise

(Tsybakov, 2004;
Mammen & Tsybakov 1999)

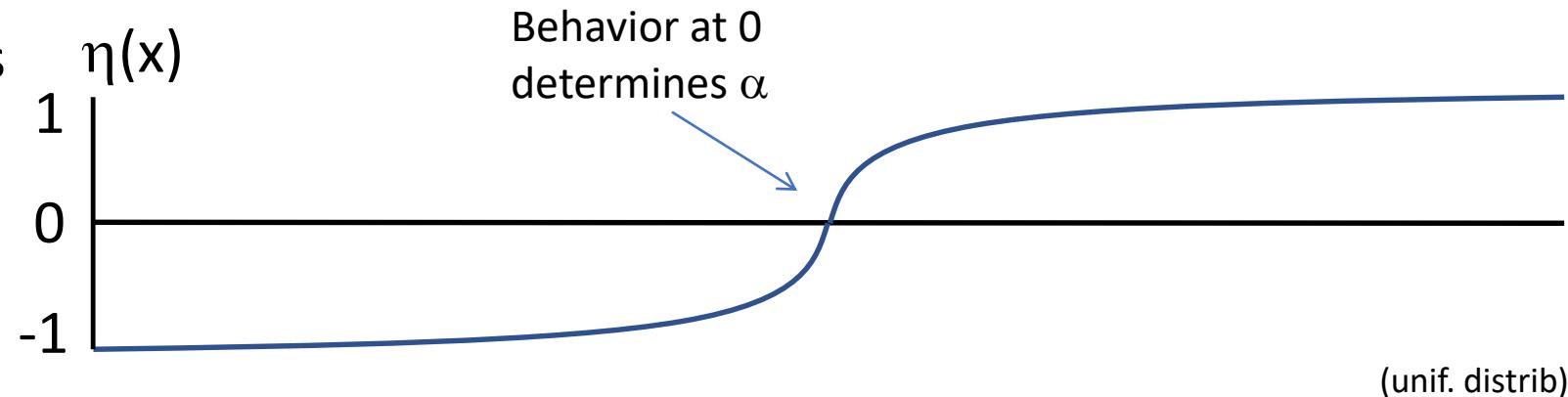
Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Definition: (Tsybakov noise)

$f^*(x) = \text{sign}(\eta(x))$ and $\exists \alpha \in (0, 1)$ s.t. $\forall \tau > 0$,
 $P_X(x : |\eta(x)| \leq \tau) \lesssim \tau^{\frac{\alpha}{1-\alpha}}$.

Example:

Thresholds



Tsybakov Noise

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Definition: (Tsybakov noise)

$f^*(x) = \text{sign}(\eta(x))$ and $\exists \alpha \in (0, 1)$ s.t. $\forall \tau > 0$,

$$P_X(x : |\eta(x)| \leq \tau) \lesssim \tau^{\frac{\alpha}{1-\alpha}}.$$

Passive OPT: $\tilde{\Theta}\left(\frac{d}{\epsilon^{2-\alpha}}\right)$.

(Massart & Nédélec, 2006)

Active OPT: $\begin{cases} \frac{d}{\epsilon^{2-2\alpha}} & \text{if } 0 < \alpha \leq 1/2 \\ \min \left\{ \frac{d}{\epsilon^{2-2\alpha}} \left(\frac{s}{d} \right)^{2\alpha-1}, \frac{d}{\epsilon} \right\} & \text{if } 1/2 < \alpha < 1 \end{cases}$

(Hanneke & Yang, 2015)

$$\sim \begin{cases} \frac{1}{\epsilon^{2-2\alpha}}, & \text{if } s < \infty \\ \frac{1}{\epsilon}, & \text{if } s = \infty \end{cases}.$$

Active Opt \ll **Passive Opt.**
(always)

Conclusions

- Many proposals for going beyond Disagreement-based Active Learning
- Each exhibits improvements in certain cases
- We still don't know the **optimal agnostic active learning algorithm**

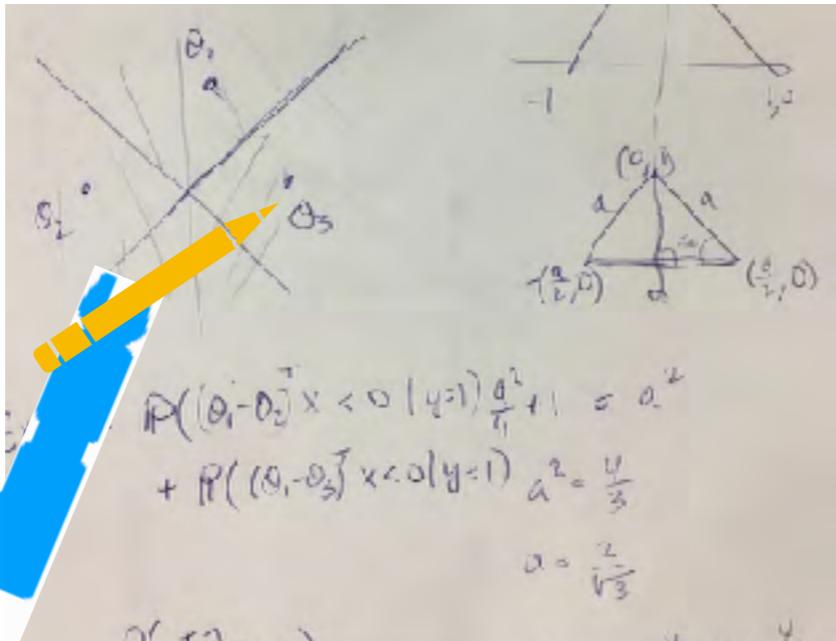
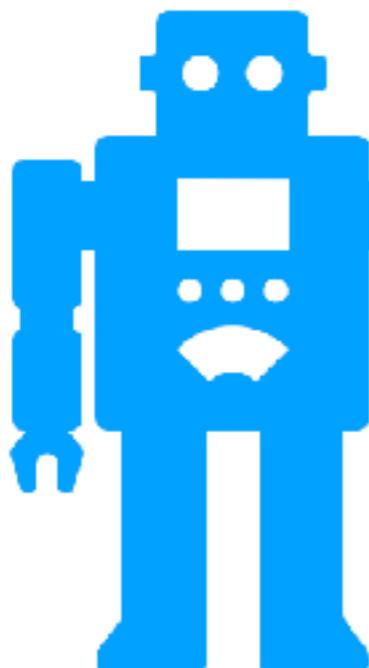
$$d \frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon/d} \log\left(\frac{1}{\epsilon}\right)$$

Questions?

Further reading:

- S. Dasgupta, A. Kalai, C. Monteleoni. Analysis of perceptron-based active learning. *COLT* 2005.
- M. F. Balcan, A. Broder, T. Zhang. Margin based active learning. *COLT* 2007.
- P. Awasthi, M. F. Balcan, P. Long. *Journal of the ACM*, 2017.
- S. Hanneke. Theoretical Foundations of Active Learning. PhD Thesis, CMU, 2009.
- S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 2012.
- C. Zhang, K. Chaudhuri. Beyond disagreement-based agnostic active learning. *NeurIPS* 2014.
- R. M. Castro, R. D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 2008.
- R. M. Castro, R.D. Nowak. Upper and lower error bounds for active learning. *Allerton* 2006.
- S. Dasgupta. Coarse sample complexity bounds for active learning. *NeurIPS* 2005.
- S. Hanneke, L. Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 2015.
- S. Hanneke. Refined error bounds for several learning algorithms. *Journal of Machine Learning Research*, 2016.
- M. F. Balcan, S. Hanneke, J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 2010.

Active Learning from Theory to Practice



Steve Hanneke
Toyota Technological
Institute at Chicago
steve.hanneke@gmail.com

Robert Nowak
UW-Madison
rdnowak@wisc.edu

ICML | 2019

Thirty-sixth International Conference on
Machine Learning

Tutorial Outline



Part 1: Introduction to Active Learning (Rob)

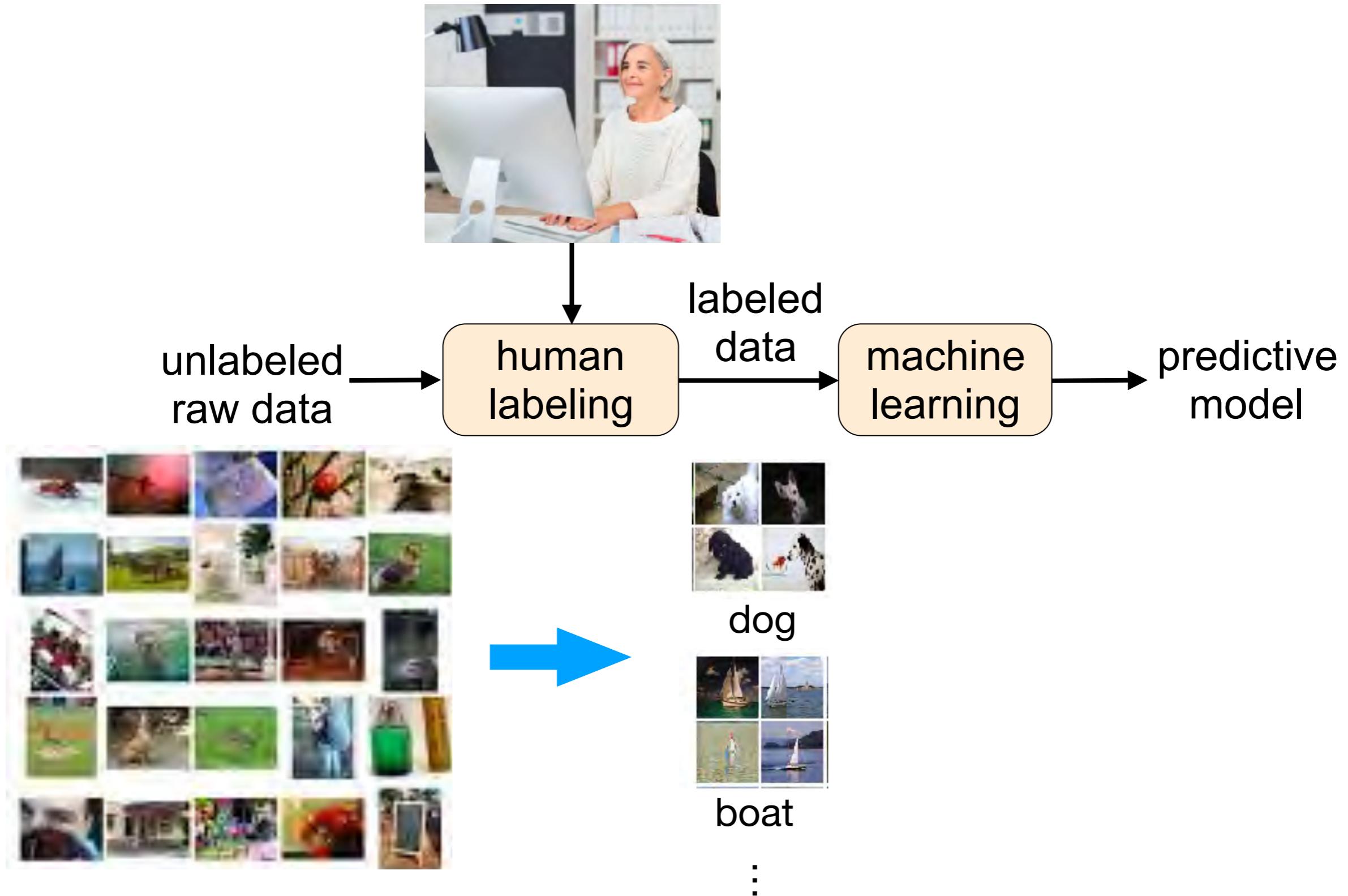
Part 2: Theory of Active Learning (Steve)

Part 3: Advanced Topics and Open Problems (Steve)

Part 4: Nonparametric Active Learning (Rob)

slides: <http://nowak.ece.wisc.edu/ActiveML.html>

Conventional (Passive) Machine Learning



ALL SYSTEMS GO

?

the guardian

Computers now better than humans at
recognising and sorting images

millions of labeled images
1000's of human hours

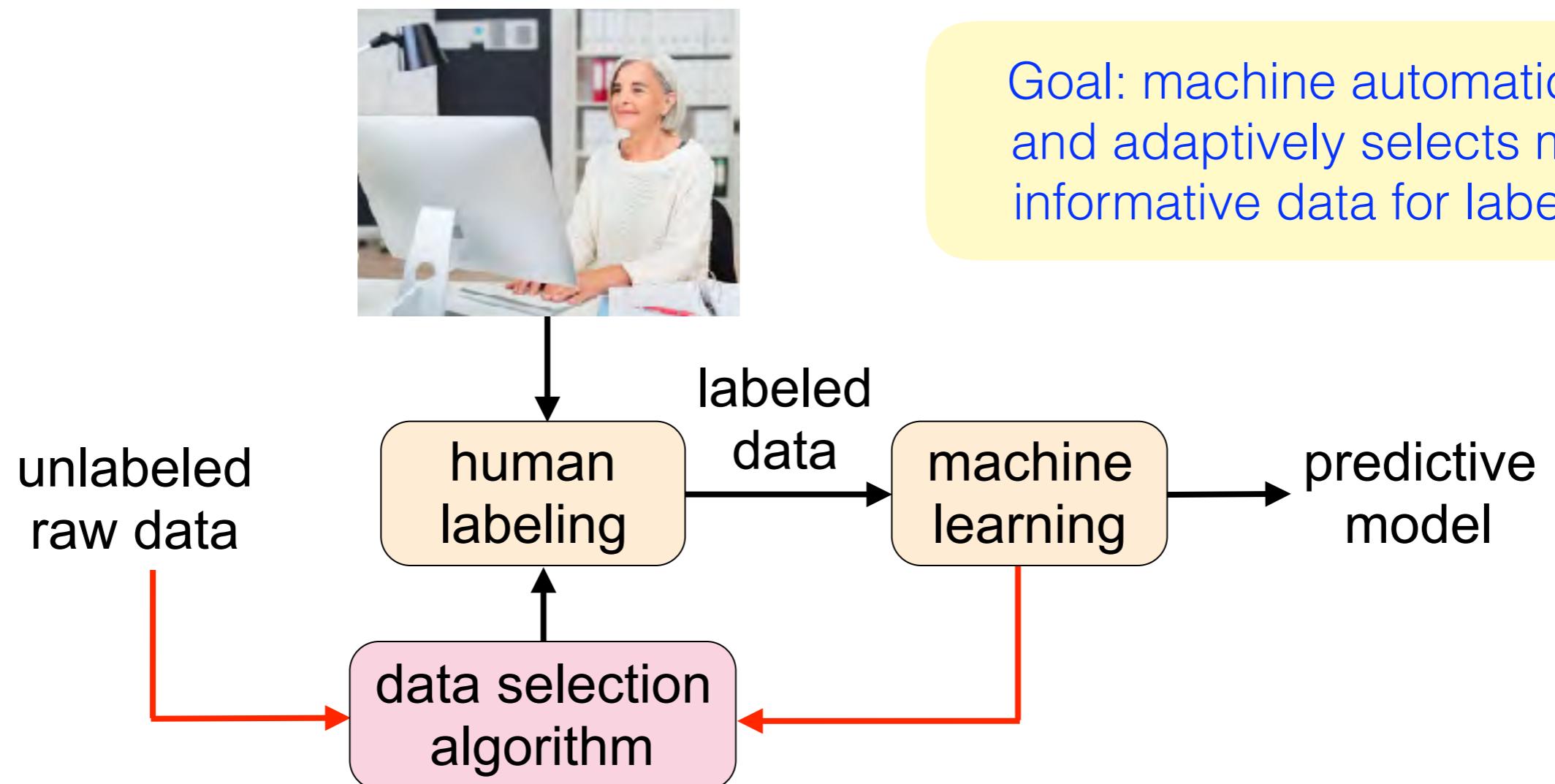
QUARTZ

**Google says its new AI-powered
translation tool scores nearly identically to
human translators**

trained on more texts than a
human could read in a lifetime

Can we train machines with less labeled
data and less human supervision?

Active Machine Learning



Motivating Application



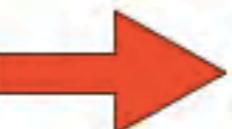
unlabeled electronic
health records (EHRs)



prediction rule
that can be applied
to unlabeled EHRs



human experts



cataracts

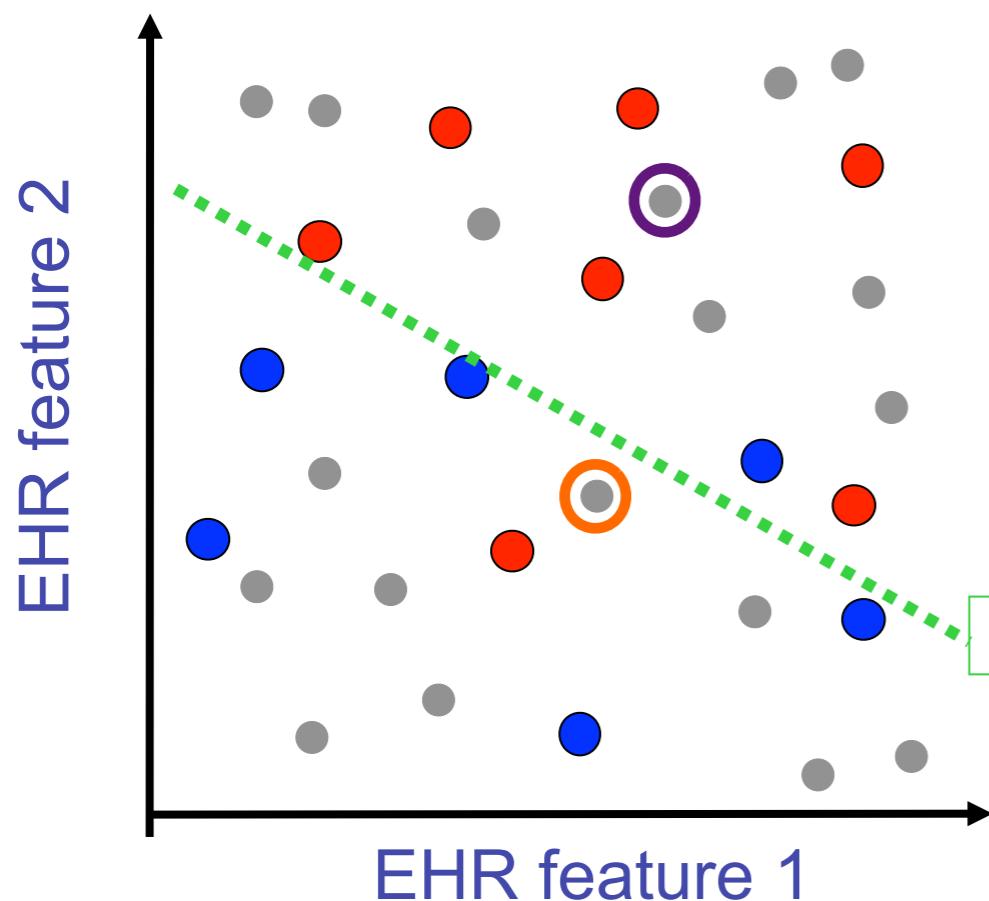


healthy



provides labels to machine learner
(several minutes / EHR)

Active Learning

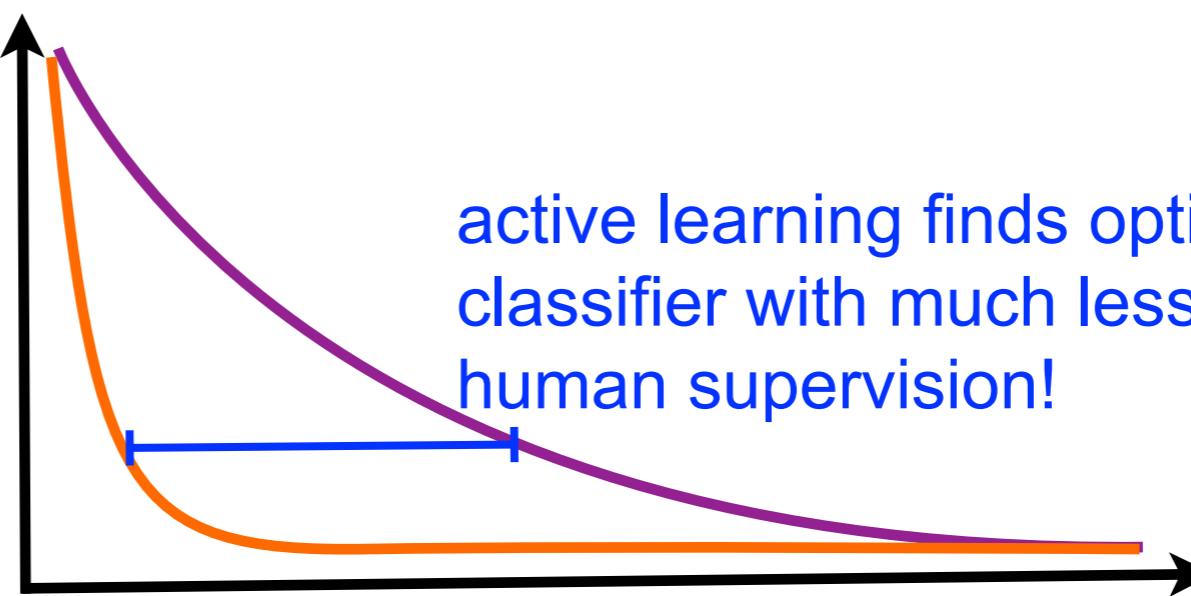


Non-adaptive strategy: Label a random sample

Active strategy: Label a sample near best decision boundary based on labels seen so far

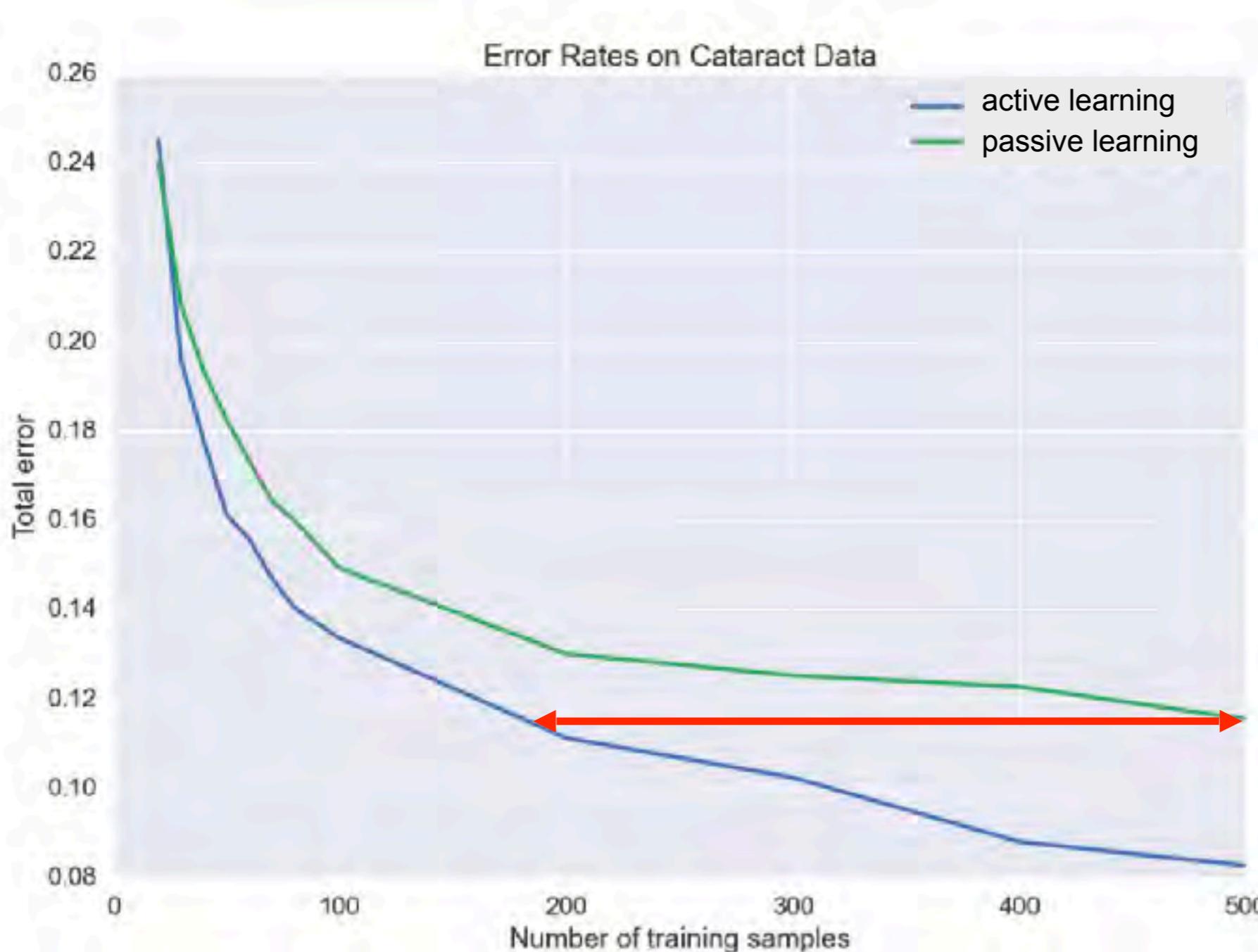
best linear classifier

error rate ϵ



labels

Active Logistic Regression



11000 patient records
8000 positive
3000 negative

6182 Numerical Features
icd9 codes
lab tests
patient data

Classification task:
cataracts or healthy

less than half as many labeled examples needed by active learning

NEXT
ASK BETTER QUESTIONS.
GET BETTER RESULTS.
FASTER. AUTOMATED.



GitHub



Paper



Docs



Blog



Team



Data

Active learning to optimize crowdsourcing and rating in New Yorker Cartoon Caption Contest



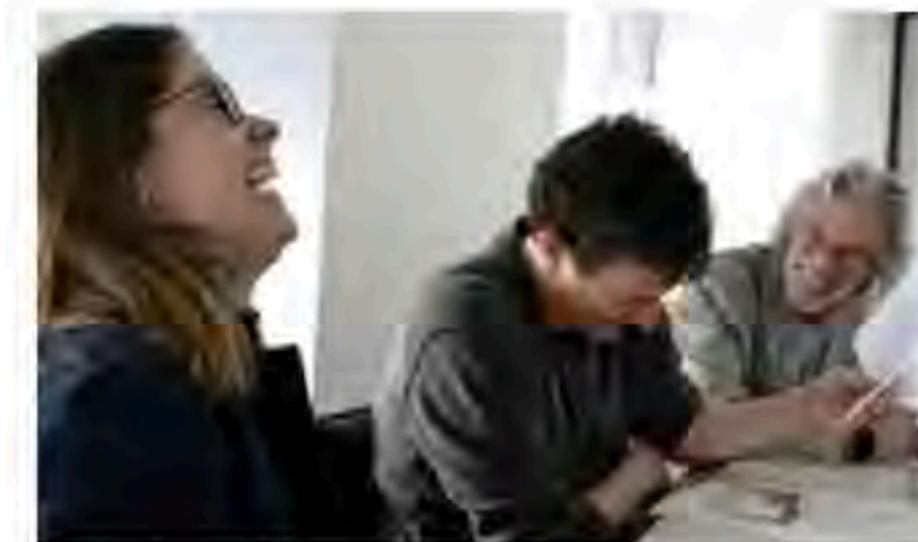
digg

BY DOING THE EXACT OPPOSITE

How New Yorker Cartoons Could Teach Computers To Be Funny

3 diggs CNET Technology

With the help of computer scientists from the University of Wisconsin at Madison, The New Yorker for the first time is using crowdsourcing algorithms to uncover the best captions.



