Active Learning

Lecture 17

COMS-4771

Using Sanjoy Dasgupta's slides

Active Learning Recap

- The learner chooses which examples it wants labeled

- The learner works harder in order to use fewer labeled examples

Basic setting

[Cohn, Atlas, and Ladner, 1992]

Underlying distribution P on the (x,y) data.

Learner has two abilities:
-- draw an unlabeled sample from the distribution
-- ask for a label *of one of these samples*

The error of any classifier h is measured on distribution P:

$$err(h) = P(h(x) \neq y)$$

Special case to simplify matters: assume the data is *separable*, ie. some concept $h \in H$ labels all points perfectly.

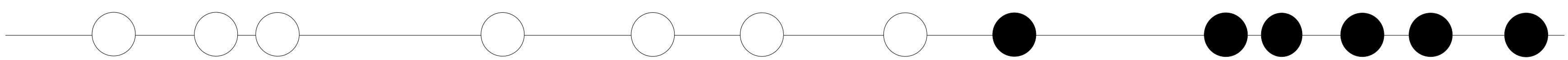
Why hope for success?

Simple hypothesis class H: thresholds on the real line

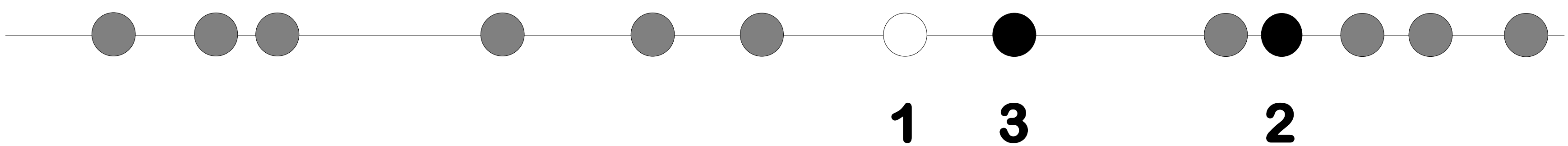
$H = \{h_w: w \text{ in } [0,1]\}$, where $h_w(x) = 1$ if $x>w$; 0 otherwise

Data is linearly separable (there is a perfect threshold)

Passive learning needs roughly $m = O(1/\varepsilon)$ random labeled points to reach a hypothesis with error rate $<\varepsilon$



Binary search needs just $\log(m) = O(\log 1/\varepsilon)$ labels



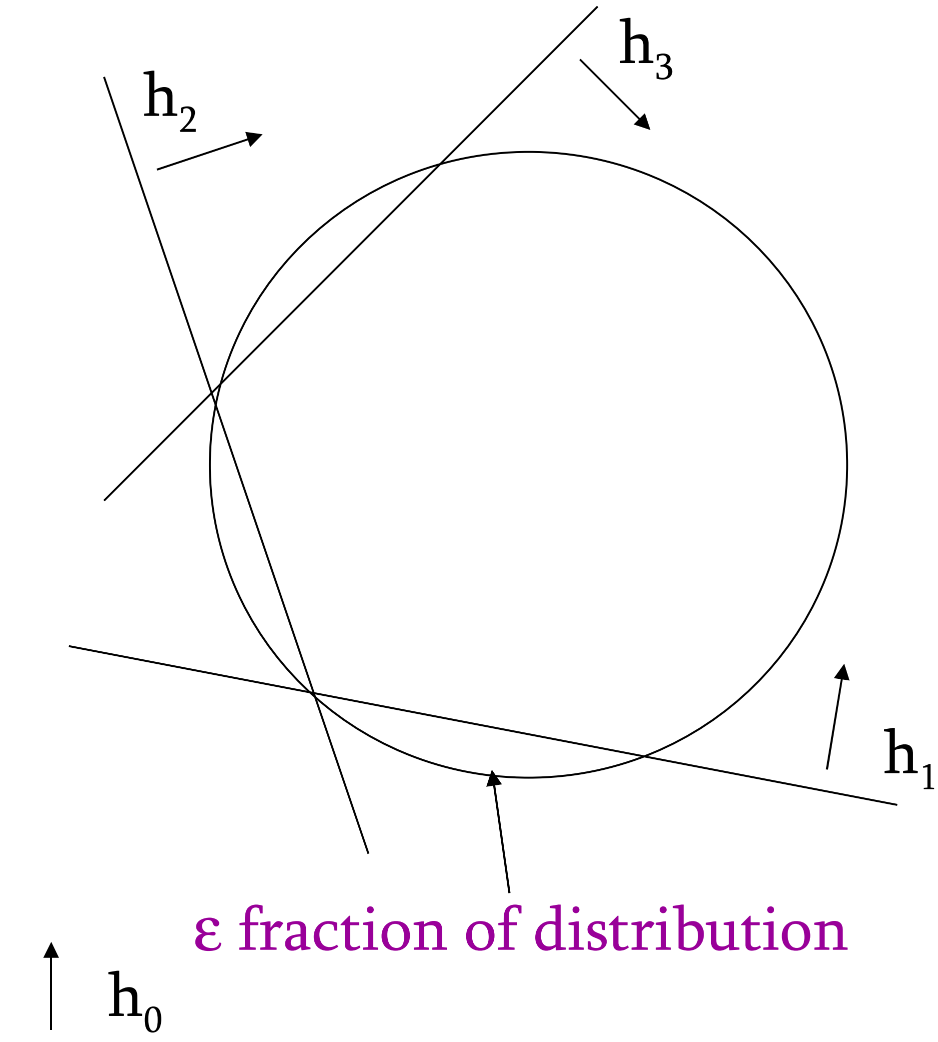
An exponential improvement!

Bad news

For linear separators in R^1, need just log 1/ε labels.
But when H = {linear separators in R^2}: some target hypotheses require 1/ε labels to be queried!

Consider *any* distribution over the circle in R^2.

Need 1/ε labels to distinguish between $h_0, h_1, h_2, \ldots, h_{1/\varepsilon}$!



Basic Notions

Current version space H_i --- part of H still under consideration by the algorithm

Region of uncertainty R_i --- region of the data space about which there is still some disagreement within H_i

Volume of R_i :

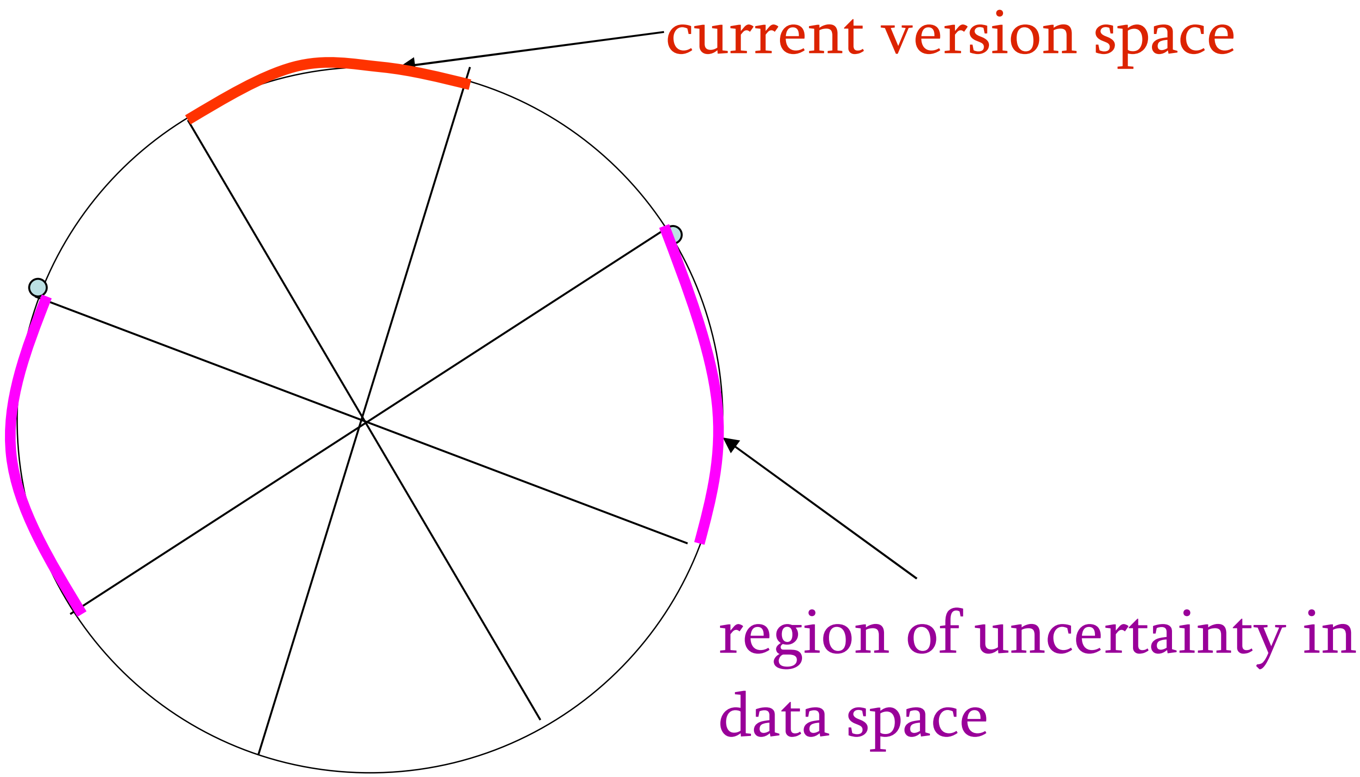
$$\text{Disagree}_P(H_i) = \Pr_{x \sim P} [\, \exists\, h_1, h_2 \in H_i : h_1(x) \neq h_2(x)]$$