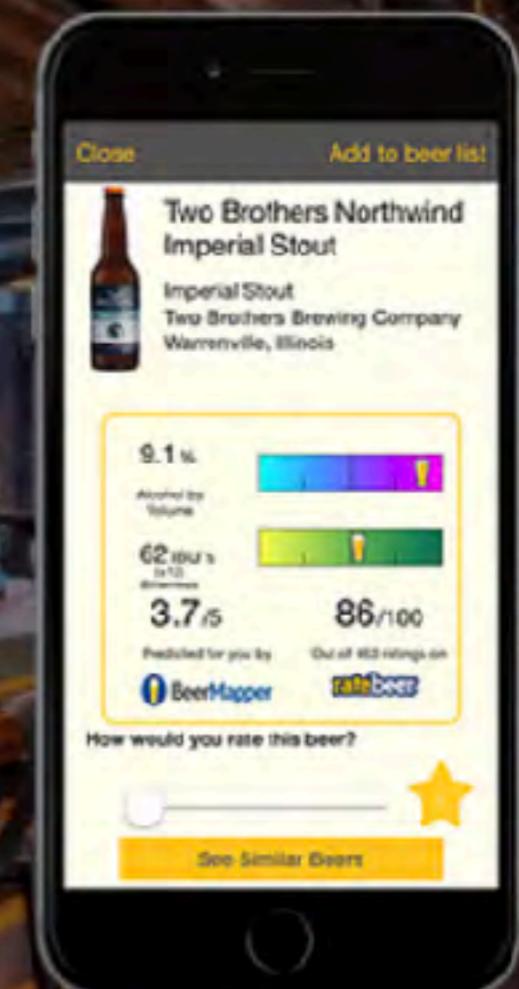
Actively learning user's beer preferences



BeerMapperSM

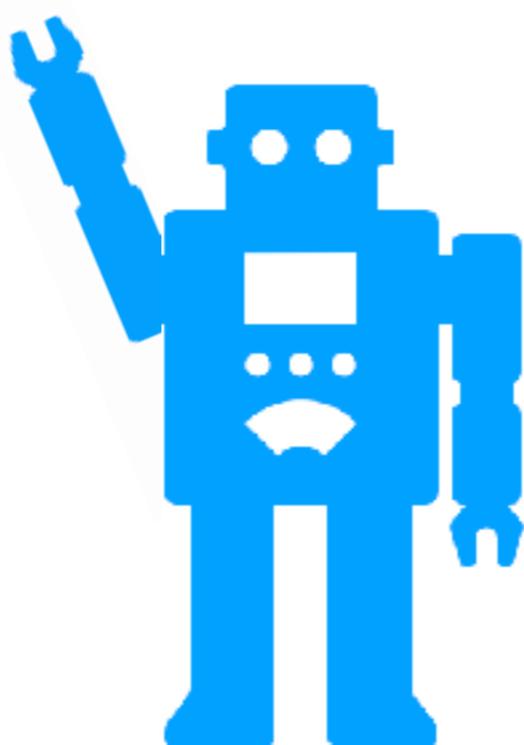
[Home](#)[Contact](#)[About](#)[FAQs](#)

Discover better beer.



The most powerful beer app on the planet.

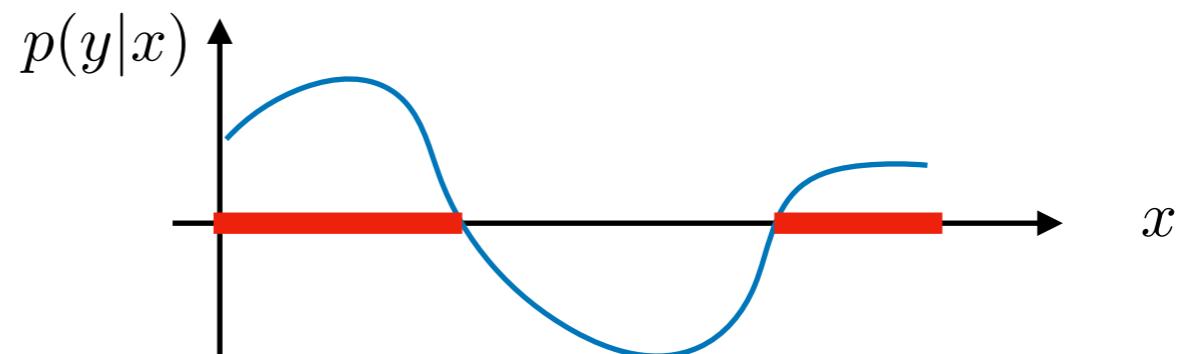
Principles of Active Learning



What and Where Information

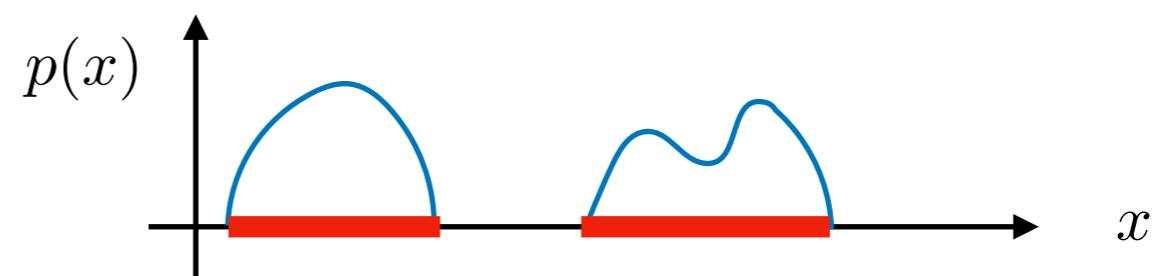
Density estimation: What is $p(y|x)$?

Classification: Where is $p(y|x) > 0$?



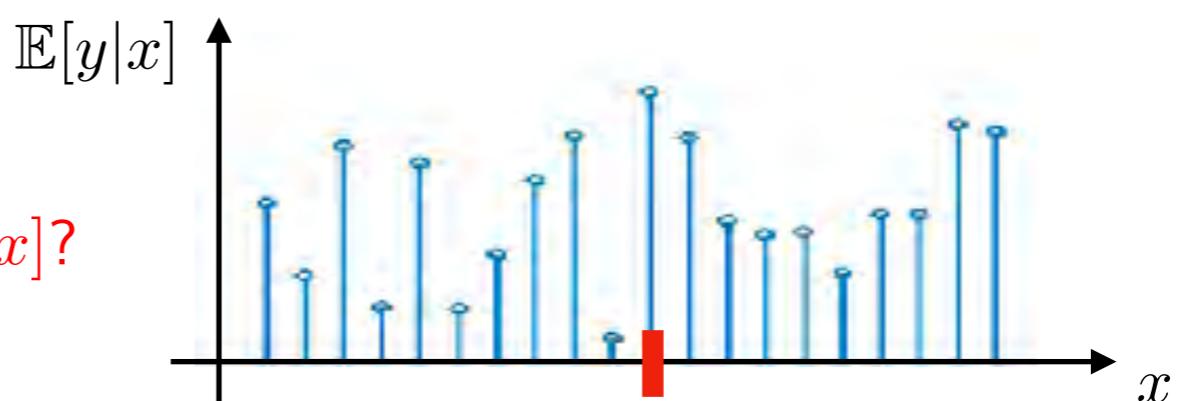
Density estimation: What is $p(x)$?

Clustering: Where is $p(x) > \epsilon$?



Function estimation: What is $\mathbb{E}[y|x]$?

Bandit optimization: Where is $\max_x \mathbb{E}[y|x]$?



Active learning is more efficient than passive learning for localized “where” information

Meta-Algorithm for Active Learning

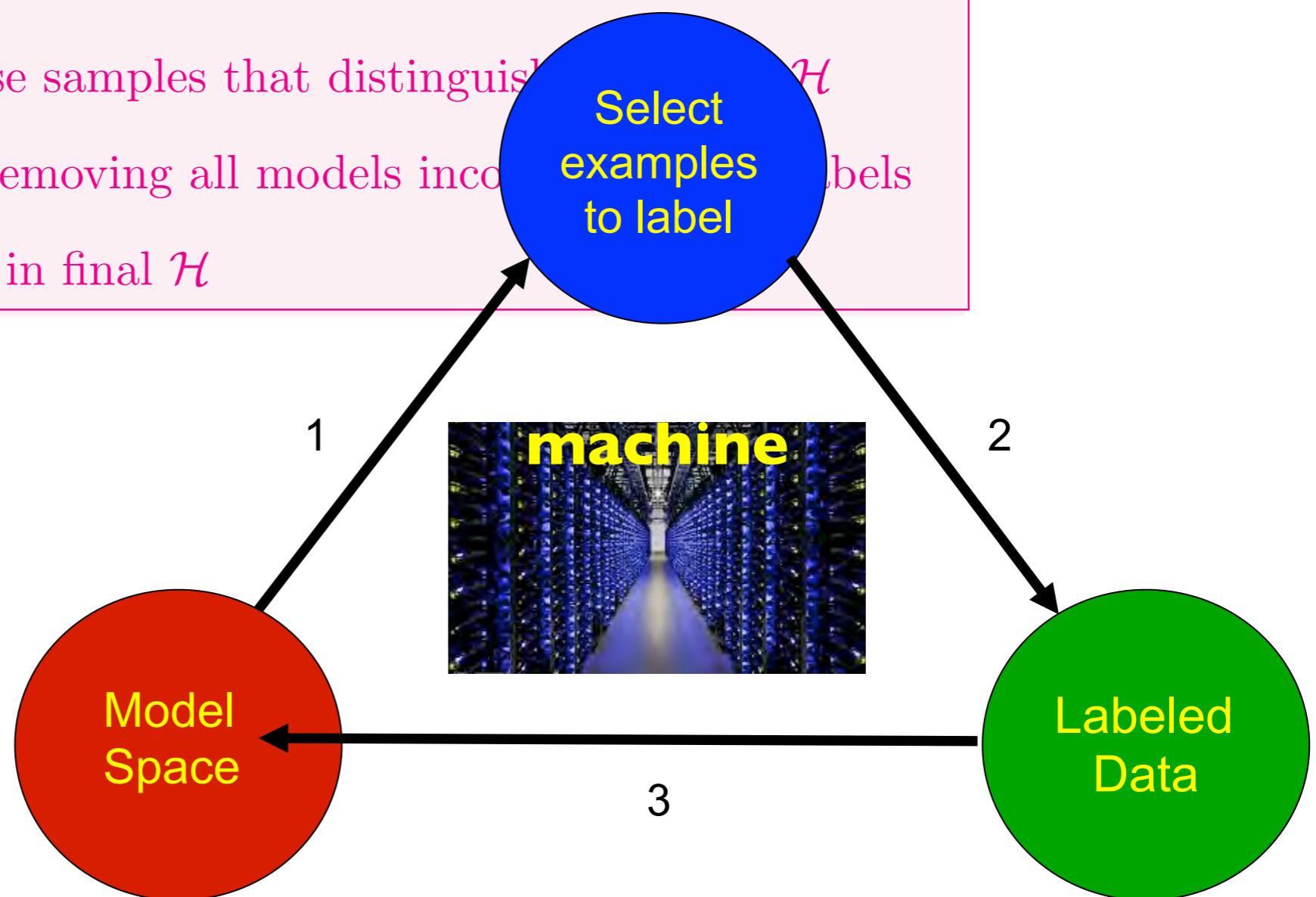
Version-Space (VS) Active Learning

initialize VS: \mathcal{H} = all models/hypotheses

while (*stopping-criterion*) not met

1. sample at random from available dataset
2. label only those samples that distinguish \mathcal{H}
3. reduce \mathcal{H} by removing all models inconsistent with labels

output: best model in final \mathcal{H}



Learning a 1-D Classifier



binary search quickly finds **decision boundary**

passive : err $\sim n^{-1}$

active : err $\sim 2^{-n}$

Vapnik-Chervonenkis (VC) Theory

Given training data $\{(x_j, y_j)\}_{j=1}^n$, learn a function f to predict y from x

Consider a possibly infinite set of hypotheses \mathcal{F} with *finite VC dimension* d and for each $f \in \mathcal{F}$ define the risk (error rate):

$$R(f) := \mathbb{P}(f(x) \neq y)$$

error rate on
training data:

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}(f(x_i) \neq y_i) \quad \text{"empirical risk"}$$

VC bound:

$$\sup_{f \in \mathcal{F}} |R(f) - \widehat{R}(f)| \leq 6 \sqrt{\frac{d \log(n/\delta)}{n}}$$

w.p. $\geq 1 - \delta$

Empirical Risk Minimization (ERM)

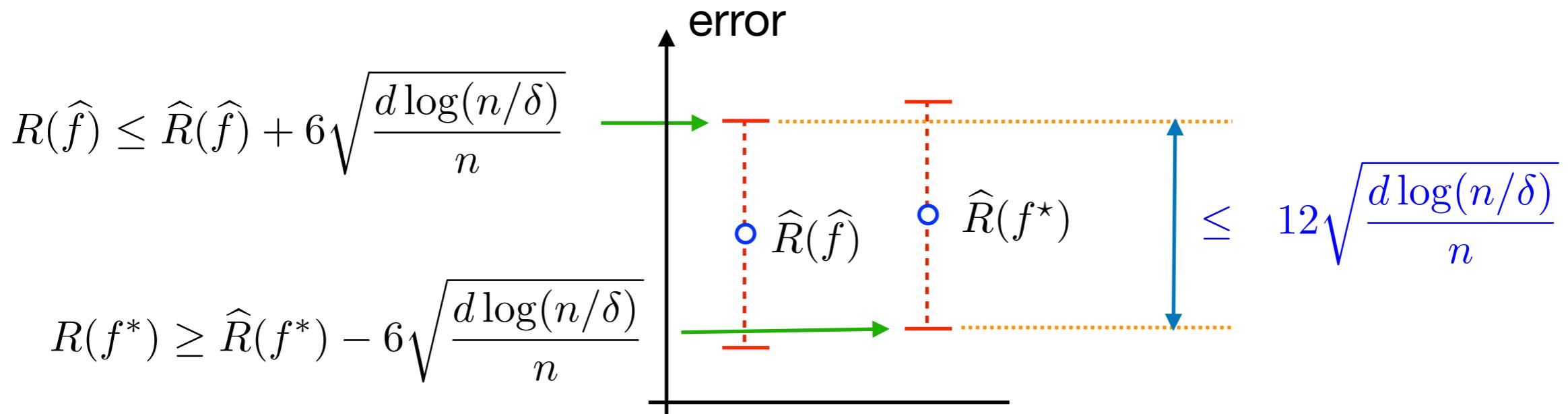
Goal: select hypothesis with true error rate within $\epsilon > 0$ of $\min_{f \in \mathcal{F}} R(f)$

$$f^* = \arg \min_{f \in \mathcal{F}} R(f) \quad \text{true risk minimizer}$$

\hat{f} minimizes empirical risk:

$$\hat{f} = \arg \min_{f \in \mathcal{F}} \hat{R}(f) \quad \text{empirical risk minimizer}$$

$$\hat{R}(\hat{f}) \leq \hat{R}(f^*)$$

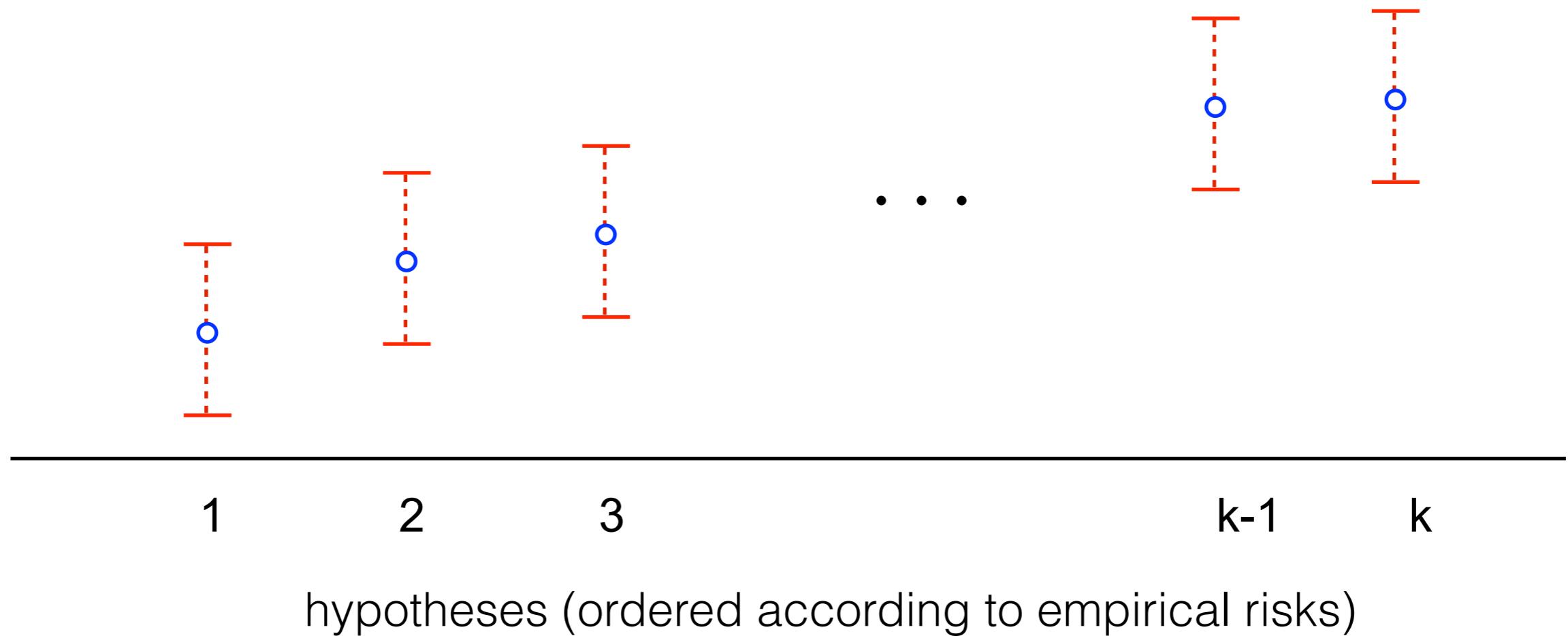


sufficient number
of training examples:

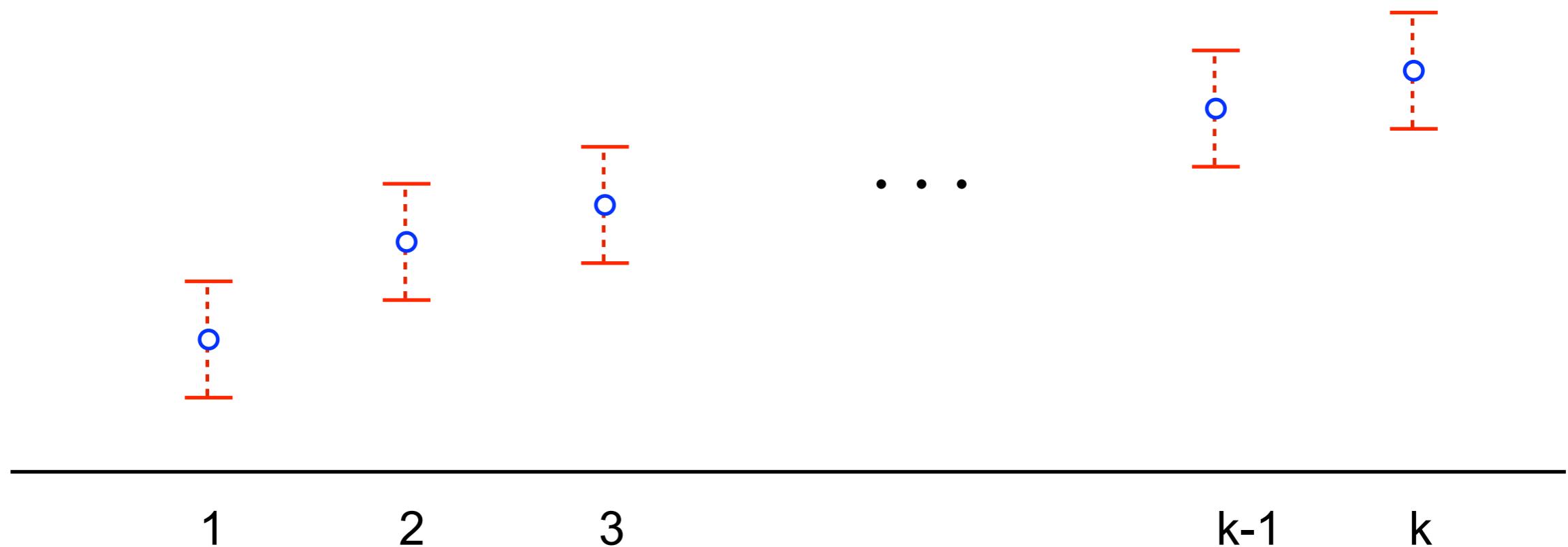
$$12\sqrt{\frac{d \log(n/\delta)}{n}} \leq \epsilon \quad \rightarrow$$

$$n = \tilde{O}\left(\frac{d \log(1/\delta)}{\epsilon^2}\right)$$

Empirical Risks and Confidence Intervals



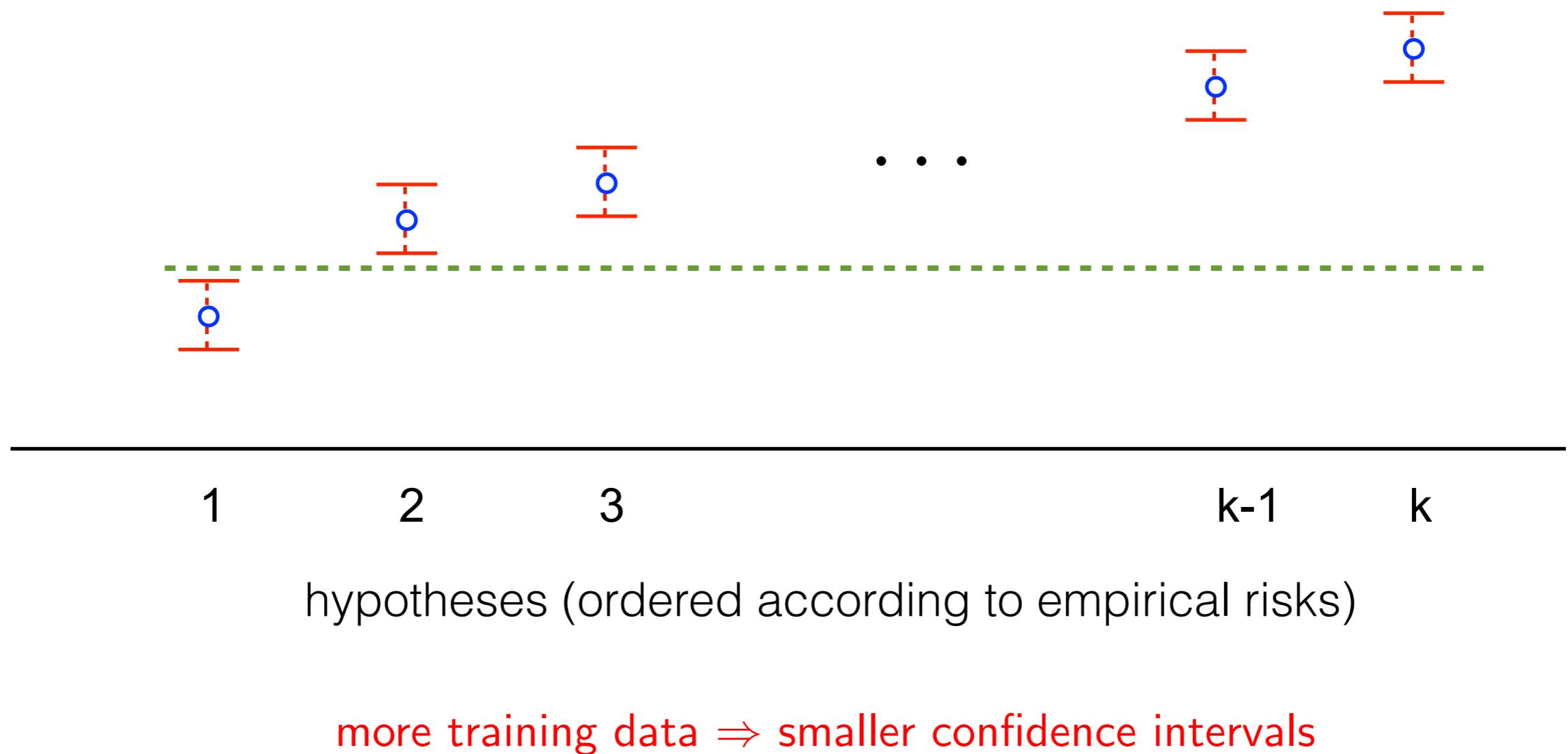
Empirical Risks and Confidence Intervals