Region of uncertainty

In the realizable case, current version space is the portion of H consistent with labels so far.

Suppose data lies on unit circle in R^2; hypotheses are linear separators.

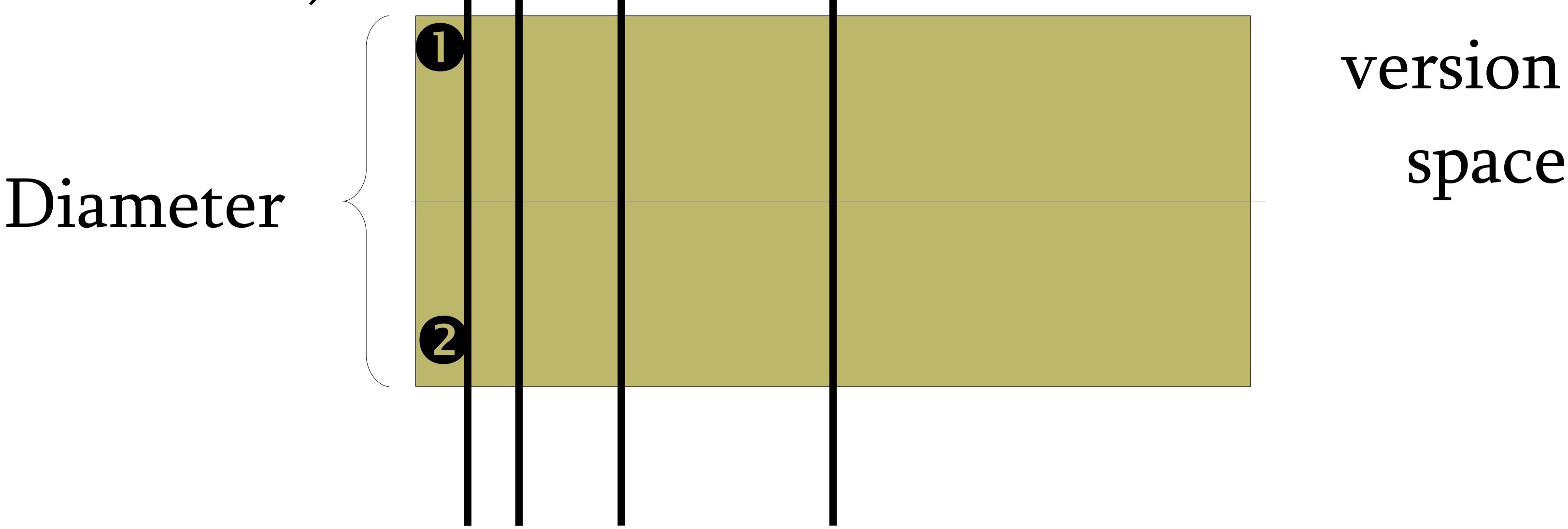
(spaces X, H superimposed)



Uncertainty sampling

First idea: Try to rapidly reduce the volume of the version space

Problem: ignores the data distribution --- reducing the volume may have little effect on the diameter (and thus error)!



Distance measure on H: $d(h, h') = \Pr_{x \sim P}[\, h(x) \neq h'(x) \,]$

What we really want to cut is the diameter with respect to d.

Query by Committee

[Seung, Opper, Sompolinsky '92; Freund et al '97]

Elegant scheme which decreases volume in a manner which is sensitive to the data distribution.

Main idea: Sample an unlabeled point; query if two random hypotheses h, h' in H_i disagree on the label.

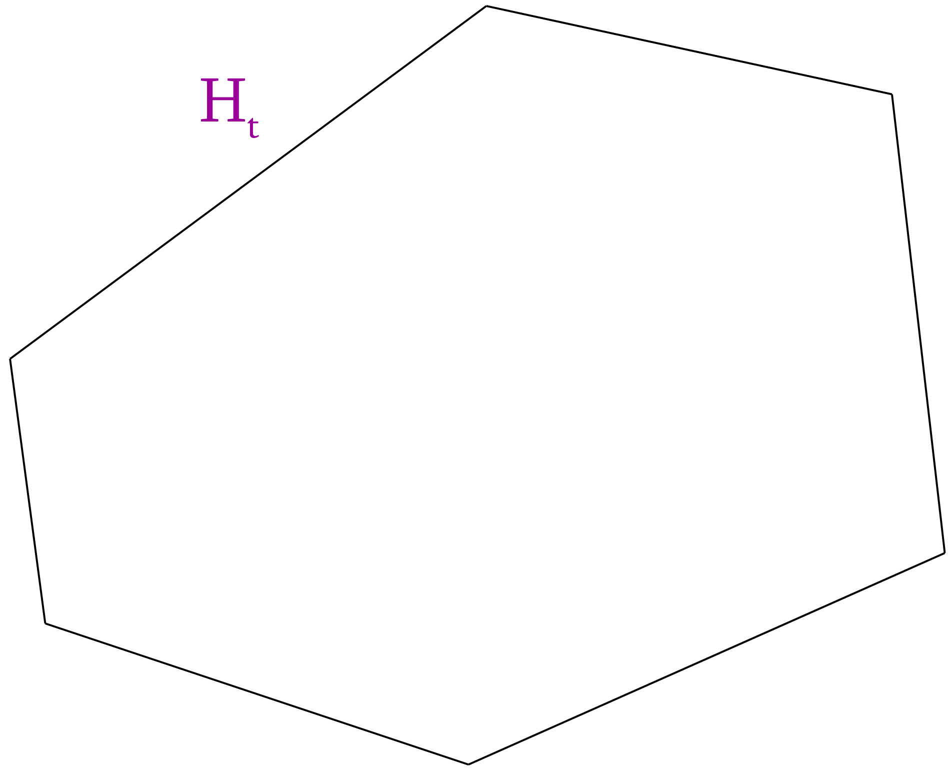
1) The stronger the disagreement on x, the higher the probability of querying it (the higher the expected reduction in volume).
2) The probability of querying when h and h' are drawn is d(h,h').

Label bound: For H = {linear separators in R^d}, P = uniform distribution, just d log 1/ε labels to reach a hypothesis with error < ε. (Compare to $O(d/\epsilon)$ in the supervised setting.)

Query by committee

Implementation: need to randomly pick h according to (π, H_t).

How do you pick a random point from a convex body?



By random walk!

2. Ball walk
3. Hit-and-run

[Gilad-Bachrach, Navot, Tishby 2005]

Online active learning

Online algorithms:

see unlabeled data streaming by, one point at a time

can query current point's label, at a cost

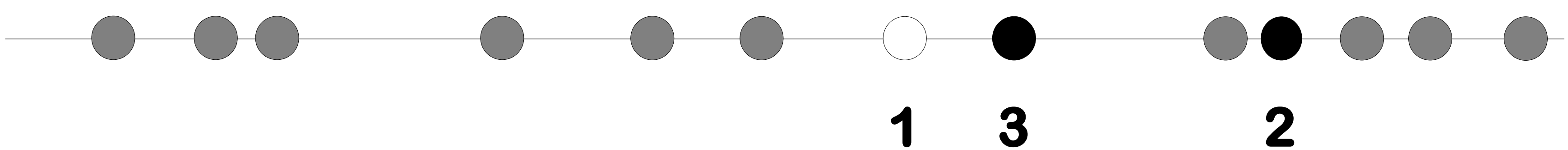
can only maintain current hypothesis (memory bound)

[Dasgupta, Kalai, Monteleoni 2005]: An active version of the perceptron algorithm.

Guarantee: In the realizable case, for linear separators under the uniform distribution, label complexity is d log 1/ε.

What if there is noise?
Need a robust active learner

A few mistakes can
induce a large error.