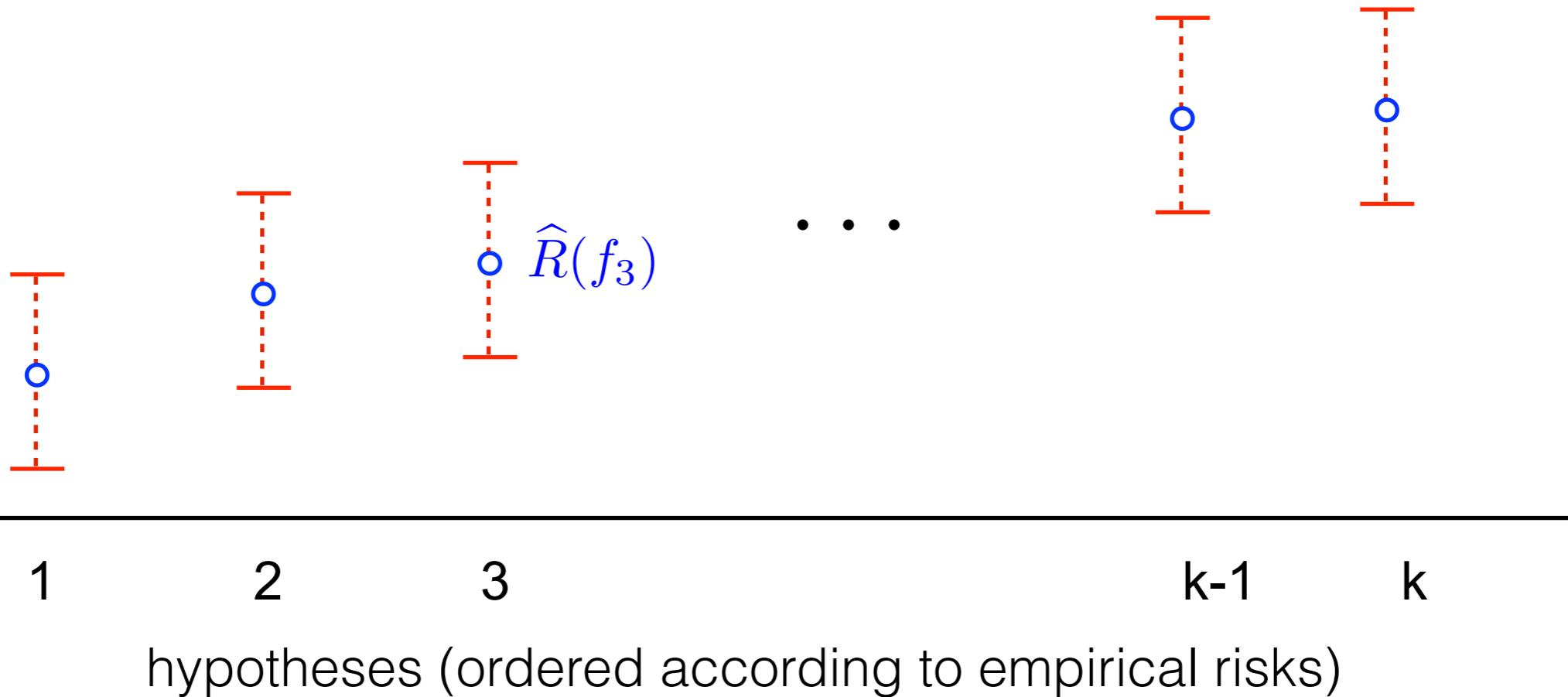
hypotheses (ordered according to empirical risks)

more training data \Rightarrow smaller confidence intervals

Empirical Risks and Confidence Intervals

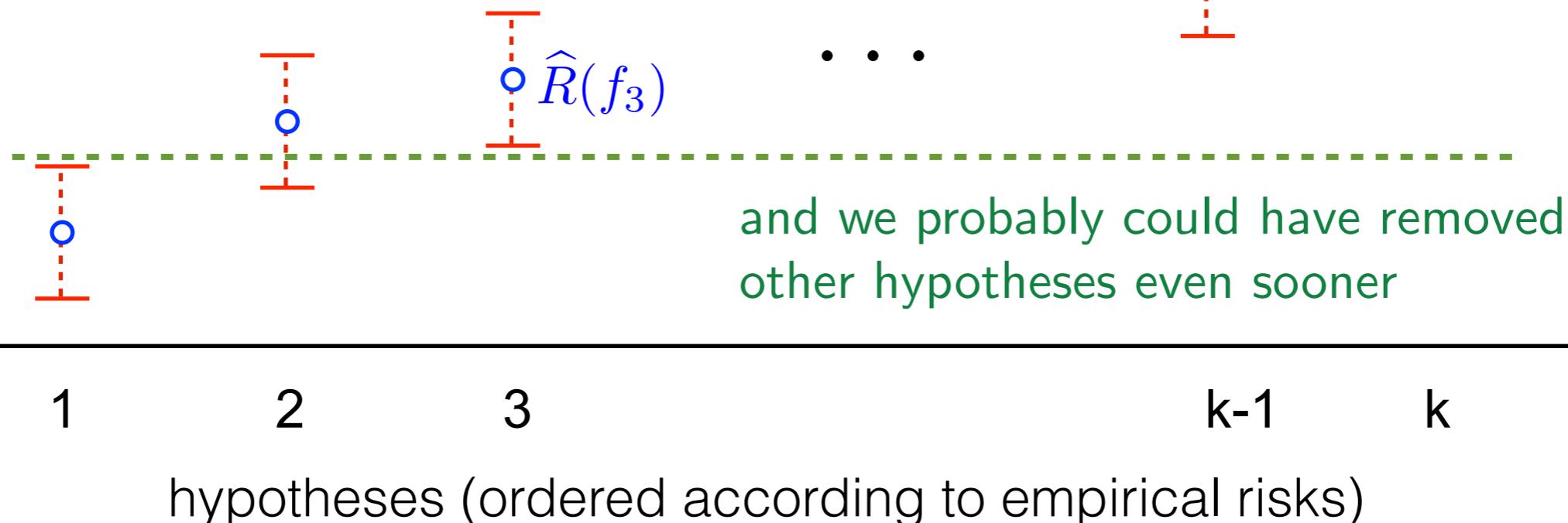


ERM is Wasting Labeled Examples

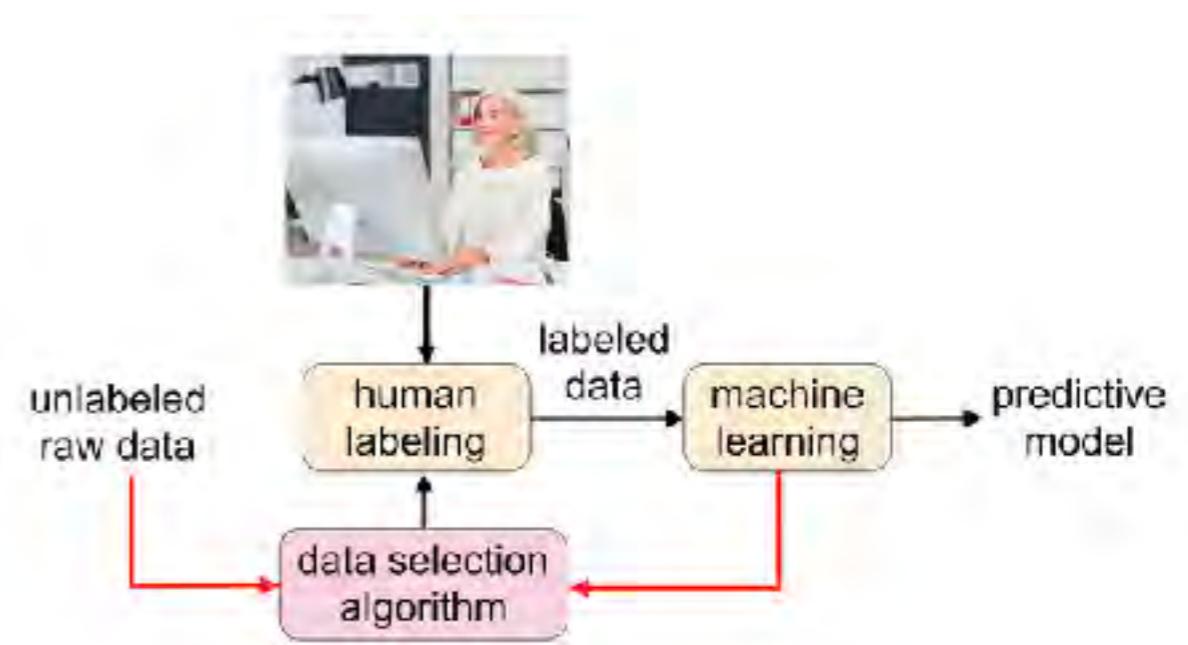


ERM is Wasting Labeled Examples

at this point we can safely remove
 f_3 from further consideration

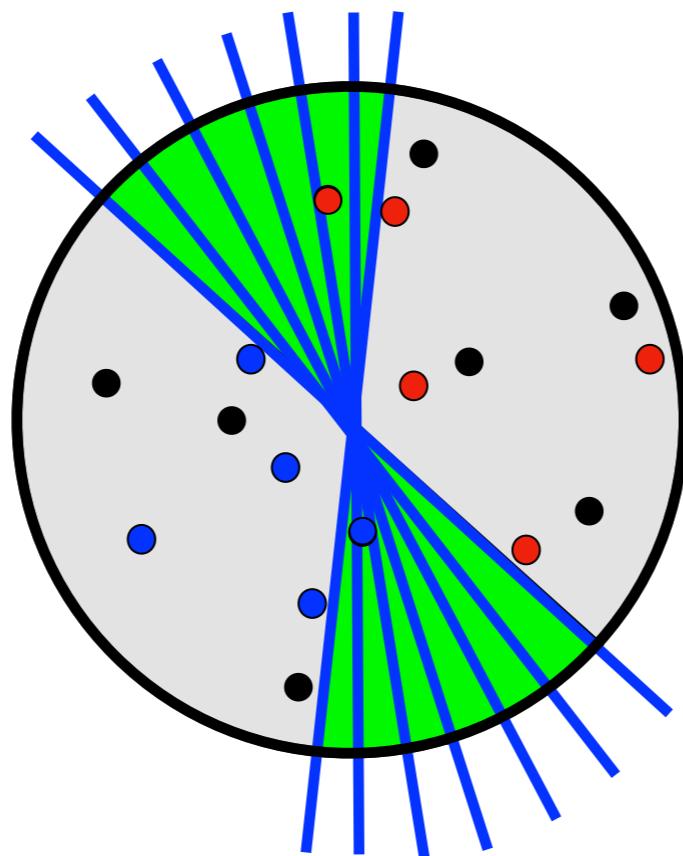


only require labels for examples that
hypotheses 1 and 2 label differently
(i.e., examples where they *disagree*)



Disagreement-Based Active Learning

consider points uniform on unit ball and linear classifiers passing through origin



only label points in the
region of disagreement \mathcal{D}

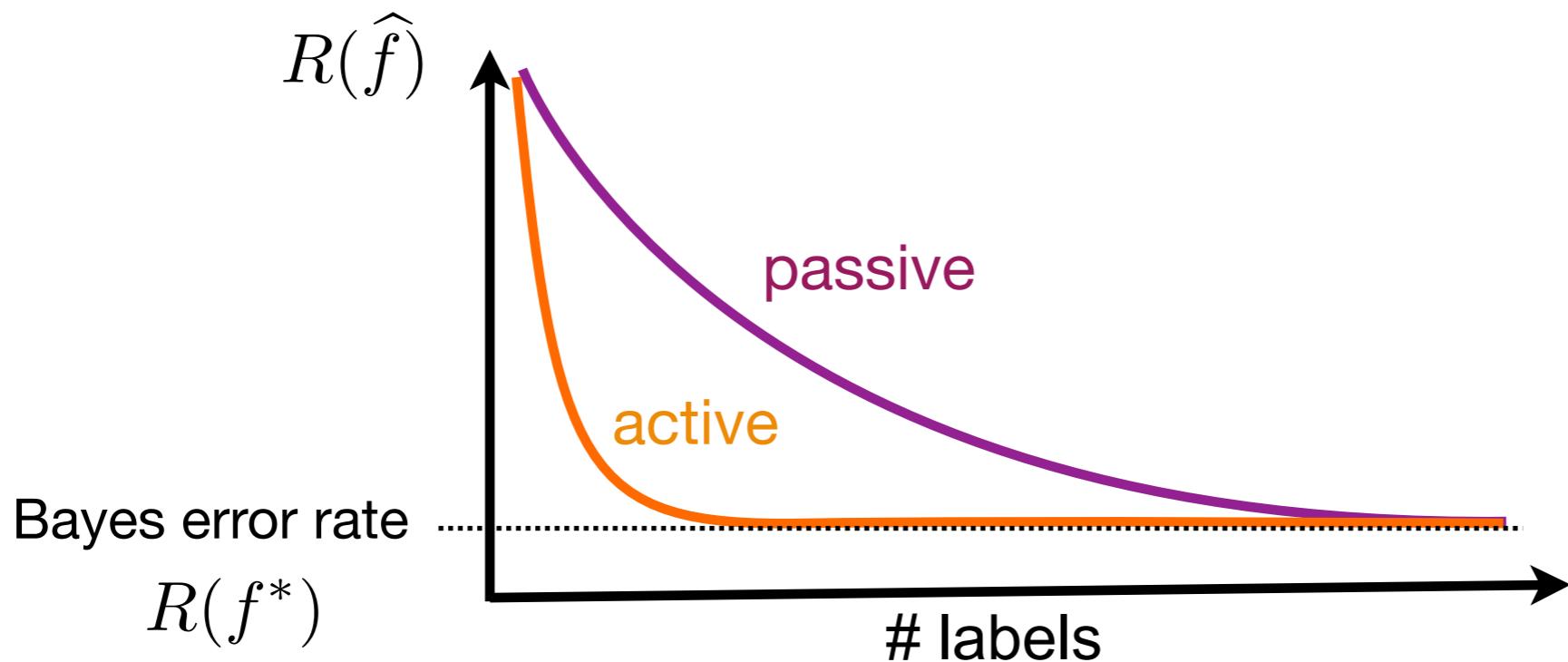
Active Binary Classification

Assuming optimal Bayes classifier f^* in VC class with dimension d and “nice” distributions (e.g., bounded label noise)

$$\epsilon = R(\hat{f}) - R(f^*)$$

passive $\epsilon \sim \frac{d}{n}$ parametric rate

active $\epsilon \sim \exp\left(-c \frac{n}{d}\right)$ exponential speed-up



Tutorial Outline

Part 1: Introduction to Active Learning (Rob)

Part 2: Theory of Active Learning (Steve)

Part 3: Advanced Topics and Open Problems (Steve)

Part 4: Nonparametric Active Learning (Rob)

slides: <http://nowak.ece.wisc.edu/ActiveML.html>

Recommended Reading (Foundations of Active Learning)

Settles, Burr. "Active learning." *Synthesis Lectures on Artificial Intelligence and Machine Learning* 6.1 (2012): 1-114.

Dasgupta, Sanjoy. "Two faces of active learning." *Theoretical computer science* 412.19 (2011): 1767-1781.

Cohn, David, Les Atlas, and Richard Ladner. "Improving generalization with active learning." *Machine learning* 15.2 (1994): 201-221.

Castro, Rui M., and Robert D. Nowak. "Minimax bounds for active learning." *IEEE Transactions on Information Theory* 54, no. 5 (2008): 2339-2353.

Zhu, Xiaojin, John Lafferty, and Zoubin Ghahramani. "Combining active learning and semi-supervised learning using gaussian fields and harmonic functions." *ICML 2003 workshop*. Vol. 3. 2003.

Dasgupta, Sanjoy, Daniel J. Hsu, and Claire Monteleoni. "A general agnostic active learning algorithm." *Advances in neural information processing systems*. 2008.

Balcan, Maria-Florina, Alina Beygelzimer, and John Langford. "Agnostic active learning." *Journal of Computer and System Sciences* 75.1 (2009): 78-89.

Nowak, Robert D. "The geometry of generalized binary search." *IEEE Transactions on Information Theory* 57, no. 12 (2011): 7893-7906.

Hanneke, Steve. "Theory of active learning." *Foundations and Trends in Machine Learning* 7, no. 2-3 (2014).

Part 2: Theory of Active Learning General Case

- Disagreement-Based Agnostic Active Learning
- Disagreement Coefficient
- Sample Complexity Bounds

**Tutorial on Active Learning:
Theory to Practice**

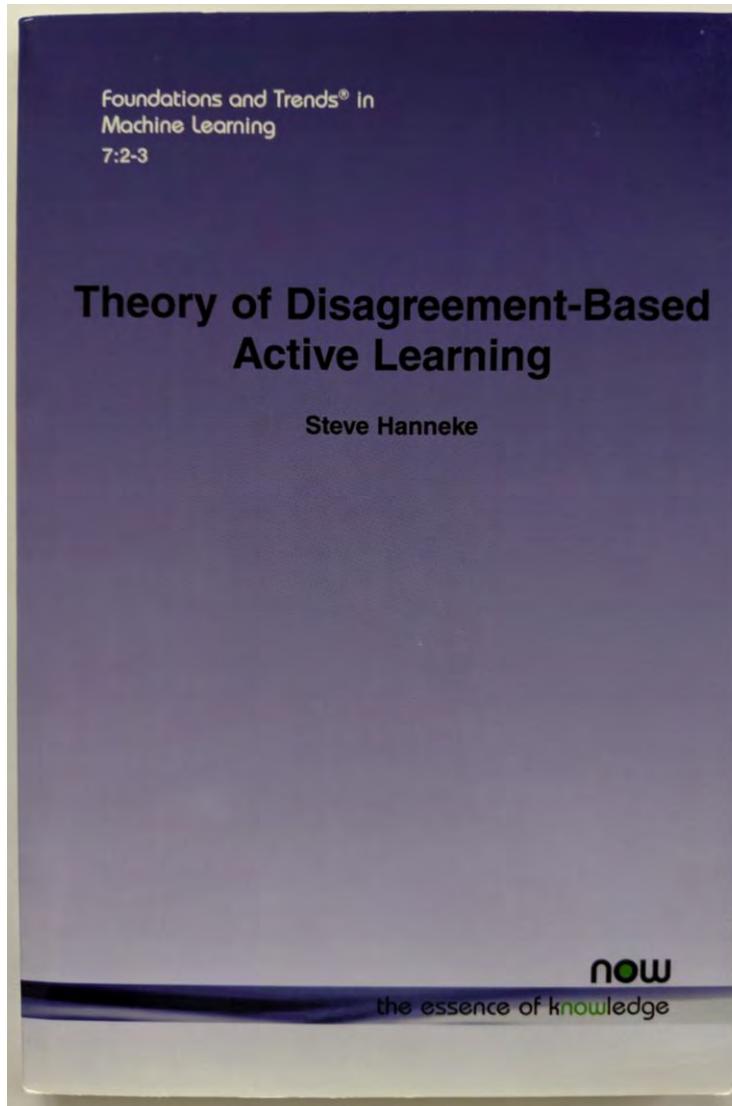
Steve Hanneke

Toyota Technological Institute at Chicago
steve.hanneke@gmail.com

Robert Nowak

University of Wisconsin - Madison
rdnowak@wisc.edu

Agnostic Active Learning



Uniform Bernstein Inequality

Bernstein's inequality:

For m iid samples

$\forall f, f'$, w.p. $1 - \delta$,

$$R(f) - R(f') \leq \hat{R}(f) - \hat{R}(f') + c\sqrt{\hat{P}(f \neq f') \frac{\log(1/\delta)}{m}} + \frac{\log(1/\delta)}{m}$$

Uniform Bernstein inequality:

w.p. $1 - \delta$, $\forall f, f' \in \mathcal{H}$,

$$R(f) - R(f') \leq \hat{R}(f) - \hat{R}(f') + c\sqrt{\hat{P}(f \neq f') \frac{d \log(m/\delta)}{m}} + \frac{d \log(m/\delta)}{m}$$

VC dimension

Roughly:

$\forall f, f' \in \mathcal{H}$,

$$R(f) - R(f') \leq \hat{R}(f) - \hat{R}(f') + \sqrt{\hat{P}(f \neq f') \frac{d}{m}}$$

Agnostic Active Learning

Balcan, Beygelzimer, & Langford (2006)

Region of disagreement:

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. **reduce** \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}

Agnostic Active Learning

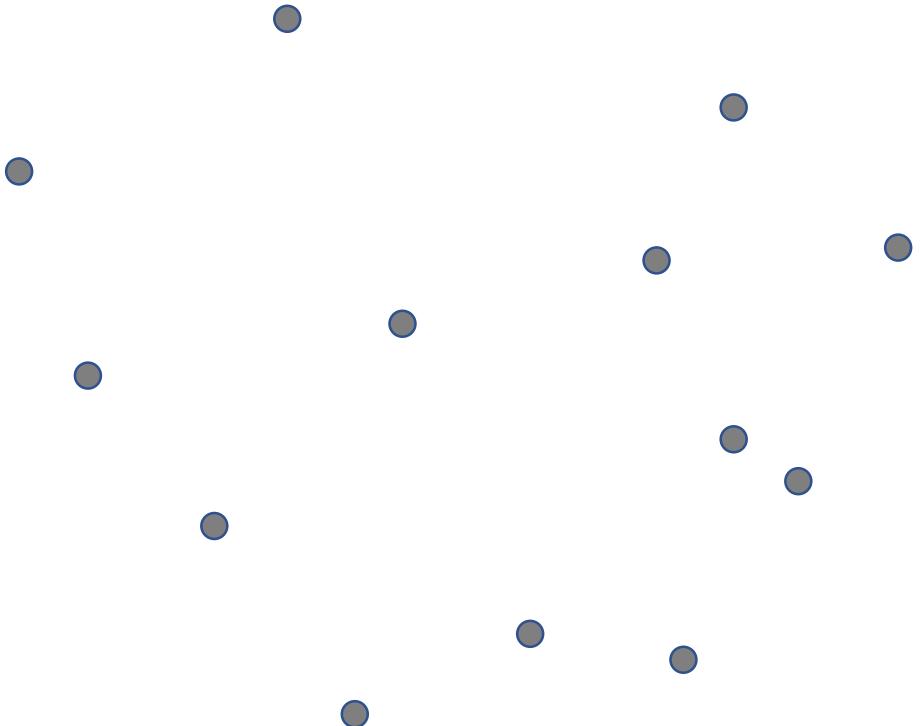
$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

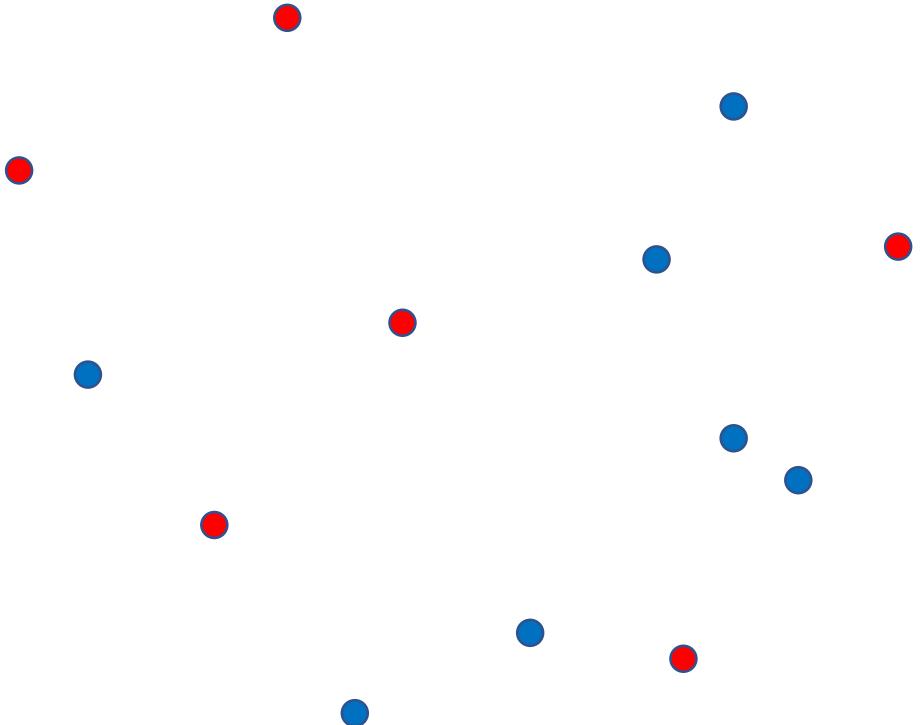
$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

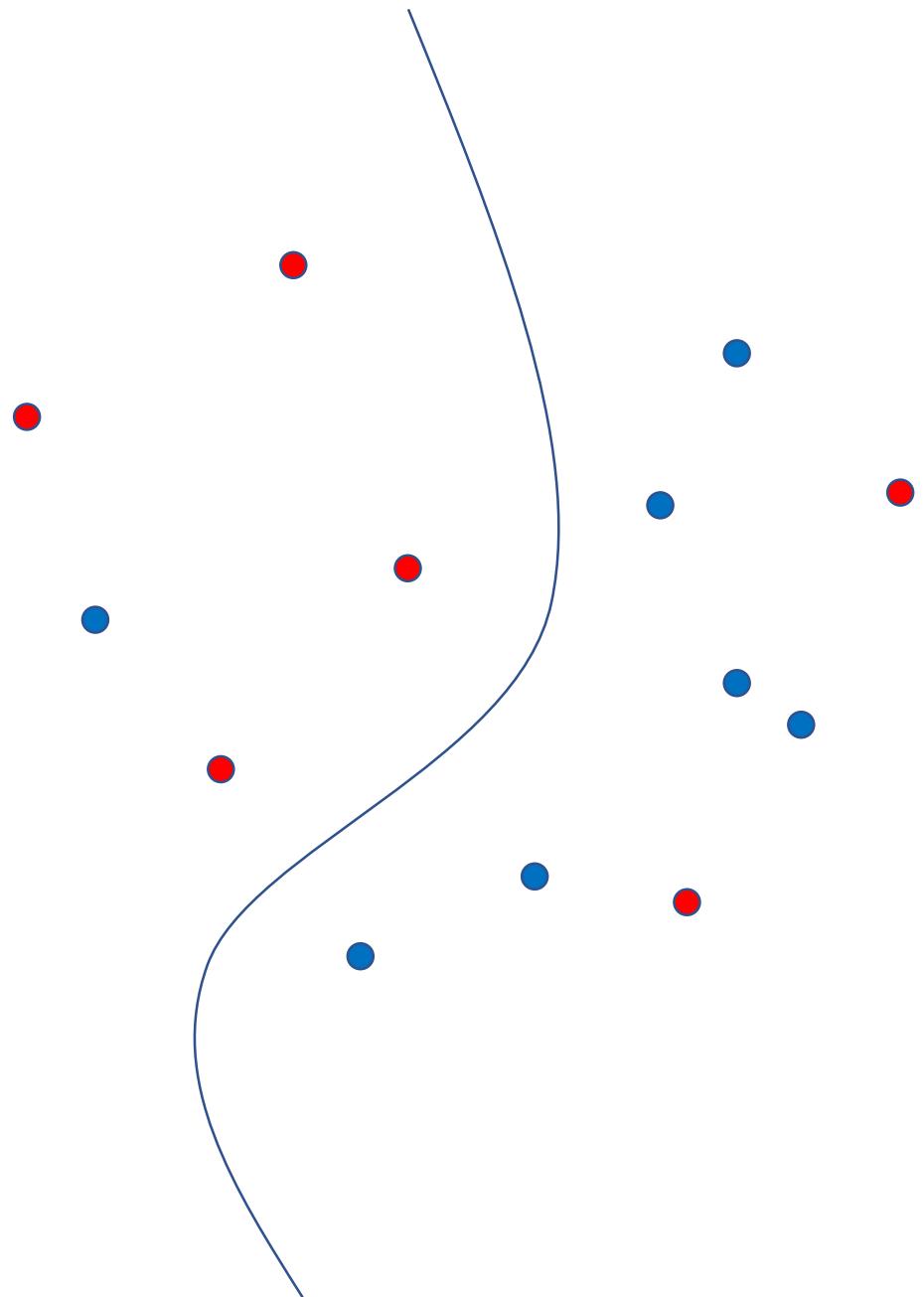
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

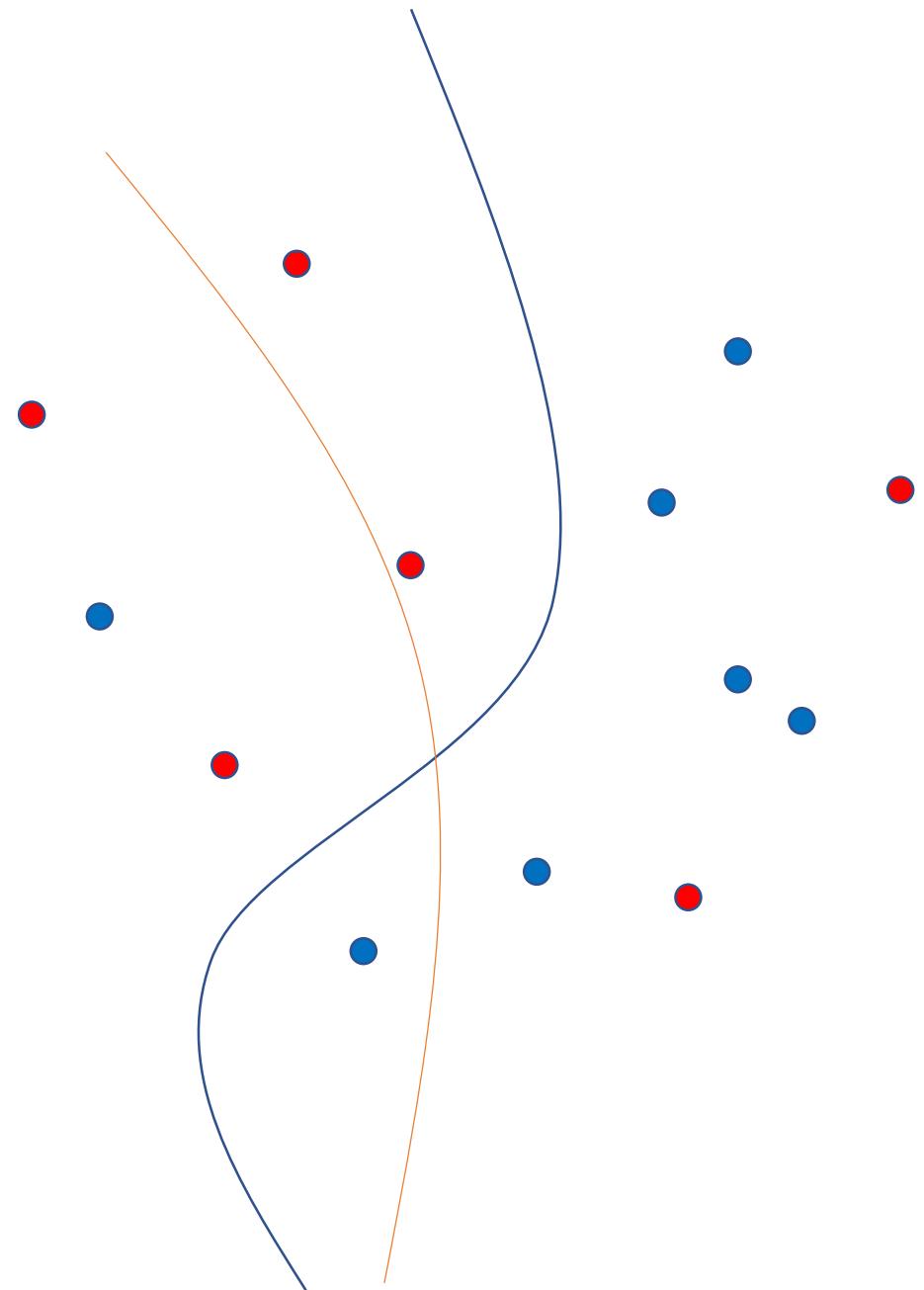
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

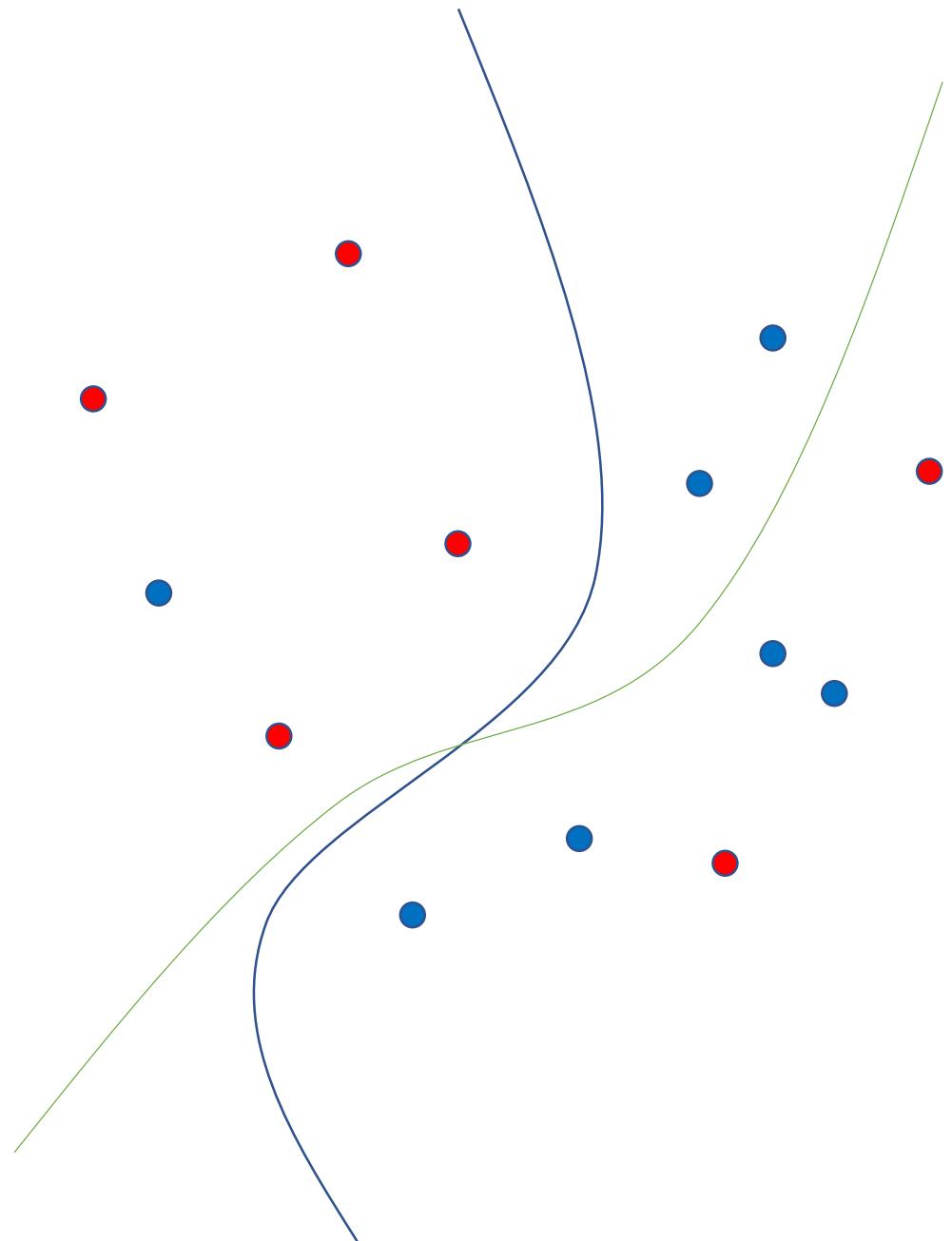
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

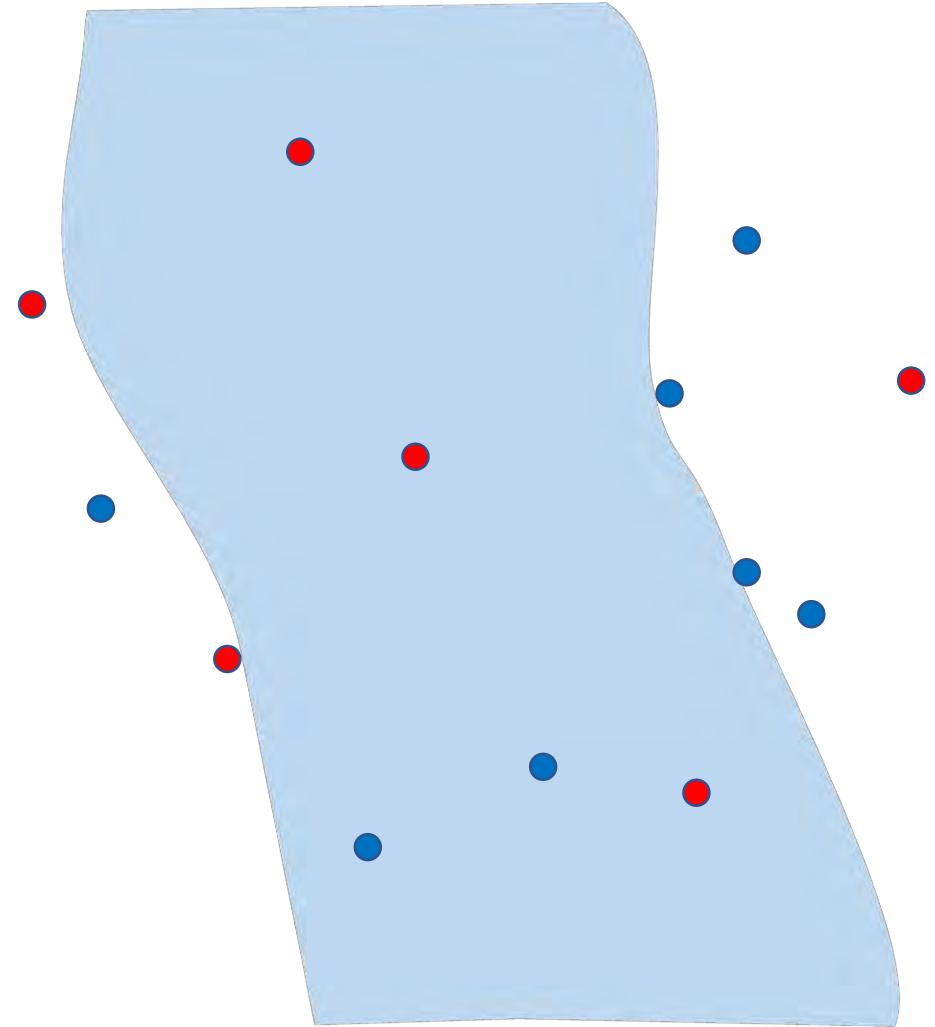
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

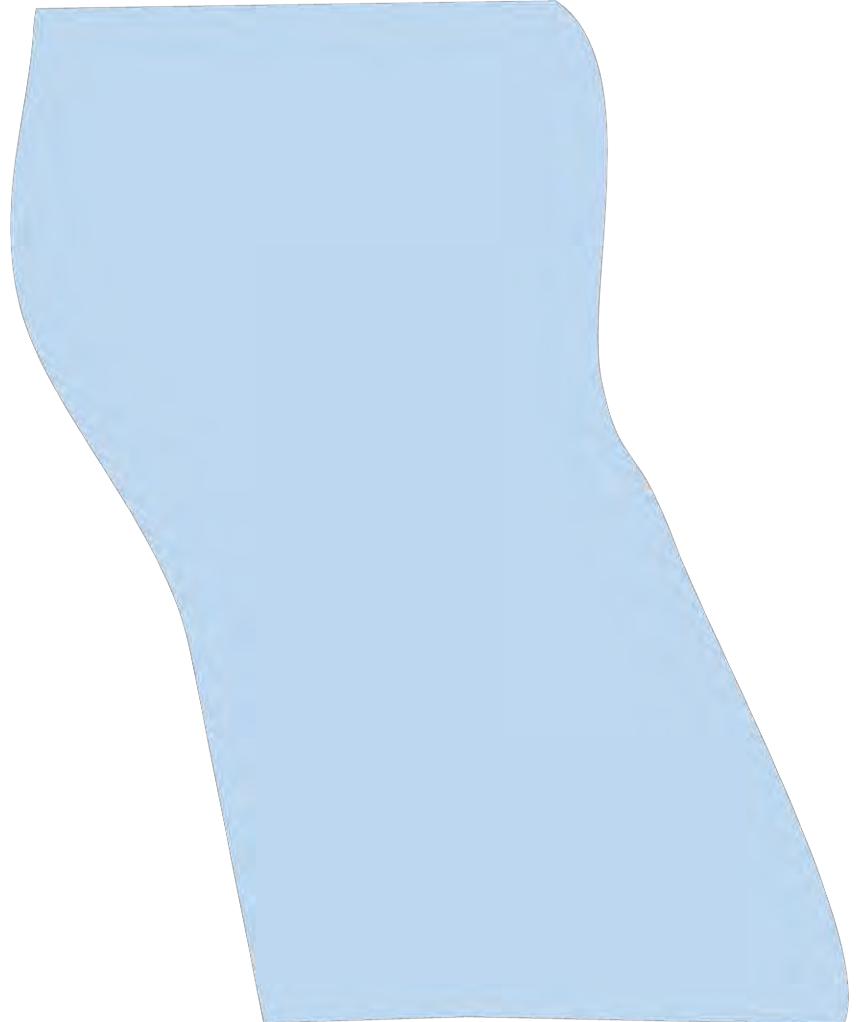
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. **reduce** \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}



Agnostic Active Learning

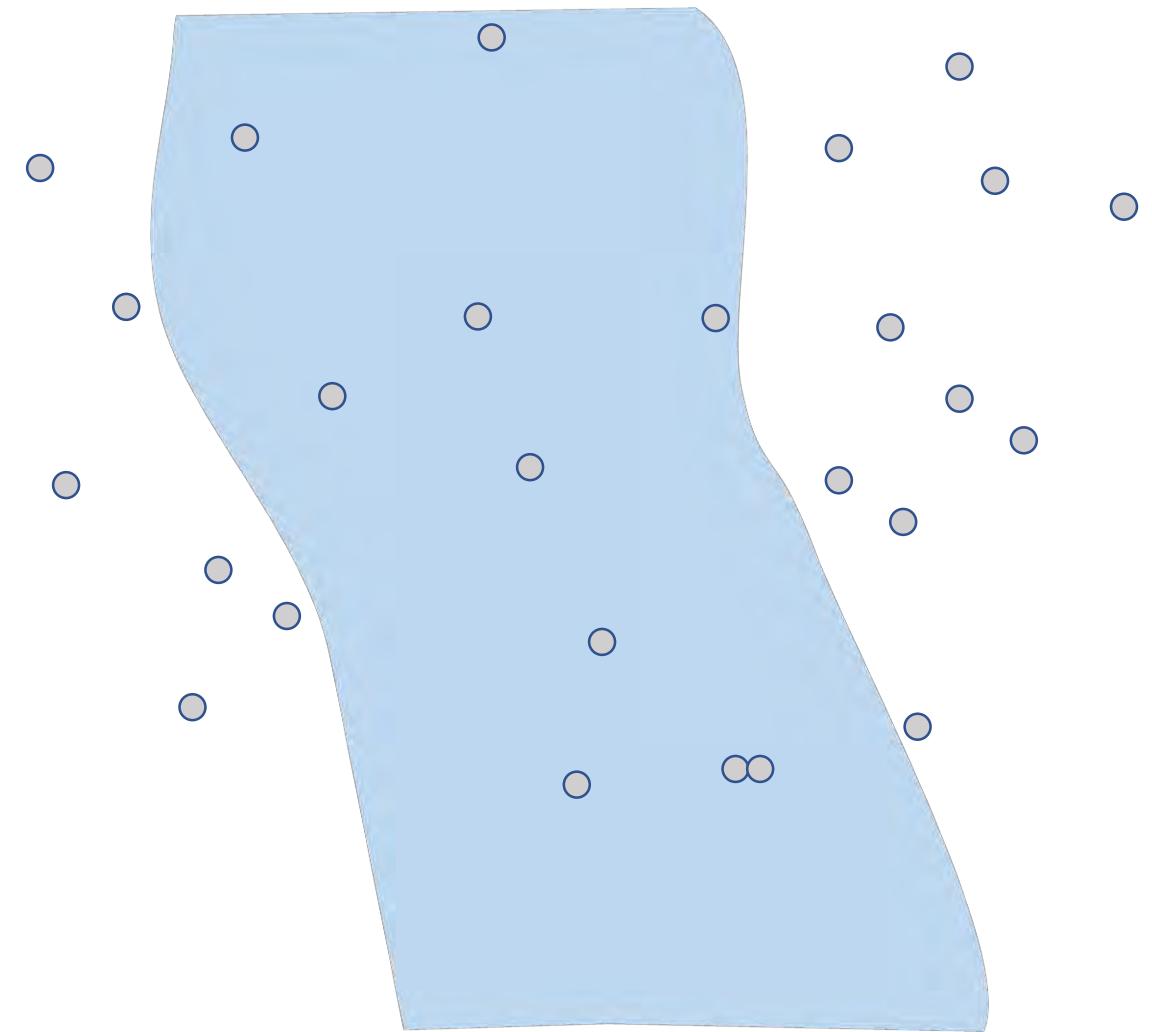
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

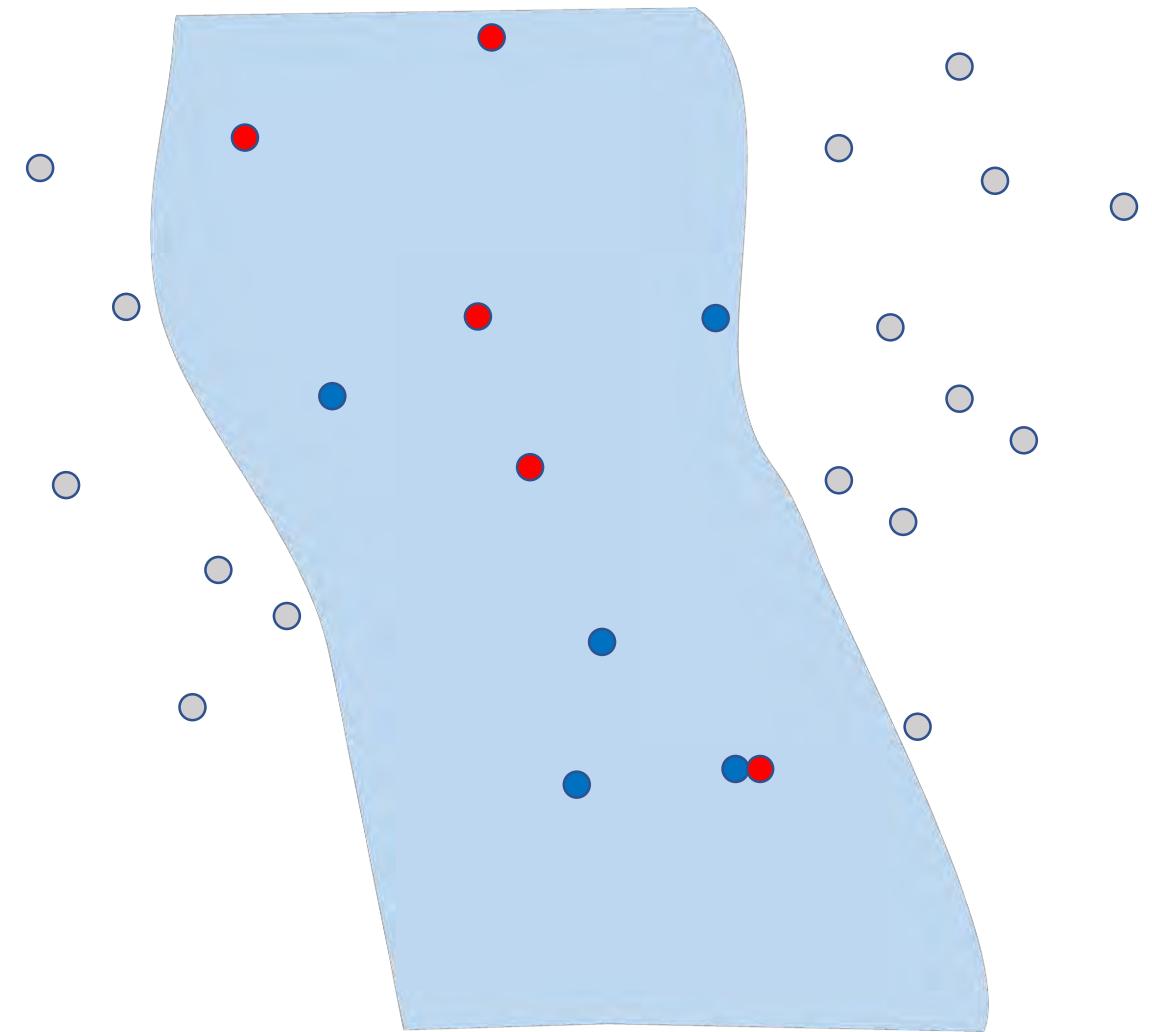
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

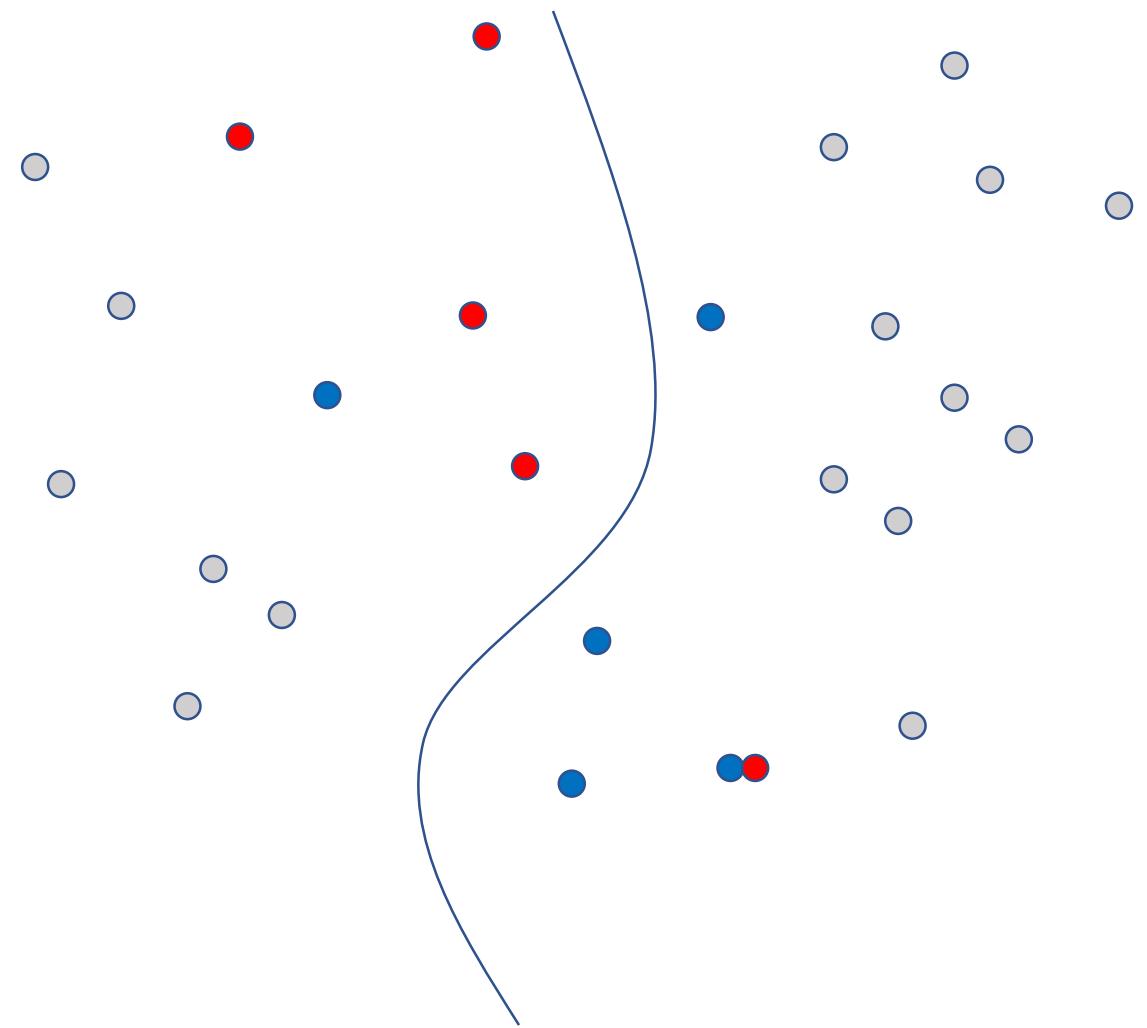
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

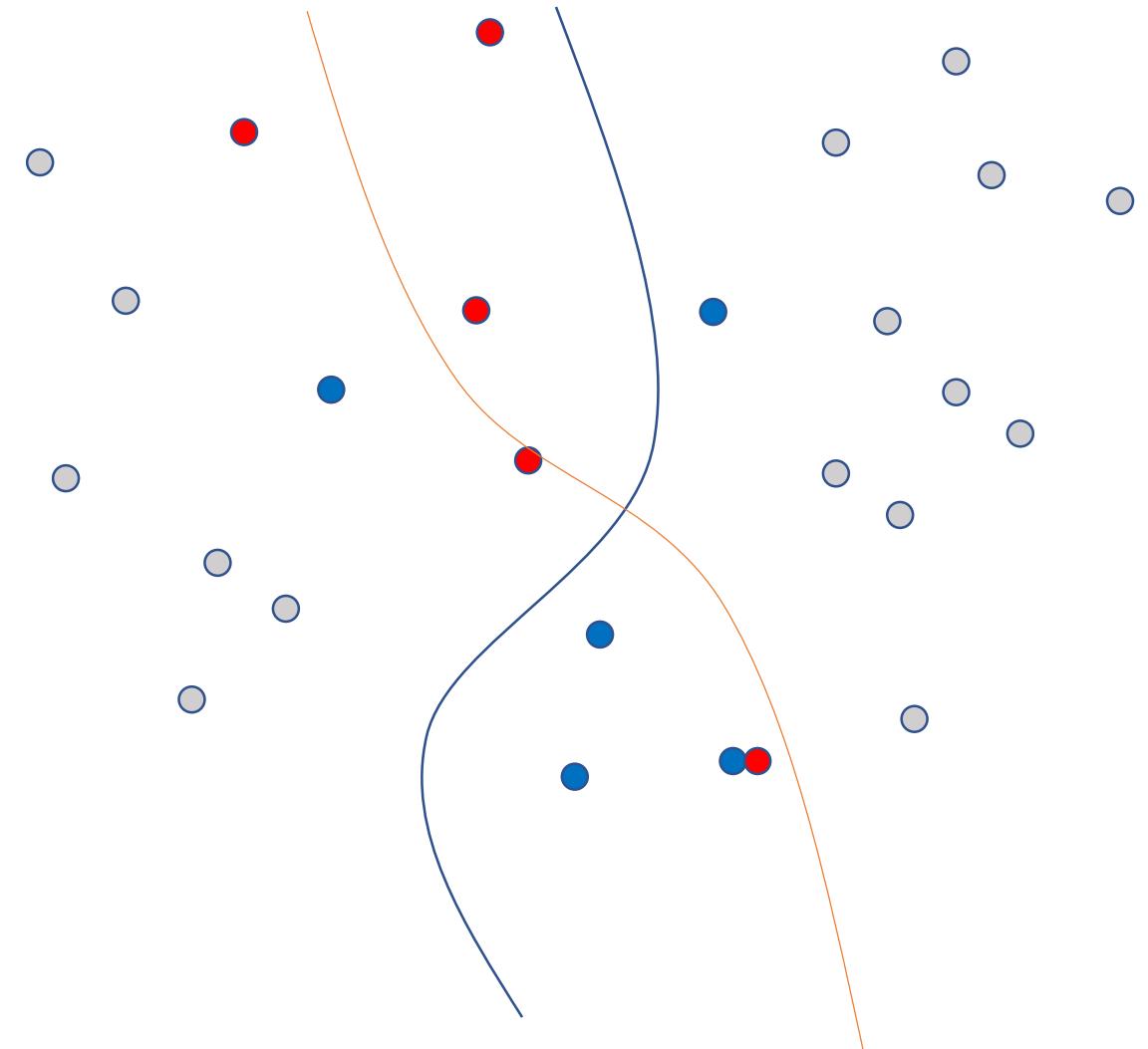
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