In fact, Active Learning is noise-seeking:

Active learners quickly go to the decision boundary and that's where noise often is.

Why?---mismatch between the input distribution and the hypotheses class; large conditional noise rate

Active learners are sensitive to noise since they try to minimize redundancy

Setup: Agnostic Learning

Hypothesis class: H

Noise rate

Goal: Find $h \in H$ with

$$err_D(h) \le opt + \varepsilon$$

Arbitrary distribution **D** over $X \times Y$

$$err_D(h) = Pr_{(x,y)\sim D}[h(x) \neq y]$$

$$opt = \min err_D(h)$$

Ideally, we don't want to make any assumptions about the mechanism producing noise!

Why is the agnostic case difficult?

Separable case:

We don't care about the query distribution we induce.

We have a promise that there is a hypothesis in H consistent with all, so any inconsistent hypothesis can be immediately discarded.

Agnostic case:

If the query distribution is far from the input distribution, a hypothesis that performs badly on the query points may be the best hypothesis in the class!

Q: Is Robust Active Learning possible?

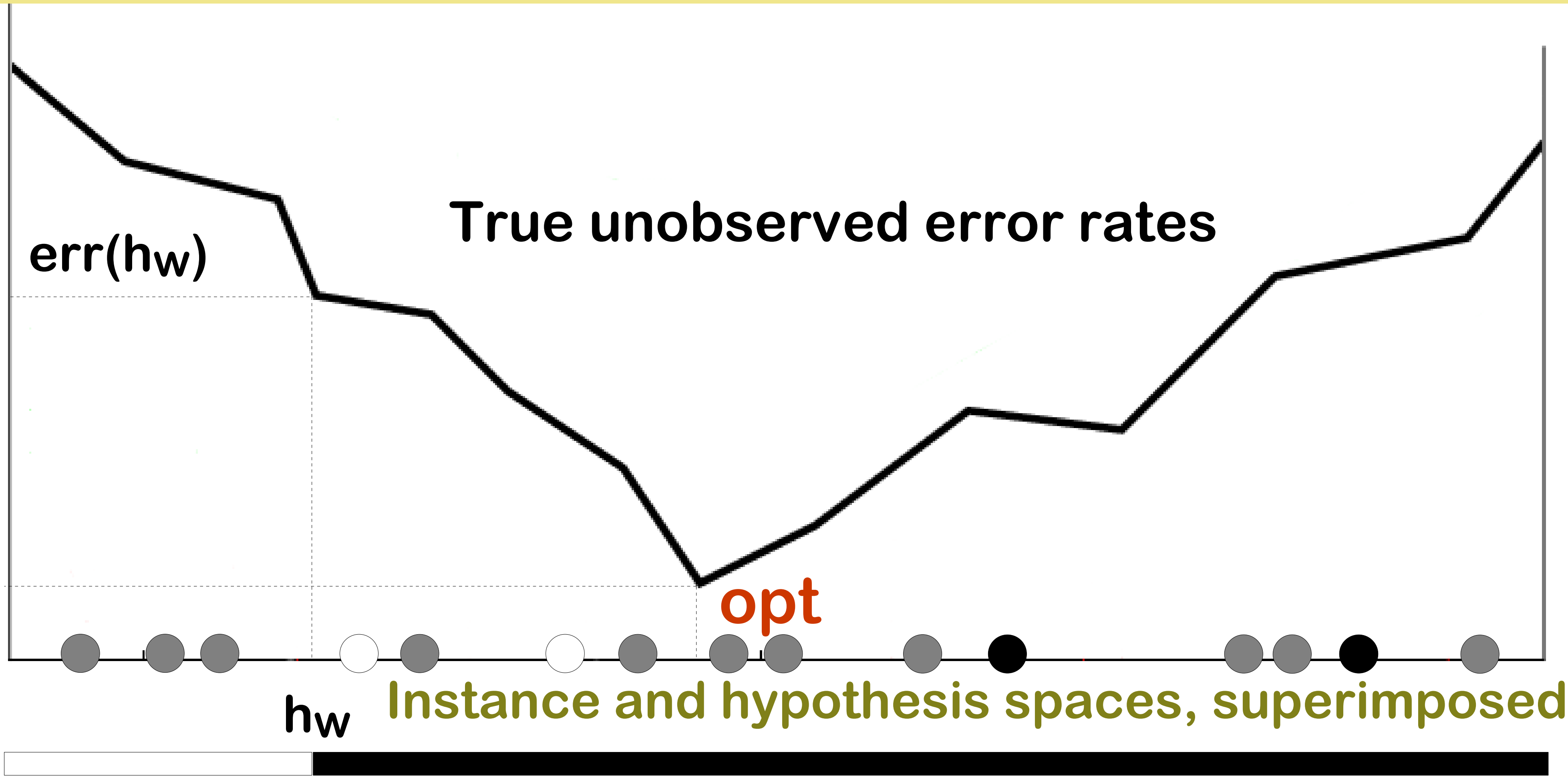
A: Yes, sometimes.