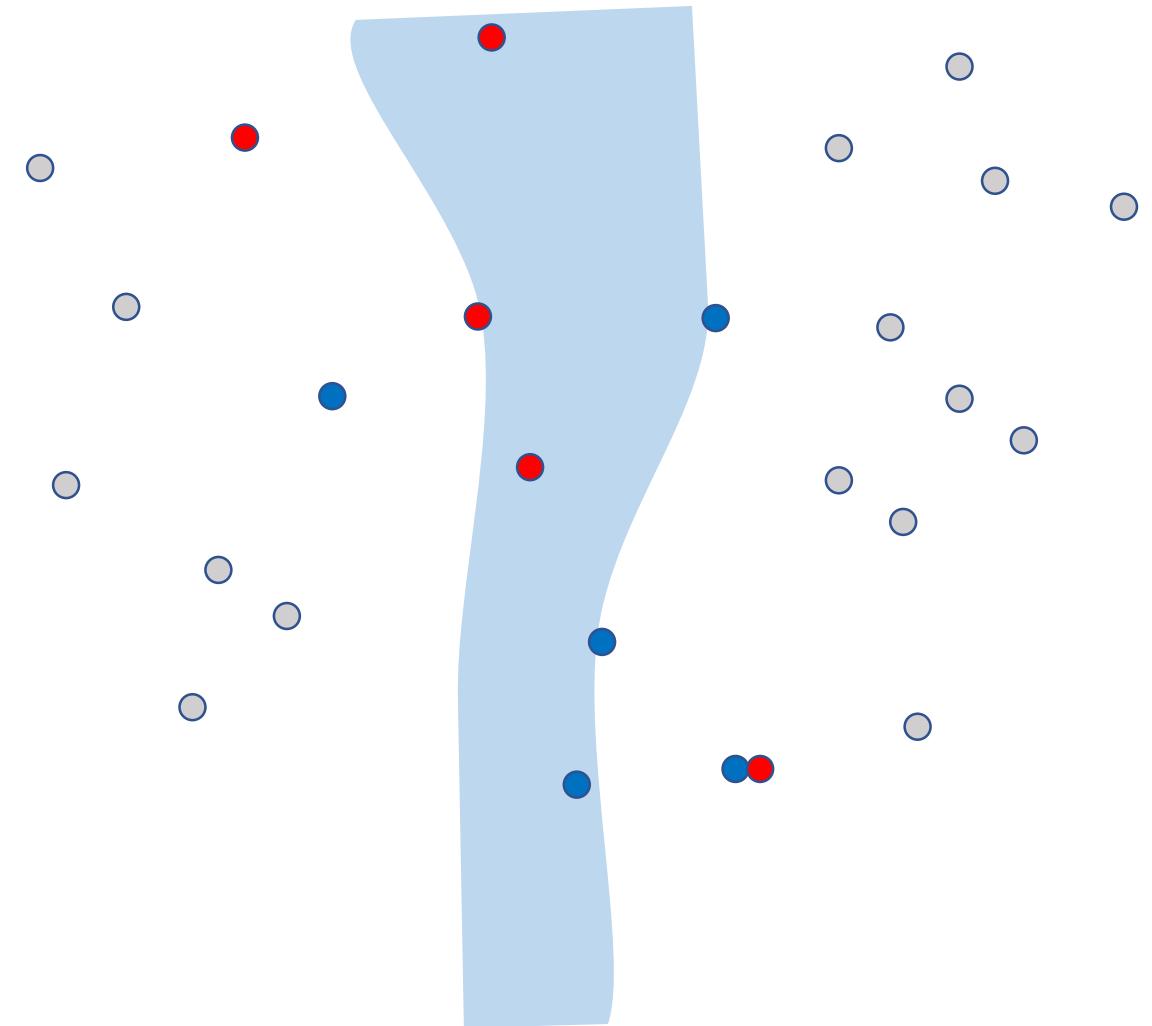
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

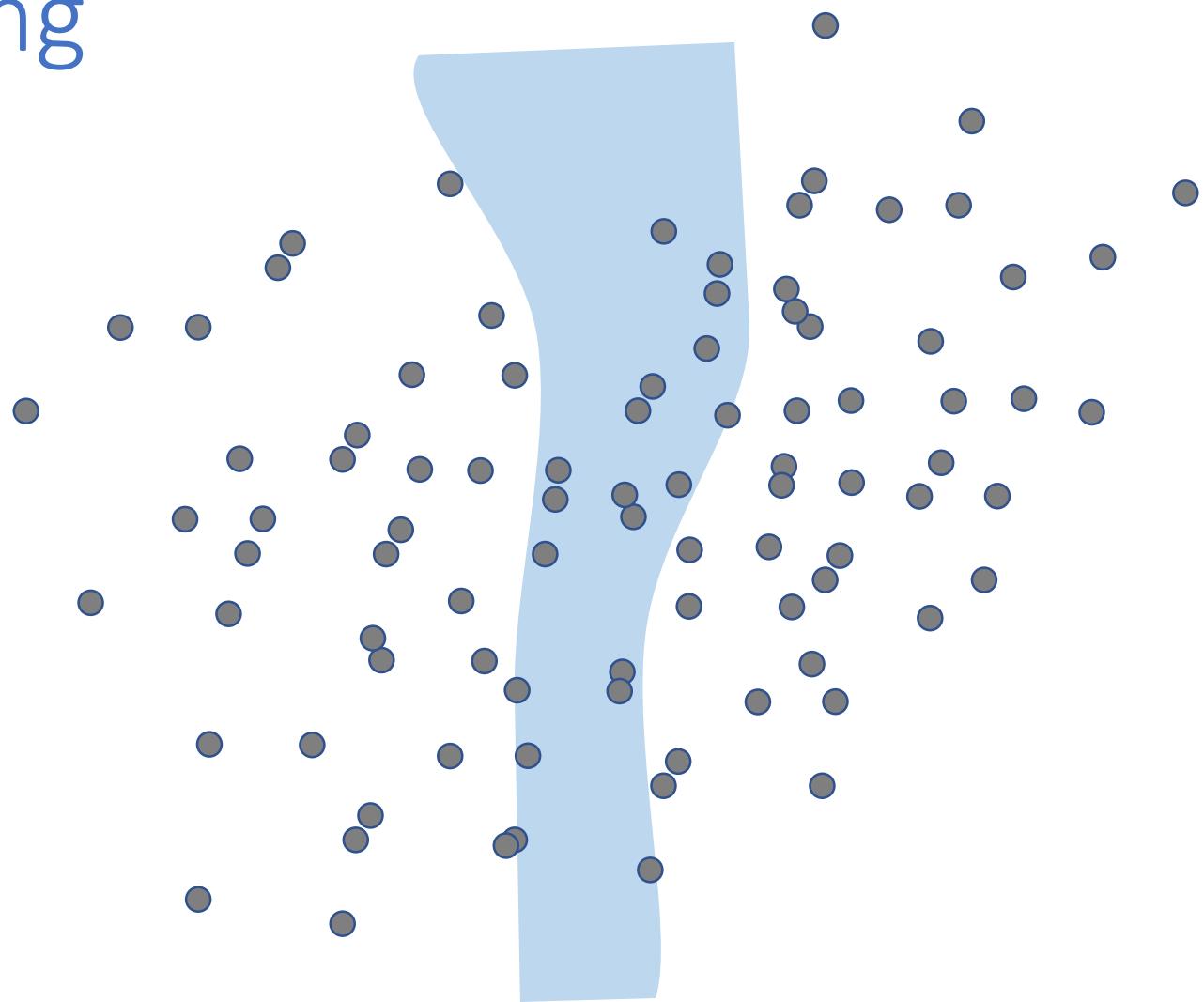
$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}



Agnostic Active Learning

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

The point:

Any t with $f^* \in \mathcal{H}$ still,
 $R(f^*|\text{DIS}(\mathcal{H}))$ still **minimal** in \mathcal{H}

$$\begin{aligned} &\Rightarrow \hat{R}_Q(f^*) - \hat{R}_Q(\hat{f}) \\ &\leq R(f^*|\text{DIS}(\mathcal{H})) - R(\hat{f}|\text{DIS}(\mathcal{H})) + \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}} \\ &\leq \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}} \\ &\Rightarrow \underline{f^* \text{ never removed.}} \end{aligned}$$

Agnostic Active Learning

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

The point:

Any t with $f^* \in \mathcal{H}$ still,
 $R(f^*|\text{DIS}(\mathcal{H}))$ still **minimal** in \mathcal{H}

$$\begin{aligned} &\Rightarrow \hat{R}_Q(f^*) - \hat{R}_Q(\hat{f}) \\ &\leq R(f^*|\text{DIS}(\mathcal{H})) - R(\hat{f}|\text{DIS}(\mathcal{H})) + \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}} \\ &\leq \sqrt{\hat{P}_Q(f^* \neq \hat{f}) \frac{d}{|Q|}} \\ &\Rightarrow \underline{f^* \text{ never removed.}} \end{aligned}$$

Next: **How many labels does it use?**

Sample Complexity Analysis

Hanneke (2007,...)

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Sample Complexity Analysis

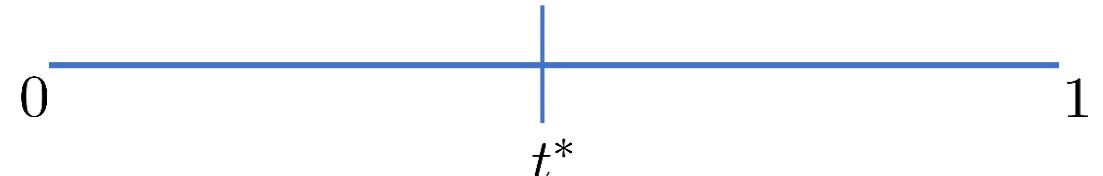
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Thresholds**, P_X Uniform(0, 1)
 $f(x) = \mathbb{I}[x \geq t]$



Sample Complexity Analysis

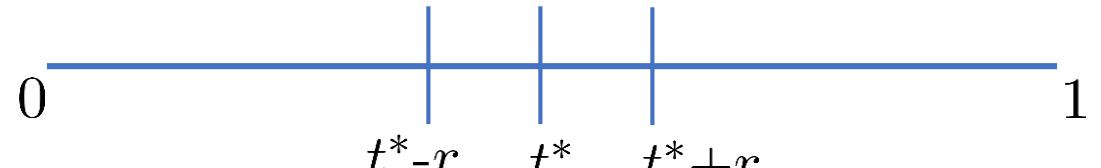
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Thresholds**, P_X Uniform(0, 1)
 $f(x) = \mathbb{I}[x \geq t]$



$$\text{DIS}(B(f^*, r)) = [t^* - r, t^* + r)$$

$$P_X(\text{DIS}(B(f^*, r))) = 2r$$

$$\theta = 2$$

Sample Complexity Analysis

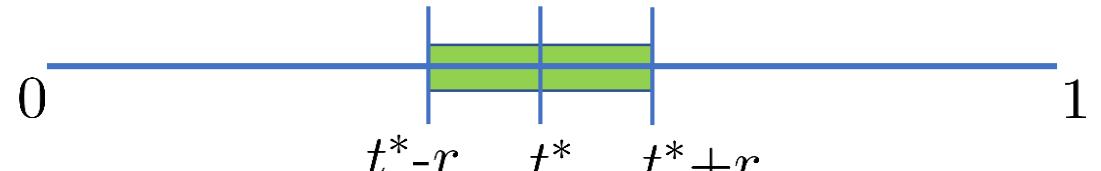
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Thresholds**, P_X Uniform(0, 1)
 $f(x) = \mathbb{I}[x \geq t]$



$$\text{DIS}(B(f^*, r)) = [t^* - r, t^* + r)$$

$$P_X(\text{DIS}(B(f^*, r))) = 2r$$

$$\Rightarrow \theta = 2$$

Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

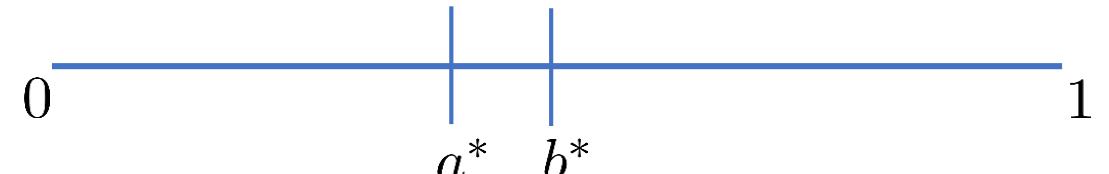
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

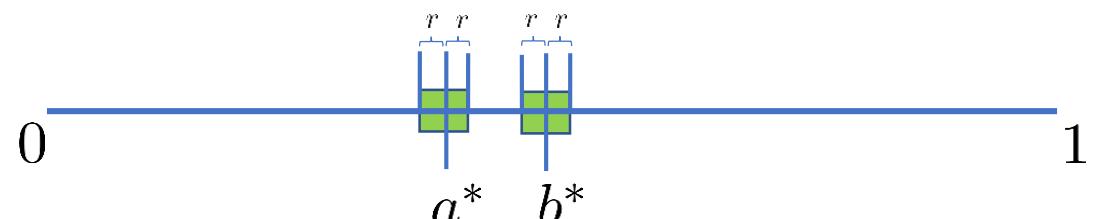
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



$$w^* := b^* - a^*$$

If $r < w^*$,

$$\text{DIS}(B(f^*, r)) = [a^* - r, a^* + r] \cup (b^* - r, b^* + r]$$

$$P_X(\text{DIS}(B(f^*, r))) = 4r$$

Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

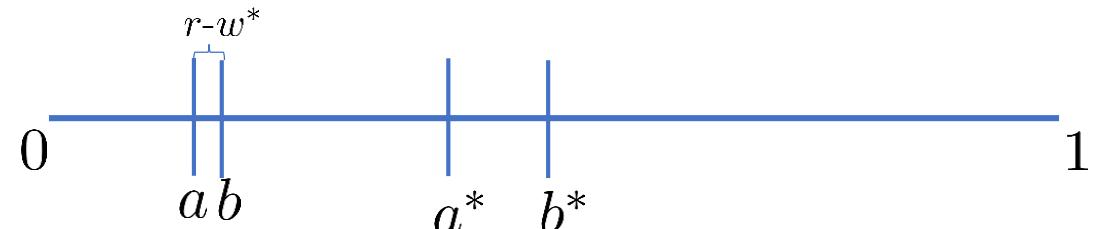
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



$$w^* := b^* - a^*$$

If $r > w^*$,

$$\text{DIS}(B(f^*, r)) = \mathcal{X}$$

$$P_X(\text{DIS}(B(f^*, r))) = 1$$

Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

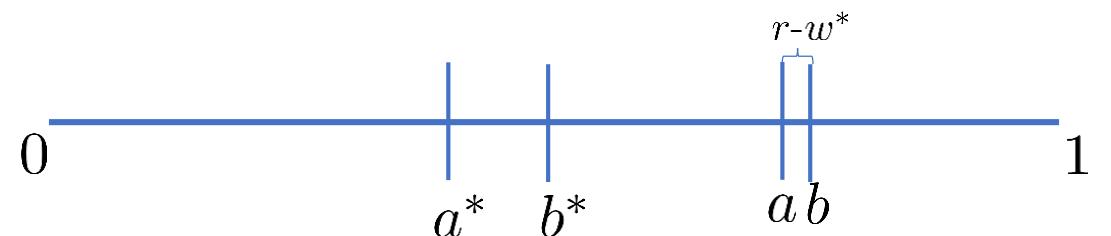
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



$$w^* := b^* - a^*$$

If $r > w^*$,

$$\text{DIS}(B(f^*, r)) = \mathcal{X}$$

$$P_X(\text{DIS}(B(f^*, r))) = 1$$

Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

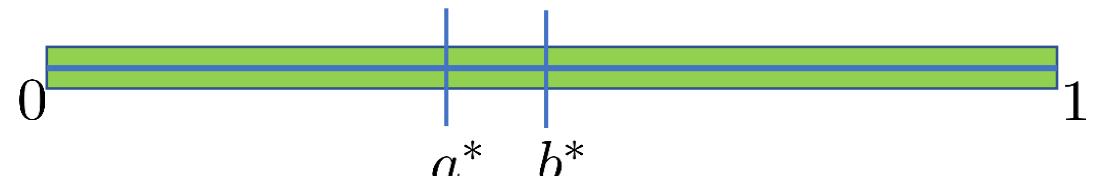
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



$$w^* := b^* - a^*$$

If $r > w^*$,

$$\text{DIS}(B(f^*, r)) = \mathcal{X}$$

$$P_X(\text{DIS}(B(f^*, r))) = 1$$

Sample Complexity Analysis

Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

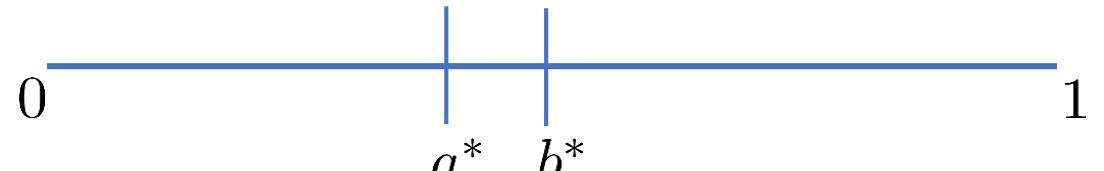
$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: **Intervals**, P_X Uniform(0, 1)

$$f(x) = \mathbb{I}[a \leq x \leq b]$$



$$w^* := b^* - a^*$$

If $r < w^*$, $P_X(\text{DIS}(B(f^*, r))) = 4r$

If $r > w^*$, $P_X(\text{DIS}(B(f^*, r))) = 1$

$$\Rightarrow \theta \leq \max\left\{4, \frac{1}{w^*}\right\}$$

Sample Complexity Analysis

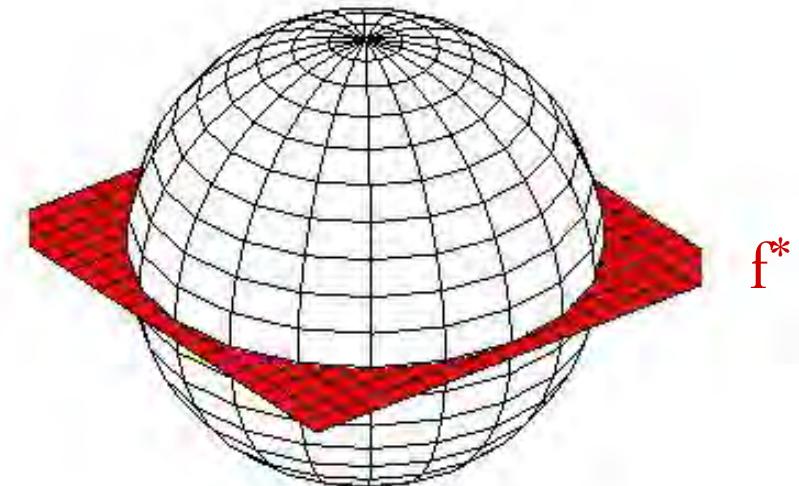
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0),
 n dimensions, uniform P_X on sphere.



Sample Complexity Analysis

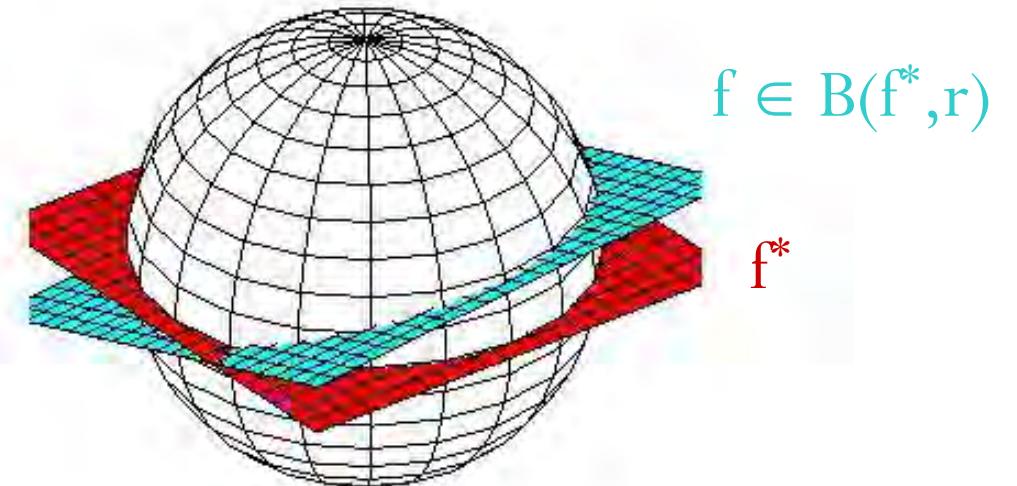
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0),
 n dimensions, uniform P_X on sphere.



Sample Complexity Analysis

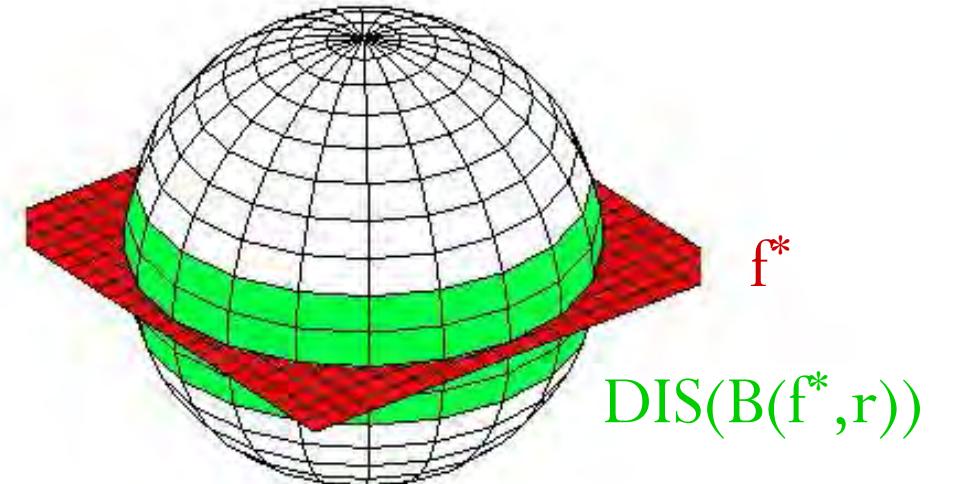
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0),
 n dimensions, uniform P_X on sphere.



Sample Complexity Analysis

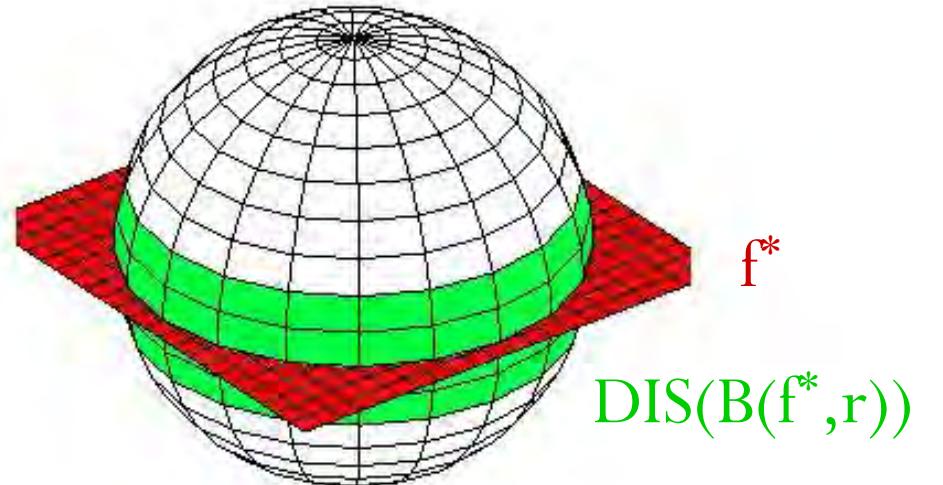
Ball: $B(f^*, r) := \{f \in \mathcal{H} : P_X(f \neq f^*) \leq r\}$

$\text{DIS}(B(f^*, r)) := \{x \in \mathcal{X} : \exists f, f' \in B(f^*, r), f(x) \neq f'(x)\}$

Disagreement coefficient:

$$\theta = \sup_{r > \epsilon} \frac{P_X(\text{DIS}(B(f^*, r)))}{r}$$

Example: homog. linear separators (bias 0),
 n dimensions, uniform P_X on sphere.



Some geometry \Rightarrow for small r ,
 $P_X(\text{DIS}(B(f^*, r))) \propto \sqrt{nr}$.
 $\Rightarrow \theta \propto \sqrt{n}$.

Sample Complexity Analysis

Bounded Noise assumption: (aka Massart noise)

$$\exists \beta < 1/2 \text{ s.t. } P(Y \neq f^*(X)|X) \leq \beta \text{ everywhere}$$

	Sample Complexity: $R(\hat{f}) \leq R(f^*) + \epsilon$	Excess Error: n labels
Passive	$\frac{d}{\epsilon}$	$\frac{d}{n}$
Active	$d\theta \log(\frac{1}{\epsilon})$	$e^{-n/d\theta}$

Sample Complexity Analysis

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Theorem: $P(Y \neq f^*(X)|X) \leq \beta$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\# \text{ labels} \approx d\theta \log\left(\frac{1}{\epsilon}\right).$$

Proof Sketch:

Round t , all $f \in \mathcal{H}$ agree on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with

$$R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*) \frac{d}{2^t}}$$

\Rightarrow surviving f after round t have $R(f) - R(f^*) \lesssim \frac{d}{2^t}$

$\Rightarrow t \gtrsim \log\left(\frac{d}{\epsilon}\right)$ suffices

Also \Rightarrow after round $t-1$, $\mathcal{H} \subseteq B(f^*, d/2^{t-1})$

$$\Rightarrow |Q| \lesssim P_X(\text{DIS}(B(f^*, d/2^{t-1})))|S| \leq \theta \frac{d}{2^{t-1}} |S| = \theta d 2$$

$$\sum_{t=1}^{\log(d/\epsilon)} \theta d = \theta d \log\left(\frac{d}{\epsilon}\right)$$



Sample Complexity Analysis

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Bounded noise:

$$\begin{aligned} R(f) - R(f^*) &= \int_{f \neq f^*} (P(Y = f^*(X)|X) - P(Y \neq f^*(X)|X)) dP_X \\ &\geq (1 - 2\beta) P_X(f \neq f^*) \end{aligned}$$

Theorem: $P(Y \neq f^*(X)|X) \leq \beta$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\# \text{ labels} \approx d\theta \log\left(\frac{1}{\epsilon}\right).$$

Proof Sketch:

Round t , all $f \in \mathcal{H}$ agree on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with
 $R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*) \frac{d}{2^t}}$

\Rightarrow surviving f after round t have $R(f) - R(f^*) \lesssim \frac{d}{2^t}$
 $\Rightarrow t \gtrsim \log\left(\frac{d}{\epsilon}\right)$ suffices

Also \Rightarrow after round $t - 1$, $\mathcal{H} \subseteq B(f^*, d/2^{t-1})$

$$\Rightarrow |Q| \lesssim P_X(\text{DIS}(B(f^*, d/2^{t-1}))) |S| \leq \theta \frac{d}{2^{t-1}} |S| = \theta d 2$$

$$\sum_{t=1}^{\log(d/\epsilon)} \theta d = \theta d \log\left(\frac{d}{\epsilon}\right)$$



Sample Complexity Analysis

Agnostic Learning: (no assumptions)

Denote $\beta = R(f^*)$

	Sample Complexity: $R(\hat{f}) \leq R(f^*) + \epsilon$	Excess Error: n labels
Passive	$d \frac{\beta}{\epsilon^2}$	$\sqrt{\frac{d\beta}{n}}$
Active	$d\theta \frac{\beta^2}{\epsilon^2}$	$\sqrt{\frac{d\beta^2\theta}{n}}$

Sample Complexity Analysis

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Theorem: $\beta = R(f^*)$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\# \text{ labels} \approx d\theta \frac{\beta^2}{\epsilon^2}.$$

Proof Sketch:

Round t , all $f \in \mathcal{H}$ agree on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with

$$R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*) \frac{d}{2^t}}$$

\Rightarrow surviving f after round t have $R(f) - R(f^*) \lesssim \sqrt{\beta \frac{d}{2^t}} + \frac{d}{2^t}$

(Roughly) $\sqrt{\beta \frac{d}{2^t}}$

$\Rightarrow t \gtrsim \log(d \frac{\beta}{\epsilon^2})$ suffices

Also \Rightarrow after round $t-1$, $\mathcal{H} \subseteq B\left(f^*, 2\beta + \sqrt{\beta \frac{d}{2^{t-1}}}\right) \subseteq B(f^*, 3\beta)$ (for large t)

$\Rightarrow |Q| \lesssim P_X(\text{DIS}(B(f^*, 3\beta)))|S| \lesssim \theta\beta|S| = \theta\beta 2^t$

$$\sum_{t=1}^{\log(d\beta/\epsilon^2)} \theta\beta 2^t \sim \theta d \frac{\beta^2}{\epsilon^2}$$



Sample Complexity Analysis

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
 3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

$$P_X(f \neq f^*) \leq R(f) + R(f^*) = 2\beta + R(f) - R(f^*)$$

Theorem: $\beta = R(f^*)$. $R(\hat{f}) \leq R(f^*) + \epsilon$ with

$$\# \text{ labels} \approx d\theta \frac{\beta^2}{\epsilon^2}.$$

Proof Sketch:

Round t , all $f \in \mathcal{H}$ agree on pts in $S \setminus Q$

Roughly, that means Step 4 only keeps f with

$$R(f) - R(f^*) \lesssim \sqrt{P_X(f \neq f^*) \frac{d}{2^t}}$$

\Rightarrow surviving f after round t have $R(f) - R(f^*) \lesssim \sqrt{\beta \frac{d}{2^t}} + \frac{d}{2^t}$

(Roughly) $\sqrt{\beta \frac{d}{2^t}}$

$\Rightarrow t \gtrsim \log(d \frac{\beta}{\epsilon^2})$ suffices

Also \Rightarrow after round $t-1$, $\mathcal{H} \subseteq B\left(f^*, 2\beta + \sqrt{\beta \frac{d}{2^{t-1}}}\right) \subseteq B(f^*, 3\beta)$ (for large t)

$$\Rightarrow |Q| \lesssim P_X(\text{DIS}(B(f^*, 3\beta)))|S| \lesssim \theta\beta|S| = \theta\beta 2^t$$

$$\sum_{t=1}^{\log(d\beta/\epsilon^2)} \theta\beta 2^t \sim \theta d \frac{\beta^2}{\epsilon^2}$$



Sample Complexity Analysis

When is θ small?

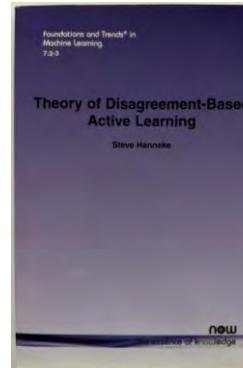
- Linear separators, P_X has a density,
 f^* boundary intersects interior of support
 $\Rightarrow \theta$ bounded
- Linear separators, P_X has a density
 $\Rightarrow \theta \ll \frac{1}{\epsilon}$
- \mathcal{H} smoothly-parametrized model,
 P_X “regular” density w/ compact support,
other technical conditions on \mathcal{H}
 $\Rightarrow \theta \propto \# \text{ parameters for } \mathcal{H}$
- ...

Sample Complexity Analysis

When is θ small?

- Linear separators, P_X has a density,
 f^* boundary intersects interior of support
 $\Rightarrow \theta$ bounded
- Linear separators, P_X has a density
 $\Rightarrow \theta \ll \frac{1}{\epsilon}$
- \mathcal{H} smoothly-parametrized model,
 P_X “regular” density w/ compact support,
other technical conditions on \mathcal{H}
 $\Rightarrow \theta \propto \# \text{ parameters for } \mathcal{H}$
- ...

Lots more 



Stopping Criterion

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Stopping criteria:

- Any-time
- Label budget
- Run out of unlabeled data
- Check $\max_{f \in \mathcal{H}} \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}} < \epsilon$

Simpler Agnostic Active Learning

Hsu (2010,...)

$Q \leftarrow \{\}$

for $m = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** a random point x

2. **optimize** $\forall y, \hat{f}_y \leftarrow \operatorname{argmin}_{f \in \mathcal{H}: f(x)=y} \hat{R}_Q(f)$

3. if $|\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_- \neq \hat{f}_+) \frac{d}{|Q|}}$

then **label** x , add it to Q

output $\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$

- Roughly same sample complexity as A^2 .
- Can implement as a **reduction** to ERM.
- In practice, replace ERM with any passive learner.

Surrogate Loss

Hanneke & Yang (2012)

$$Q \leftarrow \{\}$$

for $m = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** a random point x

2. **optimize** $\forall y, \hat{f}_y \leftarrow \operatorname{argmin}_{f \in \mathcal{H}: f(x)=y} \hat{R}_Q^\ell(f)$

3. if $|\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_- \neq \hat{f}_+) \frac{d}{|Q|}}$

then **label** x , add it to Q

output $\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$

- Roughly same sample complexity as A^2 .
- Can implement as a **reduction** to ERM.
- In practice, replace ERM with any passive learner.

Consider learner that minimizes a **surrogate loss**
 $\ell : \mathbb{R} \times \{-1, +1\} \rightarrow \mathbb{R}_+$
(e.g., hinge loss, squared loss, exponential loss, ...)

Now \mathcal{H} is **real-valued** functions

$$\hat{R}_Q^\ell(f) = \frac{1}{|Q|} \sum_{(x,y) \in Q} \ell(f(x), y)$$

Theorem: Bounded noise, plus strong assumptions on \mathcal{H}, ℓ, P
still get $R(\hat{f}) \leq R(f^*) + \epsilon$ with # labels

$$\approx \theta d \log\left(\frac{1}{\epsilon}\right)$$

Importance-Weighted Active Learning

Beygelzimer, Dasgupta,
Langford (2009)

$$Q \leftarrow \{\}$$

for $m = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** a random point x
2. **set** sampling probability p_x
3. **flip** coin with prob p_x of heads
4. if heads, **label** x , add to Q with weight $1/p_x$

output $\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$ (weighted loss)

Use importance weights to stay **unbiased**:
 $\mathbb{E}[\hat{R}_Q(f)] = R(f)$

Now Q set of triples (x, y, w)

$$\hat{R}_Q(f) = \frac{1}{|Q|} \sum_{(x,y,w) \in Q} w \mathbb{I}[f(x) \neq y]$$

- **Any** choice of Step 2 (setting p_x) is fine (just p_x not too small, else high variance)
- Can set p_x in a way to recover A^2 sample complexity
$$p_x = \mathbb{I}\left[|\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_+ \neq \hat{f}_-) \frac{d}{|Q|}} \right]$$

Importance-Weighted Active Learning

Beygelzimer, Dasgupta,
Langford (2009)

$Q \leftarrow \{\}$

for $m = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** a random point x
2. **set** sampling probability p_x
3. **flip** coin with prob p_x of heads
4. if heads, **label** x , add to Q with weight $1/p_x$

output $\hat{f} = \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$ (weighted loss)

Use importance weights to stay **unbiased**:
 $\mathbb{E}[\hat{R}_Q(f)] = R(f)$

Now Q set of triples (x, y, w)

$$\hat{R}_Q(f) = \frac{1}{|Q|} \sum_{(x,y,w) \in Q} w \mathbb{I}[f(x) \neq y]$$

- **Any** choice of Step 2 (setting p_x) is fine (just p_x not too small, else high variance)
- Can set p_x in a way to recover A^2 sample complexity
$$p_x = \mathbb{I}\left[|\hat{R}_Q(\hat{f}_+) - \hat{R}_Q(\hat{f}_-)| \leq \sqrt{\hat{P}_Q(\hat{f}_+ \neq \hat{f}_-) \frac{d}{|Q|}} \right]$$
- In practice, replace ERM with any passive learner (e.g., ERM with a surrogate loss)
- (approx) implementation in **Vowpal Wabbit** library

Questions?

Further reading:

- D. Cohn, L. Atlas, R. Ladner. Improving generalization with active learning. *Machine Learning*, 1994
- M. F. Balcan, A. Beygelzimer, J. Langford. Agnostic active learning. *Journal of Computer and System Sciences*, 2009.
- S. Hanneke. A bound on the label complexity of agnostic active learning. ICML 2007.
- S. Dasgupta, D. Hsu, C. Monteleoni. A general agnostic active learning algorithm. NeurIPS 2007.
- S. Hanneke. Rates of convergence in active learning. *The Annals of Statistics*, 2011.
- A. Beygelzimer, S. Dasgupta, J. Langford. Importance weighted active learning. ICML 2009.
- A. Beygelzimer, D. Hsu, J. Langford, T. Zhang. Agnostic active learning without constraints. NeurIPS 2010.
- S. Hanneke. Theoretical Foundations of Active Learning. PhD Thesis, CMU, 2009.
- D. Hsu. Algorithms for Active Learning. PhD Thesis, UCSD, 2010.
- Y. Wiener, S. Hanneke, R. El-Yaniv. A compression technique for analyzing disagreement-based active learning. *Journal of Machine Learning Research*, 2015.
- S. Hanneke. Refined error bounds for several learning algorithms. *Journal of Machine Learning Research*, 2016.
- E. Friedman. Active learning for smooth problems. COLT 2009.
- S. Mahalanabis. Subset and Sample Selection for Graphical Models: Gaussian Processes, Ising Models and Gaussian Mixture Models. PhD Thesis, University of Rochester, 2012.
- S. Hanneke. Theory of Disagreement-Based Active Learning. *Foundations and Trends in Machine Learning*, 2014.
- S. Hanneke, L. Yang. Surrogate losses in passive and active learning. arXiv:1207.3772.

Part 3: Beyond Disagreement-Based Active Learning – Current Directions

- Subregion-Based Active Learning
- Margin-Based Active Learning: Linear Separators
- Shattering-Based Active Learning
- Distribution-Free Analysis, Optimality
- TicToc: Adapting to Heterogeneous Noise
- Tsybakov Noise

**Tutorial on Active Learning:
Theory to Practice**

Steve Hanneke

Toyota Technological Institute at Chicago
steve.hanneke@gmail.com

Robert Nowak

University of Wisconsin - Madison
rdnowak@wisc.edu

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

A² (Agnostic Active)

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S
2. **label** points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. **optimize** $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. **reduce** \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

Subregion-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \mathcal{R}_{\epsilon'_t}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Instead, pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.
 $\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

Pick ϵ' carefully each round,
 $R(\hat{f}) - R(f^*) \leq \epsilon$ at end

e.g., Bounded noise: $\epsilon' \propto d2^{-t}$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

Subregion-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \mathcal{R}_{\epsilon'_t}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using $\#$ labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

$$\text{DIS}(\mathcal{H}) := \{x \in \mathcal{X} : \exists f, f' \in \mathcal{H}, f(x) \neq f'(x)\}$$

Subregion-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \mathcal{R}_{\epsilon'_t}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Pick region $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

$$\textbf{Agnostic case: } \varphi'_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, 2\beta + r)))}{2\beta + r}$$

Theorem:

$$\begin{aligned} R(\hat{f}) &\leq R(f^*) + \epsilon \text{ using # labels} \\ &\approx \varphi'_c d \frac{\beta^2}{\epsilon^2} \end{aligned}$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- $\mathcal{R}_{\epsilon'}(\mathcal{H}) = \text{DIS}(\mathcal{H})$ works
- Empirically (Zhang & Chaudhuri, 2014)
- Nice structure: e.g., **Linear separators**

Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r>\epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- $\mathcal{R}_{\epsilon'}(\mathcal{H}) = \text{DIS}(\mathcal{H})$ works
- Empirically (Zhang & Chaudhuri, 2014)
- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r>\epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

Zhang & Chaudhuri, 2014

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r>\epsilon} \frac{P_X(\mathcal{R}_{r/c}(B(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels

$$\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$$

Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

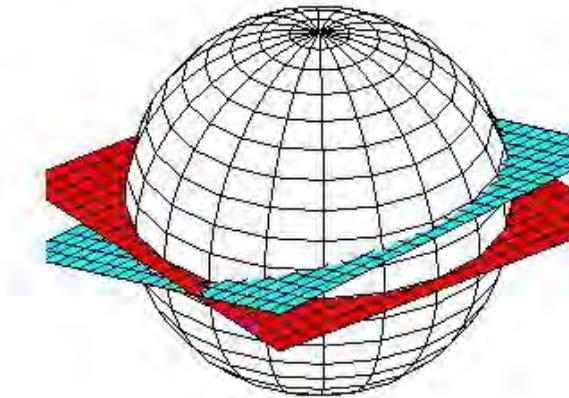
(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

Uniform P_X on d -dim sphere

For $w \in B(w^*, r)$, **project** to $\text{Span}(w, w^*)$



Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

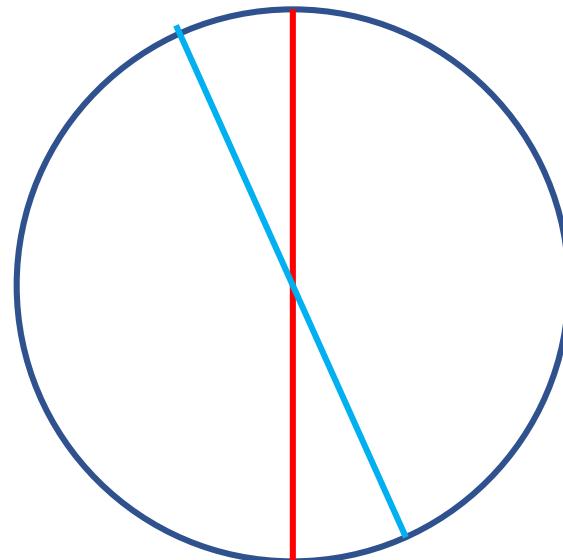
Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

Uniform P_X on d -dim sphere

For $w \in B(w^*, r)$, **project** to $\text{Span}(w, w^*)$

Most projected prob mass toward middle



Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

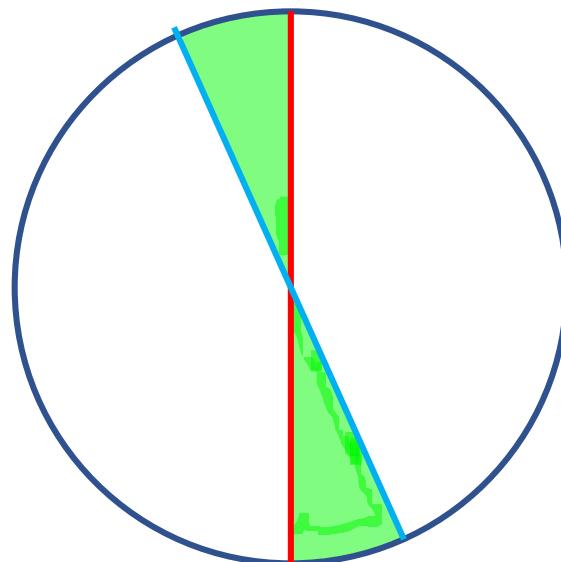
Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

Uniform P_X on d -dim sphere

For $w \in B(w^*, r)$, **project** to $\text{Span}(w, w^*)$

Most projected prob mass toward middle



Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

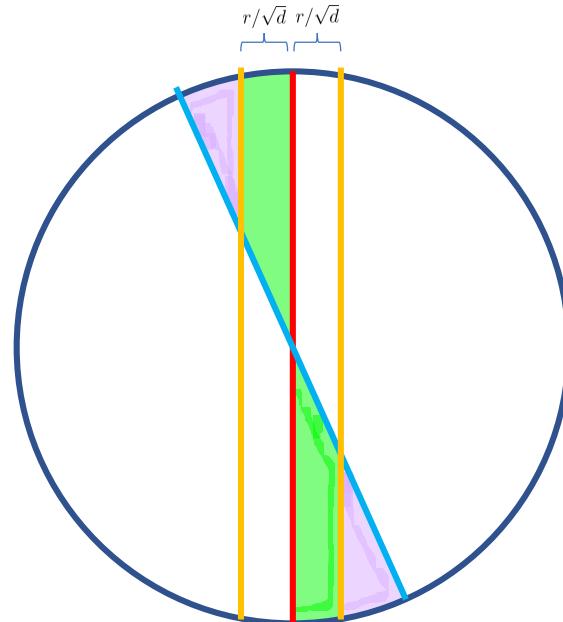
Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

Uniform P_X on d -dim sphere

For $w \in B(w^*, r)$, **project** to $\text{Span}(w, w^*)$

Most projected prob mass toward middle



$\text{DIS}(\{w, w^*\})$ in
slab of width $\approx r$

Most of its prob in
slab of width $\approx r/\sqrt{d}$

Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

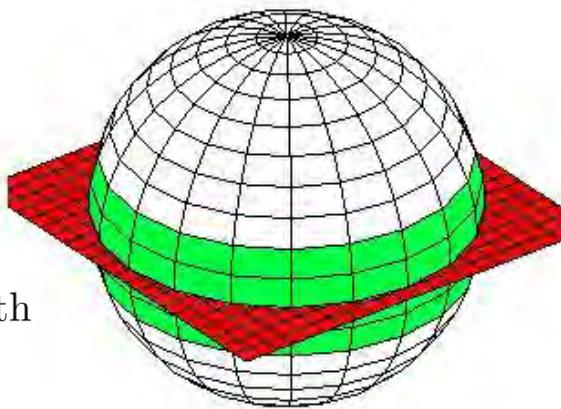
(Dasgupta, Kalai, Monteleoni, 2005;
Balcan, Broder, Zhang, 2007; ...)

$\text{DIS}(\mathcal{B}(f^*, r)) =$
slab of width $\approx r$

Take $\mathcal{R}_{r/c}(\mathcal{B}(f^*, r)) =$
slab of width $\approx r/\sqrt{d}$

Prob in slab $\approx \sqrt{d} \times \text{width}$

$\Rightarrow \varphi_c \leq \text{constant}$



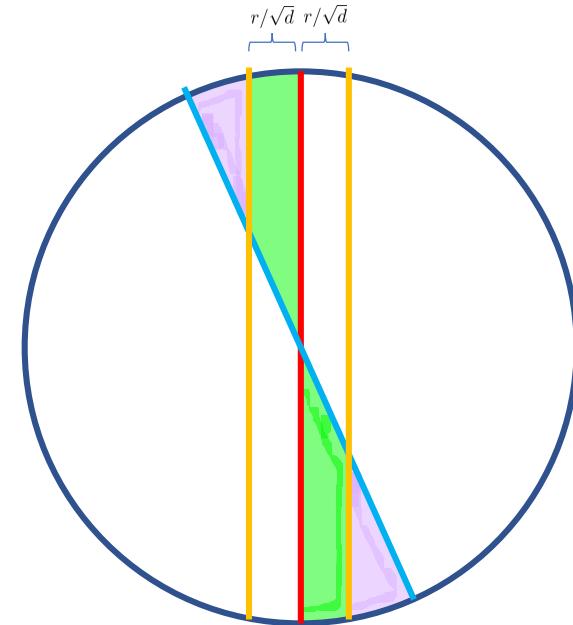
Pick **region** $\mathcal{R}_{\epsilon'}(\mathcal{H})$ s.t.

$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'$.

Uniform P_X on d -dim sphere

For $w \in \mathcal{B}(w^*, r)$, **project** to $\text{Span}(w, w^*)$

Most projected prob mass toward middle



$\text{DIS}(\{w, w^*\})$ in
slab of width $\approx r$

Most of its prob in
slab of width $\approx r/\sqrt{d}$

Subregion-Based Active Learning

How to find such an $\mathcal{R}_{\epsilon'}(\mathcal{H})$?

- Nice structure: e.g., **Linear separators**

Margin-based Active Learning

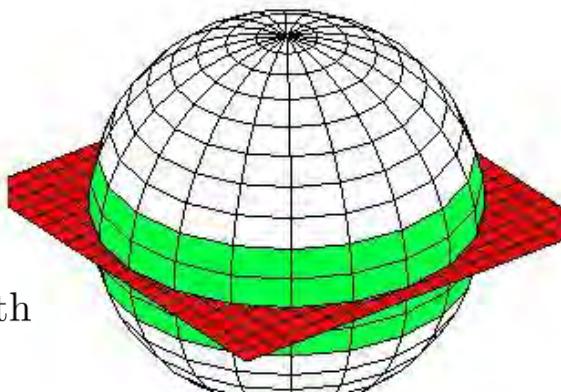
(Dasgupta, Kalai, Monteleoni, 2005;
Balcan - - - - -)

$\text{DIS}(\mathcal{B}(f^*, r)) =$
slab of width $\approx r$

Take $\mathcal{R}_{r/c}(\mathcal{B}(f^*, r)) =$
slab of width $\approx r/\sqrt{d}$

Prob in slab $\approx \sqrt{d} \times \text{width}$

$\Rightarrow \varphi_c \leq \text{constant}$



Pick **region $\mathcal{R}_{\epsilon'}(\mathcal{H})$** s.t.

$$\forall f, f' \in \mathcal{H}, P_X(x \notin \mathcal{R}_{\epsilon'}(\mathcal{H}) : f(x) \neq f'(x)) \leq \epsilon'.$$

$$\varphi_c := \sup_{r > \epsilon} \frac{P_X(\mathcal{R}_{r/c}(\mathcal{B}(f^*, r)))}{r}$$

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels
 $\approx \varphi_c d \log\left(\frac{1}{\epsilon}\right)$

\Rightarrow # labels $\approx d \log\left(\frac{1}{\epsilon}\right)$ suffice

Comparison:

Recall $\theta \approx \sqrt{d}$

$\Rightarrow A^2$ # labels $\approx d^{3/2} \log\left(\frac{1}{\epsilon}\right)$

Recall:
Passive $\approx \frac{d}{\epsilon}$

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

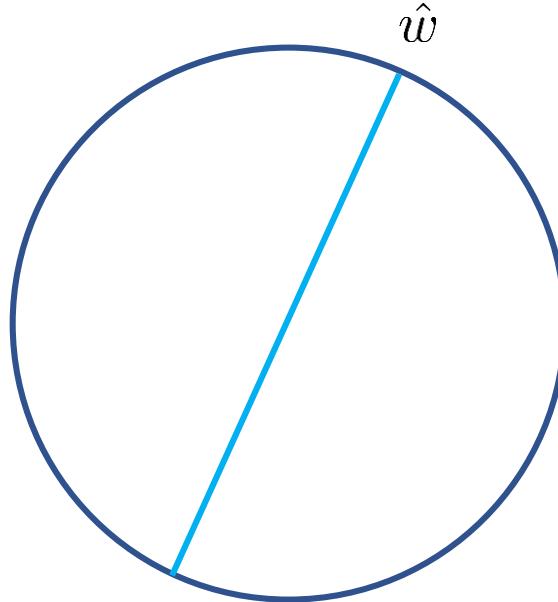
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

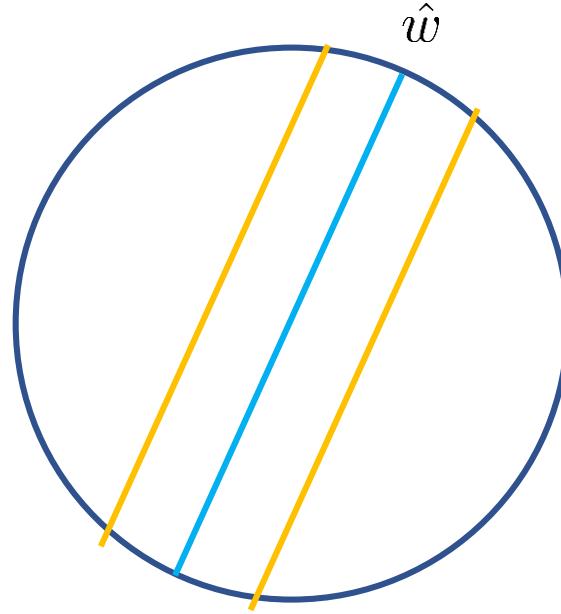
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

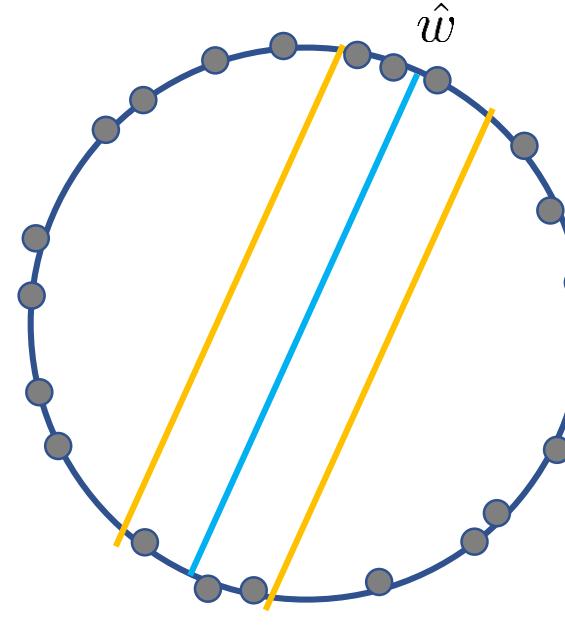
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

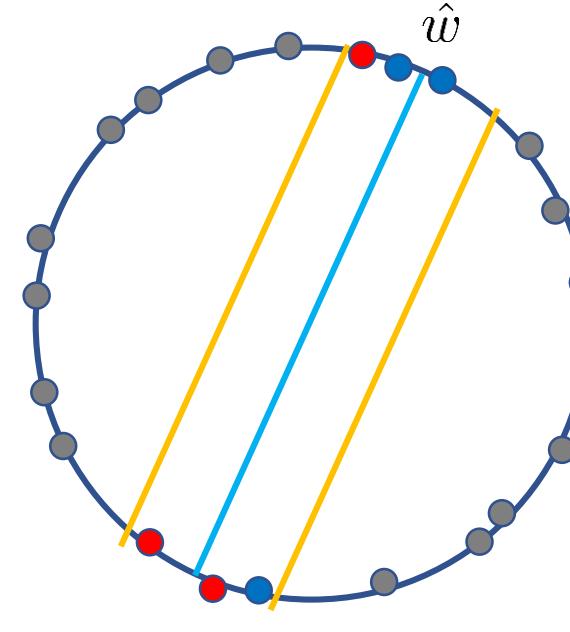
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

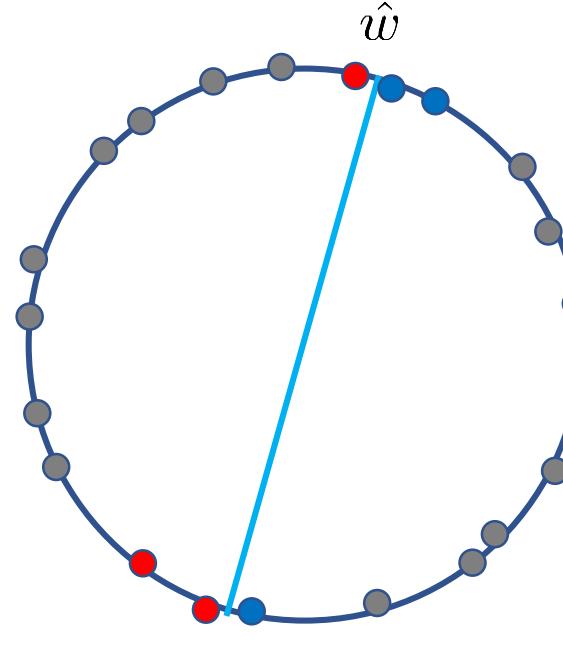
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Margin-Based Active Learning

(Balcan, Broder, Zhang, 2007; ...)