Algorithm A^2 (for Agnostic Active)

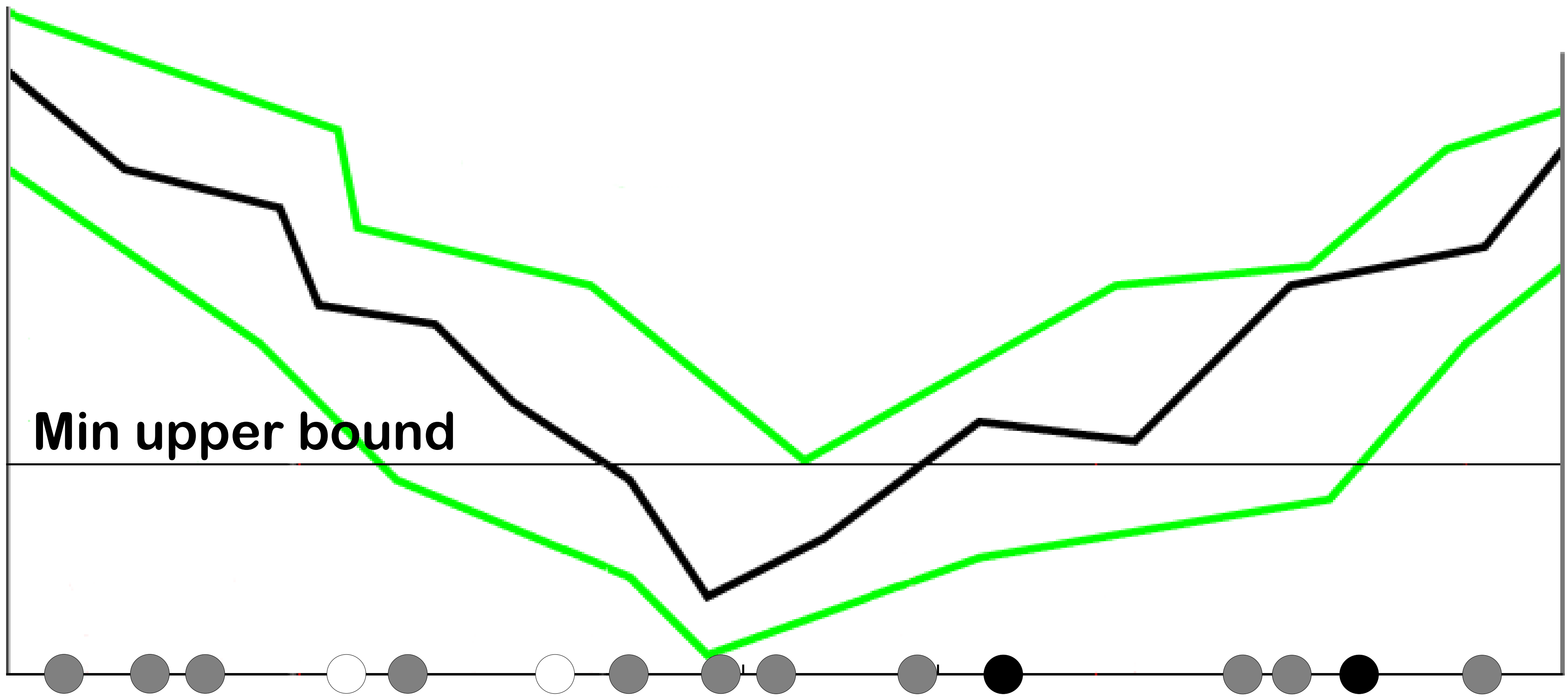
[Balcan, Beygelzimer, Langford'06]

A^2 in Action (H = thresholds on the line)