Margin-based Active Learning

Initialize \hat{w}

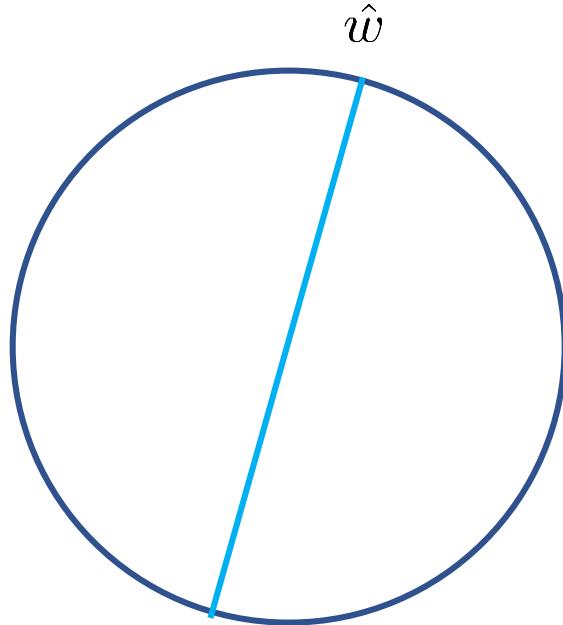
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q(w)$

output final \hat{w}



Uniform P_X on d -dim sphere

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels
 $\approx d \log\left(\frac{1}{\epsilon}\right)$

(also works for isotropic log-concave distributions)

Computational Efficiency

(Awasthi, Balcan, Long, 2014,...)

Uniform P_X on d -dim sphere

Efficient Alg

Initialize \hat{w}

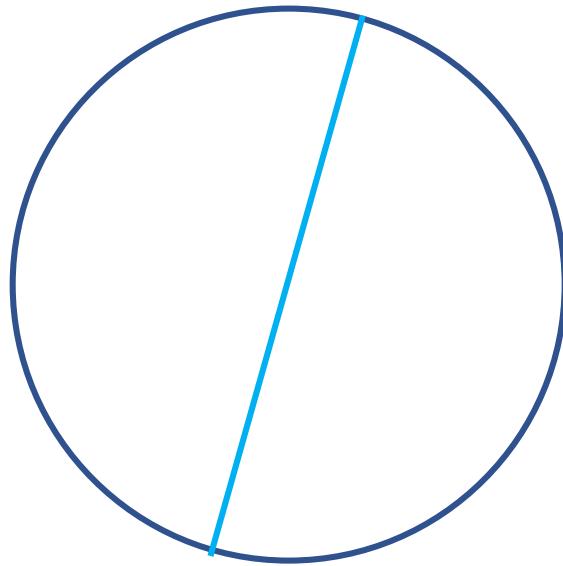
for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q^{\ell_t}(w)$

output final \hat{w}



Surrogate loss

$$\ell_t(\langle w, x \rangle, y) \approx \max\{1 - 2^t \sqrt{d}(y \langle w, x \rangle), 0\}$$

Hinge loss slope changes each round

Computational Efficiency

(Awasthi, Balcan, Long, 2014,...)

Uniform P_X on d -dim sphere

Efficient Alg

Initialize \hat{w}

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q^{\ell_t}(w)$

output final \hat{w}

Theorem: with **Bounded noise**,
 $R(\hat{f}) \leq R(f^*) + \epsilon$ using # labels
 $\approx d \log\left(\frac{1}{\epsilon}\right)$
and running in polynomial time

Surrogate loss

$$\ell_t(\langle w, x \rangle, y) \approx \max\{1 - 2^t \sqrt{d}(y \langle w, x \rangle), 0\}$$

Hinge loss slope changes each round

Computational Efficiency

(Awasthi, Balcan, Long, 2014,...)

Uniform P_X on d -dim sphere

Efficient Alg

Initialize \hat{w}

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample $d2^t$ unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t. } \langle \hat{w}, x \rangle \leq c2^{-t}/\sqrt{d}$

3. optimize $\hat{w} \leftarrow \underset{w: \|w - \hat{w}\| \leq c'2^{-t}}{\operatorname{argmin}} \hat{R}_Q^{\ell_t}(w)$

output final \hat{w}

Surrogate loss

$$\ell_t(\langle w, x \rangle, y) \approx \max\{1 - 2^t \sqrt{d}(y \langle w, x \rangle), 0\}$$

Hinge loss slope changes each round

Theorem: with **Bounded noise**,

$$R(\hat{f}) \leq R(f^*) + \epsilon \text{ using } \# \text{ labels} \\ \approx d \log\left(\frac{1}{\epsilon}\right)$$

and running in polynomial time

Theorem: with **Agnostic** case,

$$R(\hat{f}) \leq CR(f^*) \text{ in polynomial time}$$

(was first alg. known to achieve these; even passively)

(also works for isotropic log-concave distributions)

Up Next:
Shattering-Based Active Learning

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

$\text{DIS}(\mathcal{H})$ checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

A^2 (Agnostic Active)

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{DIS}(\mathcal{H}) \cap S$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}

$\text{DIS}(\mathcal{H})$ checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
4. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$.

output final \hat{f}

$\text{DIS}(\mathcal{H})$ checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Shattering-Based Active Learning

(Hanneke, 2009, 2012)

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S

2. label points in $Q = \text{all } x \in S \text{ s.t.}$

$$P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$$

3. add the remaining points $x \in S$ to Q with label

$$\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$$

4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$

5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

$\text{DIS}(\mathcal{H})$ checks for shattering 1 point.

Idea: Generalize to shattering ≥ 1 points.

Denote $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

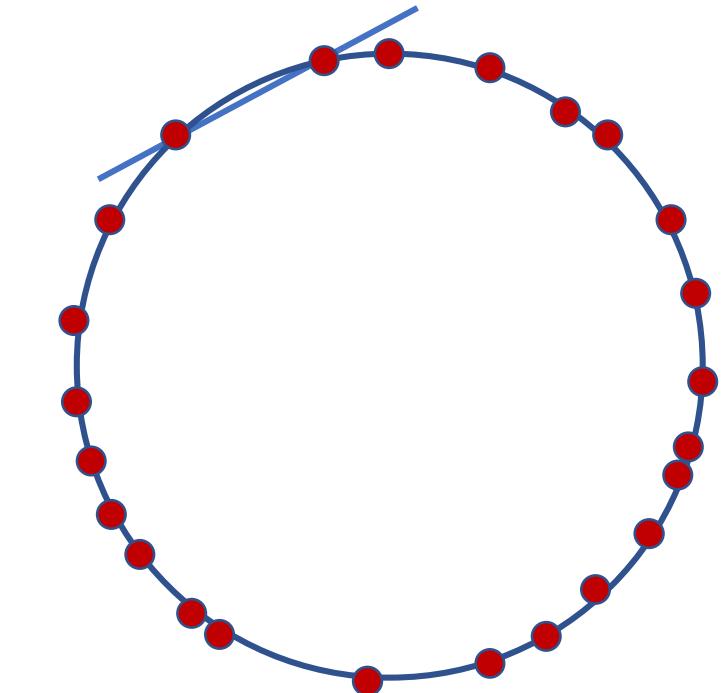
for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. add the remaining points $x \in S$ to Q with label $\hat{y}_x := \underset{y}{\operatorname{argmax}} P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. optimize $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$
5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all -1**

$\text{DIS}(\mathcal{H}) = \text{entire circle}$



Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
3. add the remaining points $x \in S$ to Q with label $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all -1**

$\text{DIS}(\mathcal{H}) = \text{entire circle}$

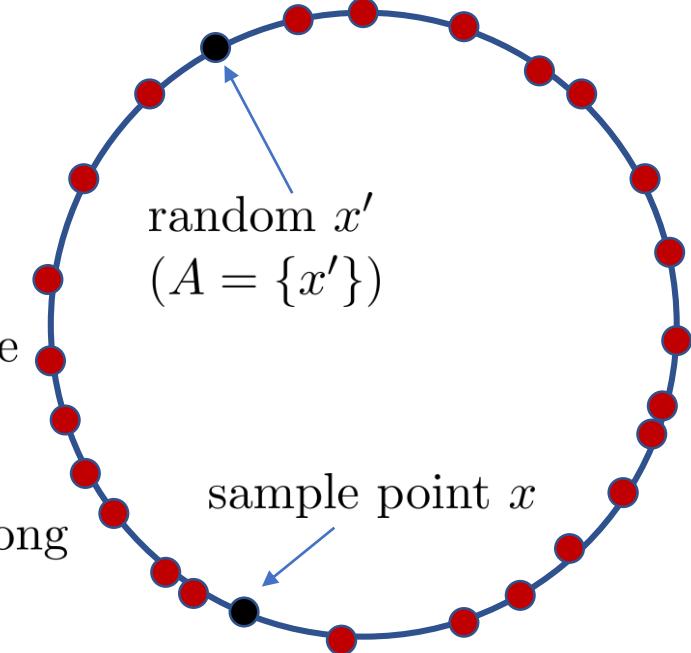
Try $k = 1$

Given sample x

Rand x' probably not close

Can't shatter $\{x, x'\}$
without a lot of points wrong

So won't query x



Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all -1**

$\text{DIS}(\mathcal{H}) = \text{entire circle}$

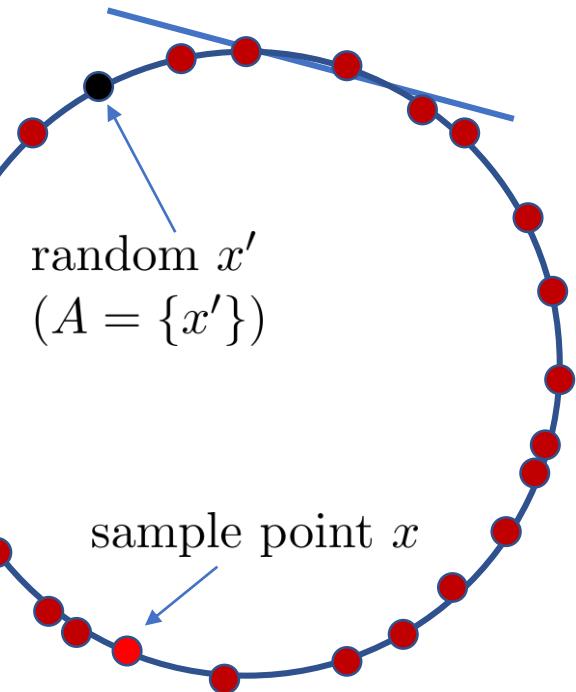
Try $k = 1$

Given sample x

Rand x' probably not close

Can't shatter $\{x, x'\}$
without a lot of points wrong

So won't query x



$\text{DIS}(\mathcal{H}_{x,-1})$ still entire circle (minus x)

$\text{DIS}(\mathcal{H}_{x,+1})$ **small** region

$$\Rightarrow \hat{y}_x = -1$$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all -1**

$\text{DIS}(\mathcal{H}) = \text{entire circle}$

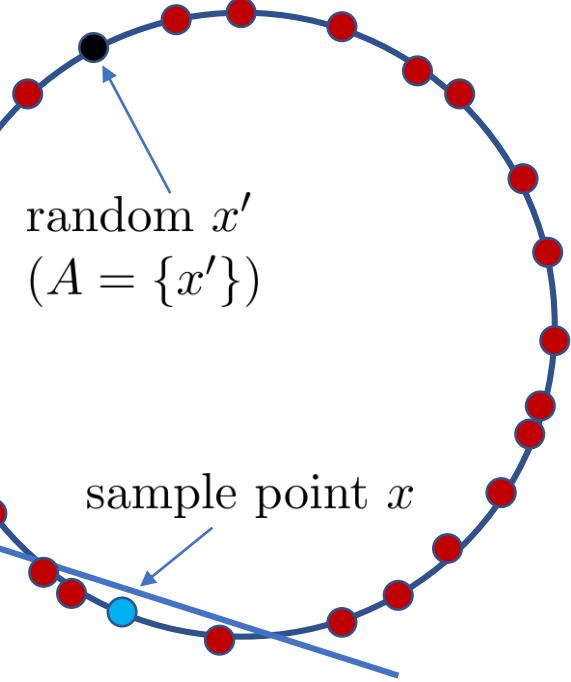
Try $k = 1$

Given sample x

Rand x' probably not close

Can't shatter $\{x, x'\}$
without a lot of points wrong

So won't query x



$\text{DIS}(\mathcal{H}_{x,-1})$ still entire circle (minus x)

$\text{DIS}(\mathcal{H}_{x,+1})$ **small** region

$$\Rightarrow \hat{y}_x = -1$$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til stopping-criterion)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Example: Linear separators, Uniform P_X on circle
Suppose true labels are **all -1**

$\text{DIS}(\mathcal{H}) = \text{entire circle}$

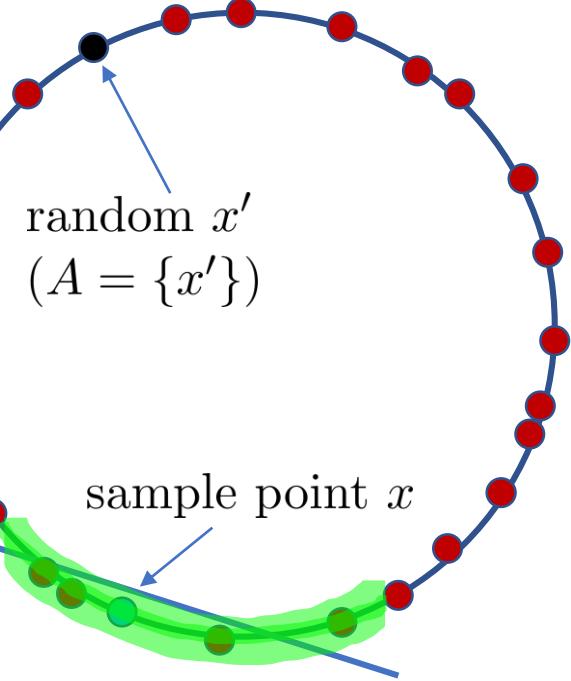
Try $k = 1$

Given sample x

Rand x' probably not close

Can't shatter $\{x, x'\}$
without a lot of points wrong

So won't query x



$\text{DIS}(\mathcal{H}_{x,-1})$ still entire circle (minus x)

$\text{DIS}(\mathcal{H}_{x,+1})$ **small** region

$$\Rightarrow \hat{y}_x = -1$$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. **sample** 2^t unlabeled points S

2. **label** points in $Q = \text{all } x \in S \text{ s.t.}$

$$P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$$

3. **add** the remaining points $x \in S$ to Q with label

$$\hat{y}_x := \underset{y}{\operatorname{argmax}} P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$$

4. **optimize** $\hat{f} \leftarrow \underset{f \in \mathcal{H}}{\operatorname{argmin}} \hat{R}_Q(f)$

5. **reduce** \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$

output final \hat{f}

Generally, need to try various k and pick one
(See the papers)

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label
 $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Generally, need to try various k and pick one
(See the papers)

$$\theta^{(k)} := \sup_{r > \epsilon} \frac{P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A)}{r}$$

$$\tilde{d} := \min \left\{ k : P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A) \xrightarrow[r \rightarrow 0]{} 0 \right\}$$

$$\tilde{\theta} := \theta^{(\tilde{d})}$$

Theorem: For Bounded noise, $R(\hat{f}) \leq R(f^*) + \epsilon$
with $\#$ labels

$$\approx C\tilde{\theta}d \log\left(\frac{1}{\epsilon}\right)$$

Note: $\tilde{\theta} \ll \frac{1}{\epsilon}$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label
 $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Generally, need to try various k and pick one
(See the papers)

$$\theta^{(k)} := \sup_{r > \epsilon} \frac{P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A)}{r}$$

$$\tilde{d} := \min \left\{ k : P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A) \xrightarrow[r \rightarrow 0]{} 0 \right\}$$

$$\tilde{\theta} := \theta^{(\tilde{d})}$$

Theorem: For Bounded noise, $R(\hat{f}) \leq R(f^*) + \epsilon$
with $\#$ labels

$$\approx C\tilde{\theta}d \log\left(\frac{1}{\epsilon}\right)$$

Note: $\tilde{\theta} \ll \frac{1}{\epsilon}$

In the example: $\tilde{\theta} = 2, \theta = \frac{1}{\epsilon}$

Shattering-Based Active Learning

Recall: \mathcal{H} **shatters** x_1, \dots, x_k if
all 2^k classifications realized by \mathcal{H}

Shattering-based Active Learning

for $t = 1, 2, \dots$ (til *stopping-criterion*)

1. sample 2^t unlabeled points S
 2. label points in $Q = \text{all } x \in S \text{ s.t. } P_X^k(A \in \mathcal{X}^k : \mathcal{H} \text{ shatters } A \cup \{x\} | \mathcal{H} \text{ shatters } A) \geq \frac{1}{2}$
 3. add the remaining points $x \in S$ to Q with label
 $\hat{y}_x := \operatorname{argmax}_y P_X^k(A \in \mathcal{X}^k : \mathcal{H}_{x,y} \text{ shatters } A | \mathcal{H} \text{ shatters } A)$
 4. optimize $\hat{f} \leftarrow \operatorname{argmin}_{f \in \mathcal{H}} \hat{R}_Q(f)$
 5. reduce \mathcal{H} : remove all f with $\hat{R}_Q(f) - \hat{R}_Q(\hat{f}) > \sqrt{\hat{P}_Q(f \neq \hat{f}) \frac{d}{|Q|}}$
- output final \hat{f}

Denoting $\mathcal{H}_{x,y} := \{h \in \mathcal{H} : h(x) = y\}$

Generally, need to try various k and pick one
(See the papers)

$$\theta^{(k)} := \sup_{r > \epsilon} \frac{P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A)}{r}$$

$$\tilde{d} := \min \left\{ k : P_X^k(A \in \mathcal{X}^k : \text{B}(f^*, r) \text{ shatters } A) \xrightarrow[r \rightarrow 0]{} 0 \right\}$$

$$\tilde{\theta} := \theta^{(\tilde{d})}$$

Theorem: For Bounded noise, $R(\hat{f}) \leq R(f^*) + \epsilon$
with $\#$ labels

$$\approx C\tilde{\theta}d \log\left(\frac{1}{\epsilon}\right)$$

Note: $\tilde{\theta} \ll \frac{1}{\epsilon}$ (may depend on f^* , P_X)

In the example: $\tilde{\theta} = 2$, $\theta = \frac{1}{\epsilon}$

Up Next:
Distribution-free Analysis

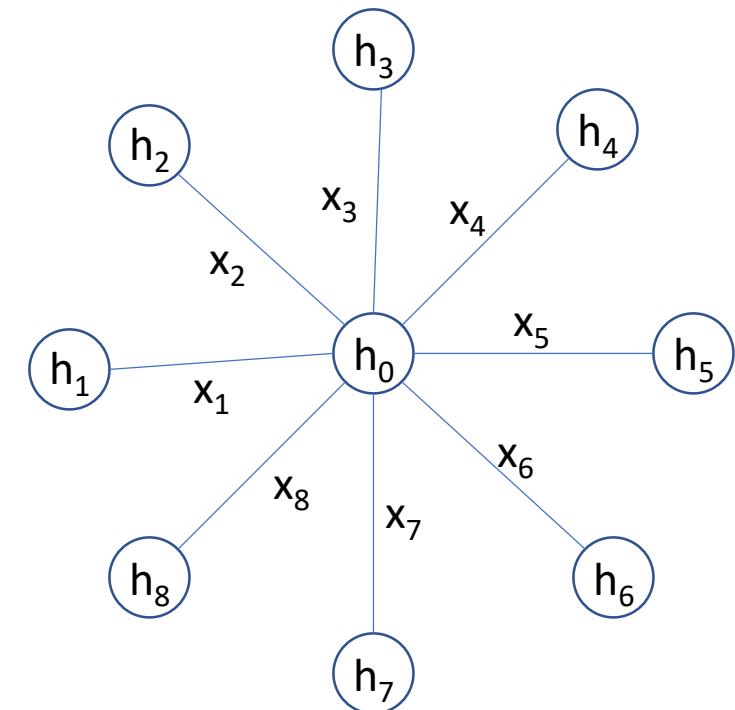
Distribution-Free Analysis

(Hanneke & Yang, 2015)

$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** s is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$,
 $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.