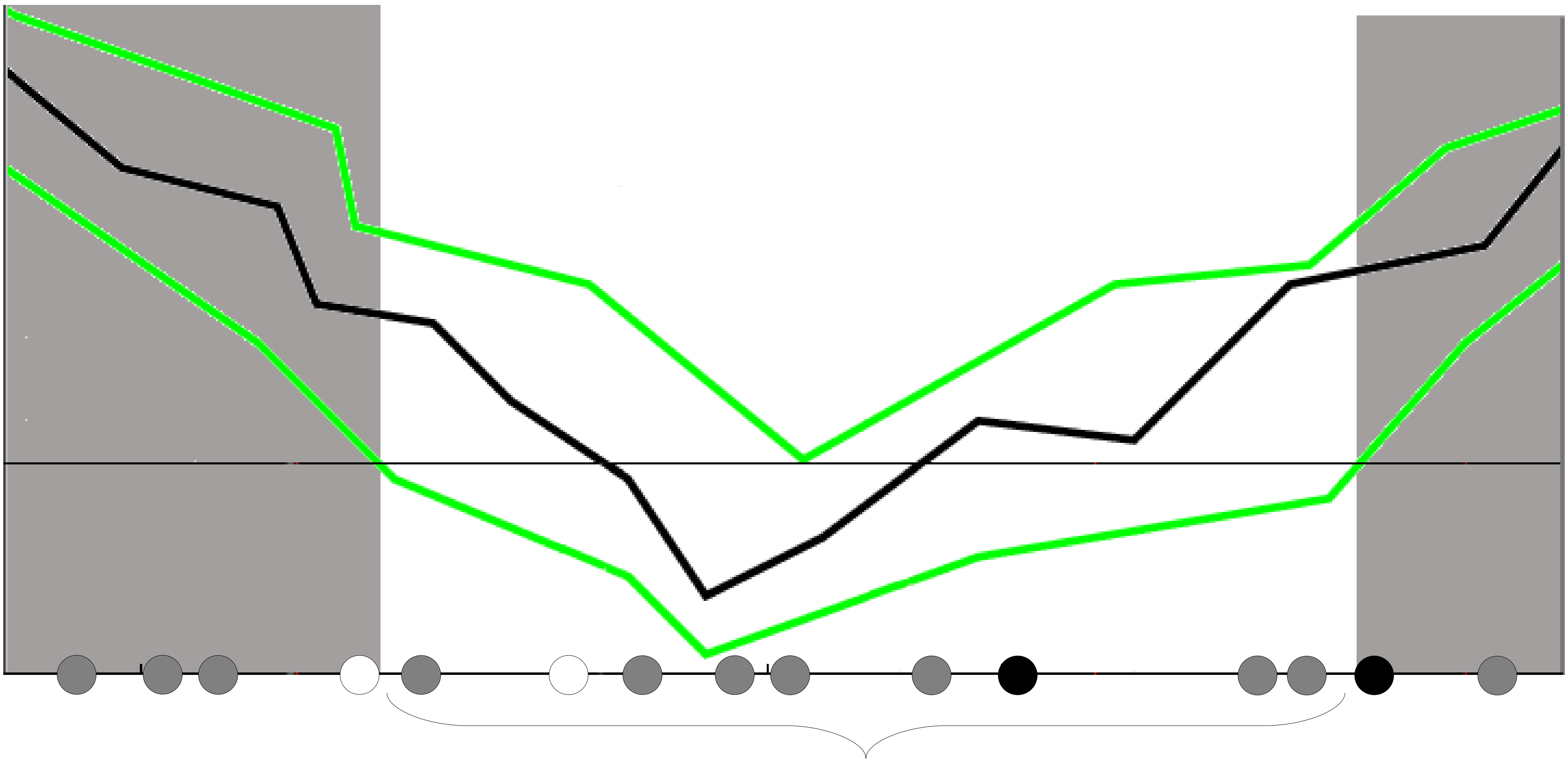
Step 1: Sample and query m examples from D (we want m large enough to cut $\text{Disagree}_D(H)$ in half)



Step 2: Estimate bounds on the error rates of surviving hypotheses (initially all of H)



Step 3: Discard those hypotheses whose lower bound on the error is larger than the smallest upper bound. Eliminate examples on which the remaining hypotheses agree