Distribution-Free Analysis

(Hanneke & Yang, 2015)

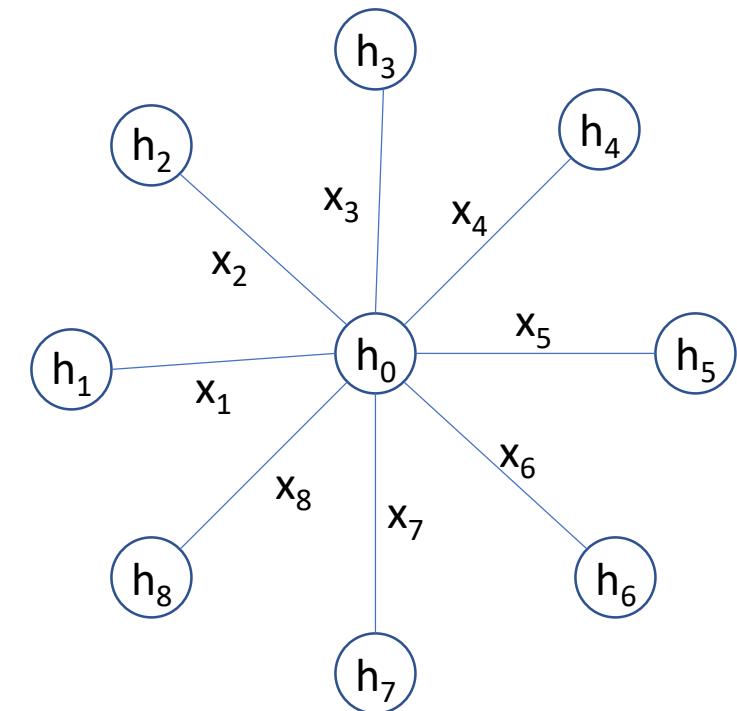
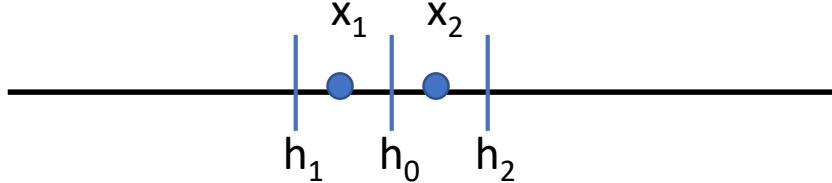
$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** s is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$, $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Example: Thresholds: $f(x) = \mathbb{I}[x \geq t]$.

$$s = 2.$$



Distribution-Free Analysis

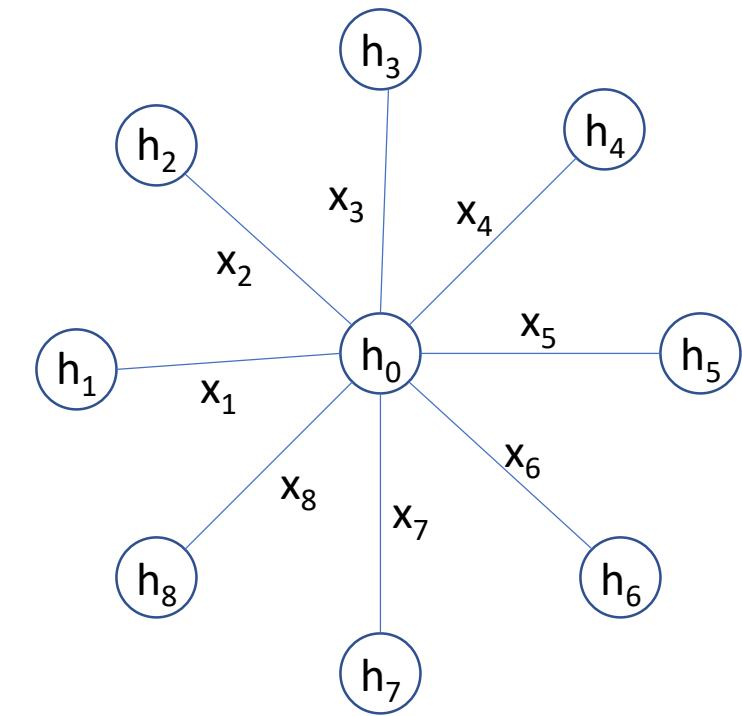
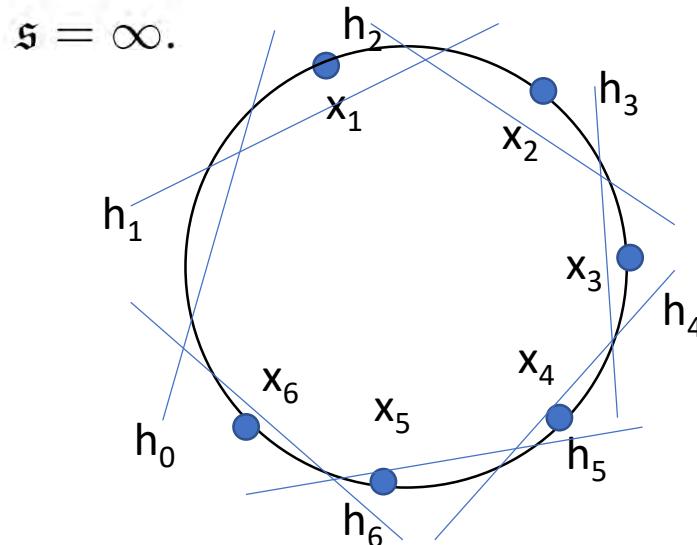
(Hanneke & Yang, 2015)

$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** s is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$, $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Example: Linear Separators in \mathbb{R}^n , $n \geq 2$:



Distribution-Free Analysis

(Hanneke & Yang, 2015)

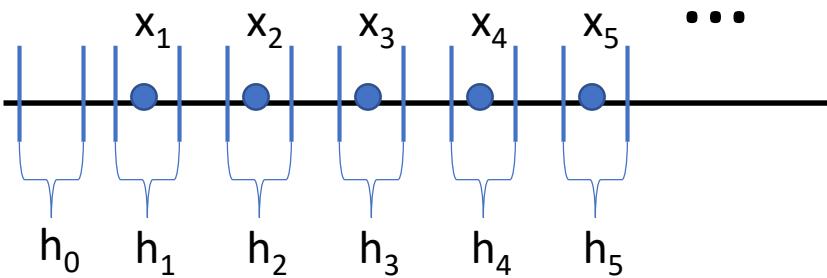
$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

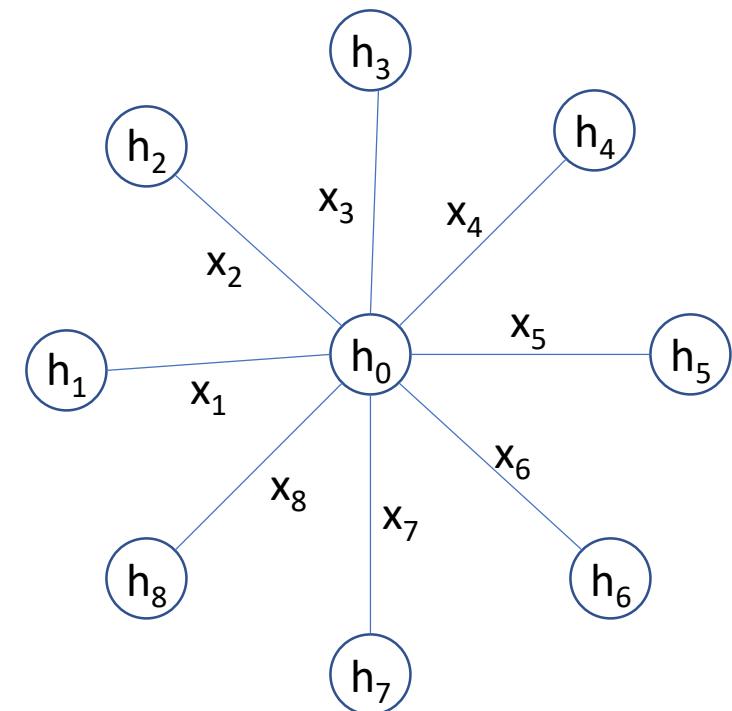
Definition: The **star number** \mathfrak{s} is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$, $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Example: Intervals: $x \mapsto \mathbb{I}[a \leq x \leq b]$

$$\mathfrak{s} = \infty.$$



Intervals of width w ($b - a = w > 0$) on $\mathcal{X} = [0, 1]$: $\mathfrak{s} \approx \lfloor \frac{1}{w} \rfloor$.



Distribution-Free Analysis

(Hanneke & Yang, 2015;
Hanneke, 2016)

$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** \mathfrak{s} is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$,
 $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

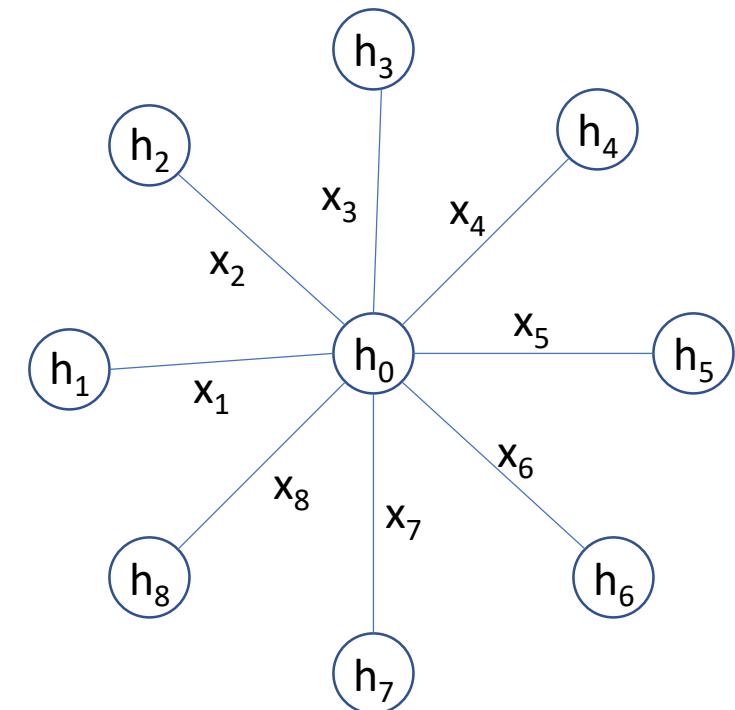
Theorem: $\sup_{P_X} \sup_{f^* \in \mathcal{H}} \theta = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \varphi_c = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \tilde{\theta} = \min\{\mathfrak{s}, \frac{1}{\epsilon}\} =: \mathfrak{s}_\epsilon$

Corollary:

Bounded noise # labels $\approx \mathfrak{s}_\epsilon d \log(\frac{1}{\epsilon})$

Agnostic ($\beta = R(f^*)$) # labels $\approx \mathfrak{s}_\beta d \frac{\beta^2}{\epsilon^2}$

Achieved by A^2



Distribution-Free Analysis

(Hanneke & Yang, 2015;
Hanneke, 2016)

$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** \mathfrak{s} is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$,
 $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Theorem: $\sup_{P_X} \sup_{f^* \in \mathcal{H}} \theta = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \varphi_c = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \tilde{\theta} = \min\{\mathfrak{s}, \frac{1}{\epsilon}\} =: \mathfrak{s}_\epsilon$

Corollary:

Bounded noise # labels $\approx \mathfrak{s}_\epsilon d \log(\frac{1}{\epsilon})$

Agnostic ($\beta = R(f^*)$) # labels $\approx \mathfrak{s}_\beta d \frac{\beta^2}{\epsilon^2}$

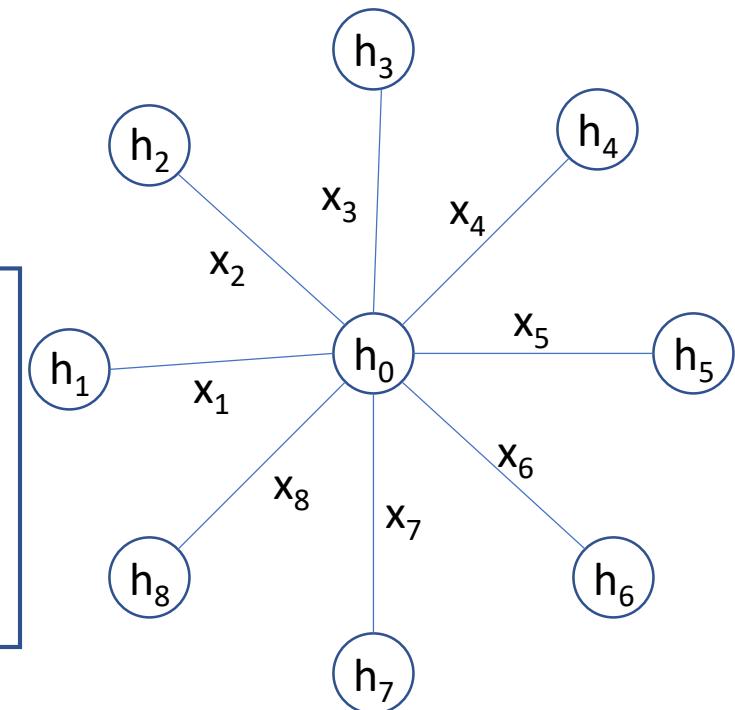
Achieved by A^2

Different alg., Bounded noise

labels $\approx \mathfrak{s}_{\epsilon/d} \log(\frac{1}{\epsilon})$

Near-matching **lower bound**:

$\mathfrak{s}_\epsilon + d \log(\frac{1}{\epsilon})$



Distribution-Free Analysis

(Hanneke & Yang, 2015;
Hanneke, 2016)

$\theta, \varphi, \tilde{\theta}$ depend on f^*, P_X .

Can we do sample complexity analysis **without** distribution-dependence?

Definition: The **star number** \mathfrak{s} is the largest k s.t. $\exists h_0, h_1, \dots, h_k \in \mathcal{H}$,
 $\exists x_1, \dots, x_k \in \mathcal{X}$ s.t. $\forall i \in \{1, \dots, k\}$, $\{x_j : h_i(x_j) \neq h_0(x_j)\} = \{x_i\}$.

Theorem: $\sup_{P_X} \sup_{f^* \in \mathcal{H}} \theta = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \varphi_c = \sup_{P_X} \sup_{f^* \in \mathcal{H}} \tilde{\theta} = \min\{\mathfrak{s}, \frac{1}{\epsilon}\} =: \mathfrak{s}_\epsilon$

Corollary:

Bounded noise # labels $\approx \mathfrak{s}_\epsilon d \log(\frac{1}{\epsilon})$

Agnostic ($\beta = R(f^*)$) # labels $\approx \mathfrak{s}_\beta d \frac{\beta^2}{\epsilon^2}$

Achieved by A^2

Different alg., Bounded noise

labels $\approx \mathfrak{s}_{\epsilon/d} \log(\frac{1}{\epsilon})$

Near-matching **lower bound**:

$\mathfrak{s}_\epsilon + d \log(\frac{1}{\epsilon})$

Open Question:

Agnostic ($\beta = R(f^*)$)

labels

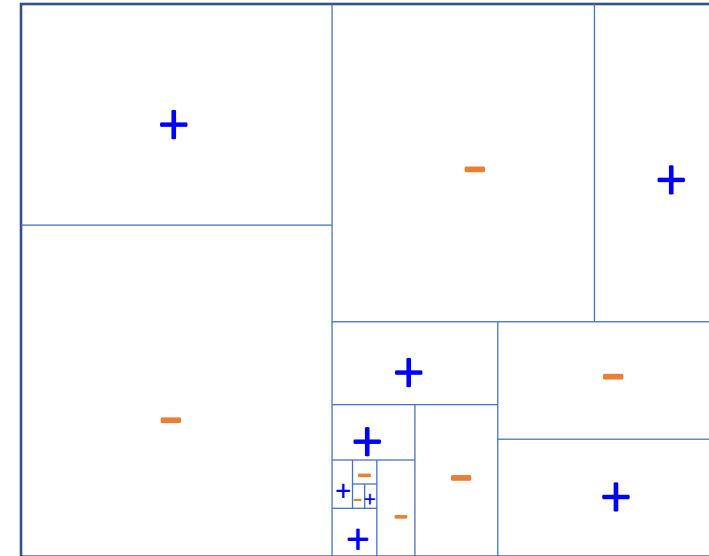
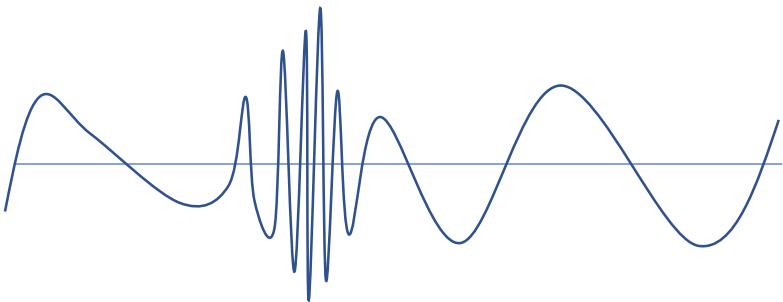
$\approx d \frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon/d} \log(\frac{1}{\epsilon})$?

lower bound:

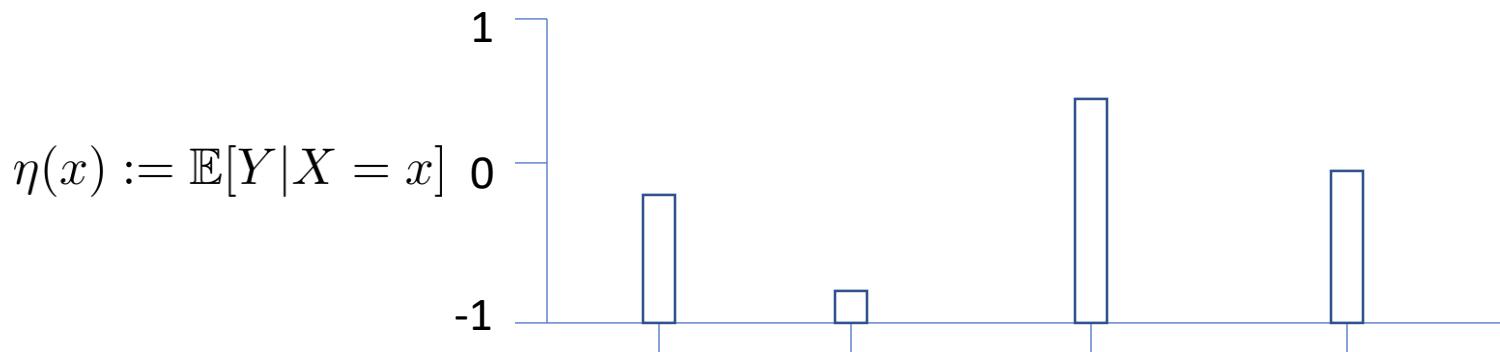
$d \frac{\beta^2}{\epsilon^2} + \mathfrak{s}_\epsilon + d \log(\frac{1}{\epsilon})$

Adapting to Heterogeneous Noise

So far: Active learning for spatial heterogeneity of **opt function**:



Also consider: Spatial heterogeneity of **noise**:



Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$

Input: Label budget n

Output: Classifier \hat{f}_n .

1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \dots$
3. $X_{s_m} \leftarrow \text{GETSEED}(\mathbb{L}, m)$
4. $\mathcal{L}_m \leftarrow \text{TICTOC}(X_{s_m}, m)$
5. if \mathcal{L}_m exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. If we've made n queries
7. Return $\hat{f}_n \leftarrow \text{LEARN}(\mathbb{L})$

An active learning alg.
(e.g. A^2)

Main new part

A passive learning alg.

Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$

Input: Label budget n

Output: Classifier \hat{f}_n .

1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \dots$
3. $X_{s_m} \leftarrow \text{GETSEED}(\mathbb{L}, m)$
4. $\mathcal{L}_m \leftarrow \text{TICTOC}(X_{s_m}, m)$
5. if \mathcal{L}_m exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. If we've made n queries
7. Return $\hat{f}_n \leftarrow \text{LEARN}(\mathbb{L})$

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Suppose f^* is the **global** optimal function: $f^*(x) = \text{sign}(\eta(x))$

TICTOC(X, m):

Query X (or nearby) to try to guess $f^*(X)$

If can figure it out, return that label

If can't figure it out by τ_m queries give up (don't return a label)

Focus queries on less-noisy points.

Double advantage:

- Focusing on the points we actually care about:

$$R(f|x) - R(f^*|x) = |\eta(x)|\mathbb{I}[f(x) \neq f^*(x)]$$

(small $|\eta(x)| \Rightarrow$ not much effect on $R(f|x)$ if $f(x) = f^*(x)$ or not).

- And those points require fewer queries to determine $f^*(X_i)$!

$\sim \frac{1}{\eta(X_i)^2}$ queries
to determine $f^*(X_i)$.

Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$

Input: Label budget n

Output: Classifier \hat{f}_n .

1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \dots$
3. $X_{s_m} \leftarrow \text{GETSEED}(\mathbb{L}, m)$
4. $\mathcal{L}_m \leftarrow \text{TICTOC}(X_{s_m}, m)$
5. if \mathcal{L}_m exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. If we've made n queries
7. Return $\hat{f}_n \leftarrow \text{LEARN}(\mathbb{L})$

Theorem: Bounded noise: # labels

$$\approx \mathfrak{s}_{\epsilon/d} \log\left(\frac{1}{\epsilon}\right)$$

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Suppose f^* is the **global** optimal function: $f^*(x) = \text{sign}(\eta(x))$

TICTOC(X, m):

Query X (or nearby) to try to guess $f^*(X)$

If can figure it out, return that label

If can't figure it out by τ_m queries give up (don't return a label)

Focus queries on less-noisy points.

Double advantage:

- Focusing on the points we actually care about:

$$R(f|x) - R(f^*|x) = |\eta(x)| \mathbb{I}[f(x) \neq f^*(x)]$$

(small $|\eta(x)| \Rightarrow$ not much effect on $R(f|x)$ if $f(x) = f^*(x)$ or not).

- And those points require fewer queries to determine $f^*(X_i)$!

$\sim \frac{1}{\eta(X_i)^2}$ queries
to determine $f^*(X_i)$.

Active Learning with TicToc

(Hanneke & Yang, 2015)

Algorithm: $\mathbb{A}(n)$

Input: Label budget n

Output: Classifier \hat{f}_n .

1. $\mathbb{L} \leftarrow \{\}$
2. For $m = 1, 2, \dots$
3. $X_{s_m} \leftarrow \text{GETSEED}(\mathbb{L}, m)$
4. $\mathcal{L}_m \leftarrow \text{TICTOC}(X_{s_m}, m)$
5. if \mathcal{L}_m exists, $\mathbb{L} \leftarrow \mathbb{L} \cup \{(s_m, \mathcal{L}_m)\}$
6. If we've made n queries
7. Return $\hat{f}_n \leftarrow \text{LEARN}(\mathbb{L})$

Theorem: Agnostic ($\beta = R(f^*)$)

and suppose f^* = global best:

labels

$$\approx d \frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon/d} \log\left(\frac{1}{\epsilon}\right)$$

Confirms agnostic sample complexity conjecture
but with extra assumption f^* = global opt.

Near-match lower bound: $d \frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon} + d \log\left(\frac{1}{\epsilon}\right)$

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Suppose f^* is the **global** optimal function: $f^*(x) = \text{sign}(\eta(x))$

TICTOC(X, m):

Query X (or nearby) to try to guess $f^*(X)$

If can figure it out, return that label

If can't figure it out by τ_m queries give up (don't return a label)

Focus queries on less-noisy points.

Double advantage:

- Focusing on the points we actually care about:

$$R(f|x) - R(f^*|x) = |\eta(x)| \mathbb{I}[f(x) \neq f^*(x)]$$

(small $|\eta(x)| \Rightarrow$ not much effect on $R(f|x)$ if $f(x) = f^*(x)$ or not).

- And those points require fewer queries to determine $f^*(X_i)$!

$\sim \frac{1}{\eta(X_i)^2}$ queries
to determine $f^*(X_i)$.

Principles of Active Learning

1. Query in dense regions where \hat{f} could disagree a lot with f^*
2. Query in regions with low noise

Tsybakov Noise

The alg. adapts to heterogeneity in the noise.

Let's try it with a model that explicitly describes heterogeneous noise:

Tsybakov Noise

Tsybakov Noise

(Tsybakov, 2004;
Mammen & Tsybakov 1999)

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Definition: (Tsybakov noise)

$f^*(x) = \text{sign}(\eta(x))$ and $\exists \alpha \in (0, 1)$ s.t. $\forall \tau > 0$,

$$P_X(x : |\eta(x)| \leq \tau) \lesssim \tau^{\frac{\alpha}{1-\alpha}}.$$

Tsybakov Noise

(Tsybakov, 2004;
Mammen & Tsybakov 1999)

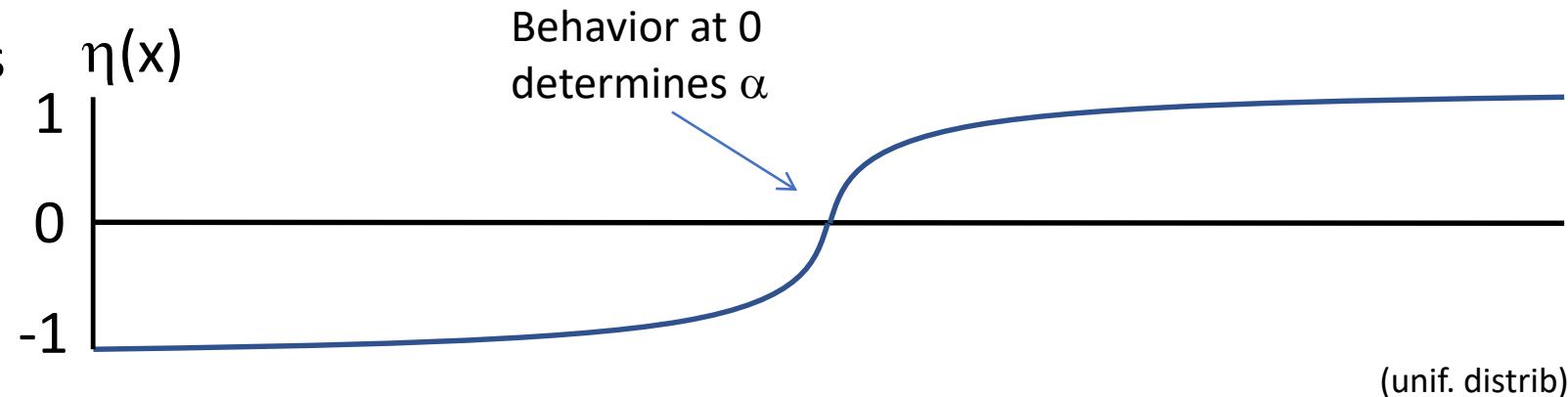
Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Definition: (Tsybakov noise)

$f^*(x) = \text{sign}(\eta(x))$ and $\exists \alpha \in (0, 1)$ s.t. $\forall \tau > 0$,
 $P_X(x : |\eta(x)| \leq \tau) \lesssim \tau^{\frac{\alpha}{1-\alpha}}$.

Example:

Thresholds



Tsybakov Noise

Denote $\eta(x) = \mathbb{E}[Y|X = x]$

Definition: (Tsybakov noise)

$f^*(x) = \text{sign}(\eta(x))$ and $\exists \alpha \in (0, 1)$ s.t. $\forall \tau > 0$,

$$P_X(x : |\eta(x)| \leq \tau) \lesssim \tau^{\frac{\alpha}{1-\alpha}}.$$

Passive OPT: $\tilde{\Theta}\left(\frac{d}{\epsilon^{2-\alpha}}\right)$.

(Massart & Nédélec, 2006)

Active OPT: (roughly) $\begin{cases} \frac{d}{\epsilon^{2-2\alpha}} & \text{if } 0 < \alpha \leq 1/2 \\ \min \left\{ \frac{d}{\epsilon^{2-2\alpha}} \left(\frac{s}{d} \right)^{2\alpha-1}, \frac{d}{\epsilon} \right\} & \text{if } 1/2 < \alpha < 1 \end{cases}$

(Hanneke & Yang, 2015)

$$\sim \begin{cases} \frac{1}{\epsilon^{2-2\alpha}}, & \text{if } s < \infty \\ \frac{1}{\epsilon}, & \text{if } s = \infty \end{cases}.$$

Active Opt \ll Passive Opt.
(always)

Conclusions

- Many proposals for going beyond Disagreement-based Active Learning
- Each exhibits improvements in certain cases
- We still don't know the **optimal agnostic active learning algorithm**

$$d \frac{\beta^2}{\epsilon^2} + \mathfrak{s}_{\epsilon/d} \log\left(\frac{1}{\epsilon}\right)$$

Questions?

Further reading:

- S. Dasgupta, A. Kalai, C. Monteleoni. Analysis of perceptron-based active learning. *COLT* 2005.
- M. F. Balcan, A. Broder, T. Zhang. Margin based active learning. *COLT* 2007.
- P. Awasthi, M. F. Balcan, P. Long. *Journal of the ACM*, 2017.
- S. Hanneke. Theoretical Foundations of Active Learning. PhD Thesis, CMU, 2009.
- S. Hanneke. Activized learning: Transforming passive to active with improved label complexity. *Journal of Machine Learning Research*, 2012.
- C. Zhang, K. Chaudhuri. Beyond disagreement-based agnostic active learning. *NeurIPS* 2014.
- R. M. Castro, R. D. Nowak. Minimax bounds for active learning. *IEEE Transactions on Information Theory*, 2008.
- R. M. Castro, R.D. Nowak. Upper and lower error bounds for active learning. *Allerton* 2006.
- S. Dasgupta. Coarse sample complexity bounds for active learning. *NeurIPS* 2005.
- S. Hanneke, L. Yang. Minimax analysis of active learning. *Journal of Machine Learning Research*, 2015.
- S. Hanneke. Refined error bounds for several learning algorithms. *Journal of Machine Learning Research*, 2016.
- M. F. Balcan, S. Hanneke, J. Wortman Vaughan. The true sample complexity of active learning. *Machine Learning*, 2010.