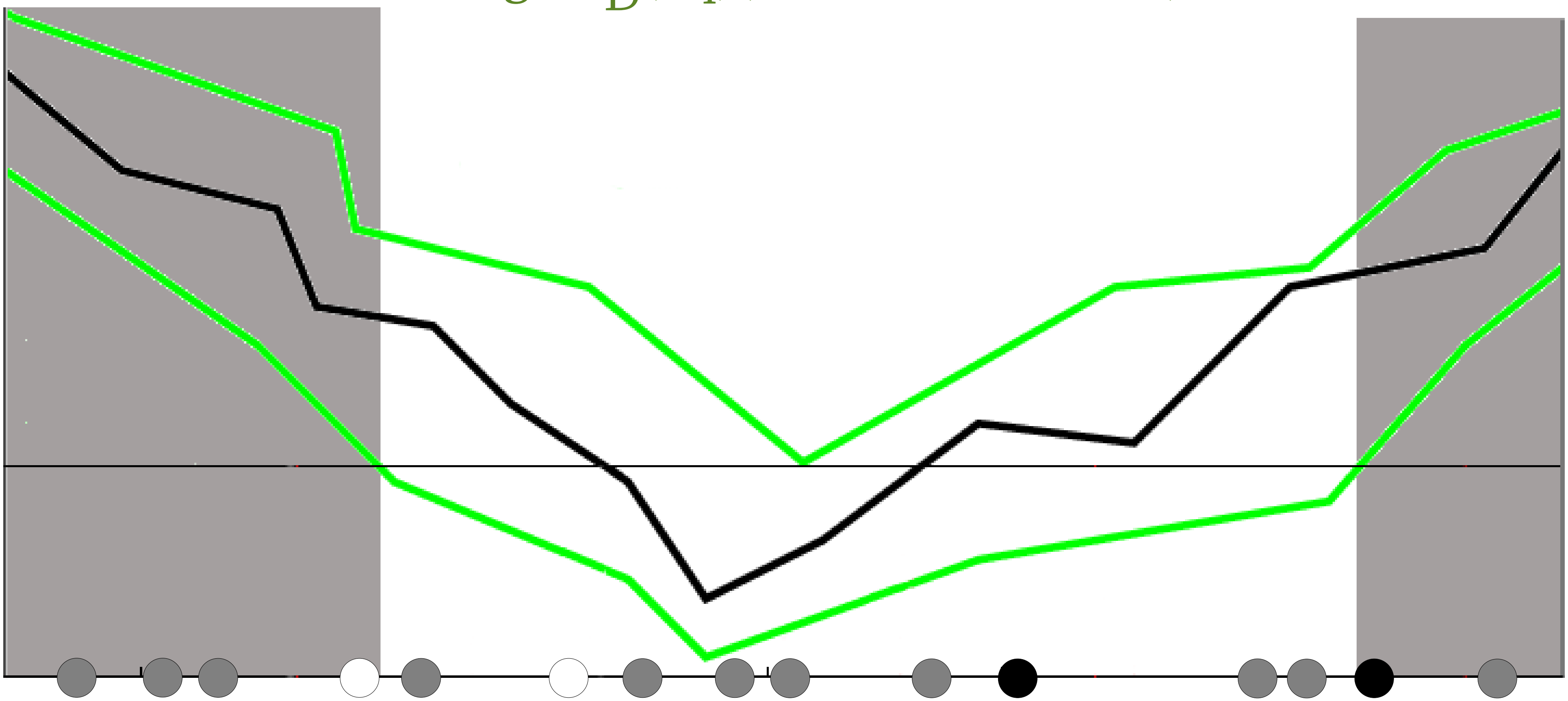


New region of uncertainty Ri

Recurse with the new H_i, D_i and R_i.

All hypotheses h in H_i agree on $X-R_i$, so we can stop once $err_{D_i}(h) Disagree_D(H_i)$ is approximated to precision ϵ, or

$$Disagree_D(H_i)(\min UB - \min LB) \le \epsilon$$



D_i = D restricted to R_i

Theorem (thresholds, low noise): For *any* input distribution, any ϵ and opt $< \epsilon/16$, label complexity is $O(\log 1/\epsilon)$.

Theorem (thresholds, high noise): If opt $> \epsilon$, label complexity is $O(\mathrm{opt}^2/\epsilon^2)$.

Theorem (Linear separators in R^d, low noise): For distributions within multiplicative factor of the uniform, any ϵ and opt $< \frac{\epsilon}{16\sqrt{d}}$, label complexity is $O(d^2 \log 1/\epsilon)$.

Linear separators in R^d

Uniform distribution:

Concentrated near
the equator