Active Learning and Selective Sensing

Closing the loop between data analysis and acquisition

Rui Castro

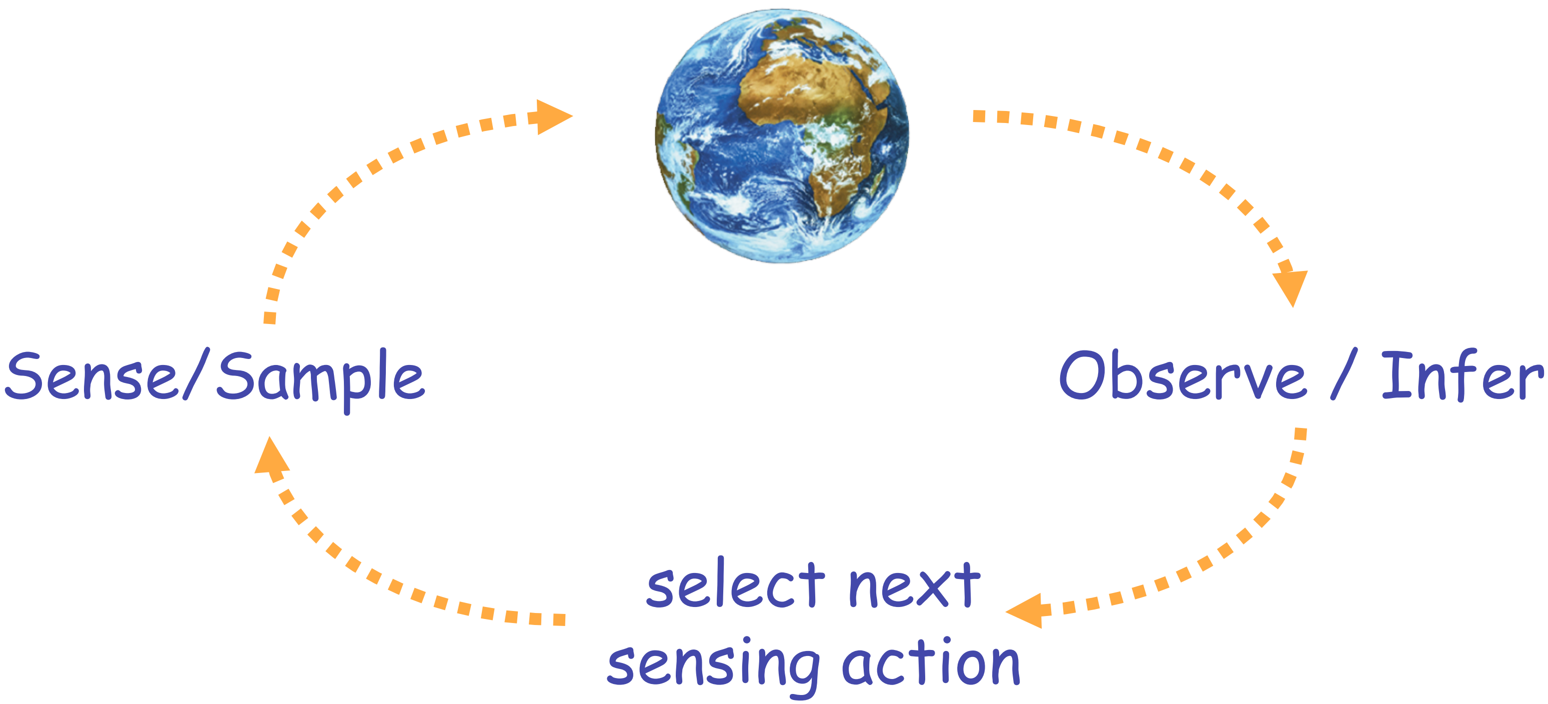
Department of Electrical Engineering

COLUMBIA UNIVERSITY

http://www.ee.columbia.edu/~rmcastro

Motivation

How do we learn about the World?



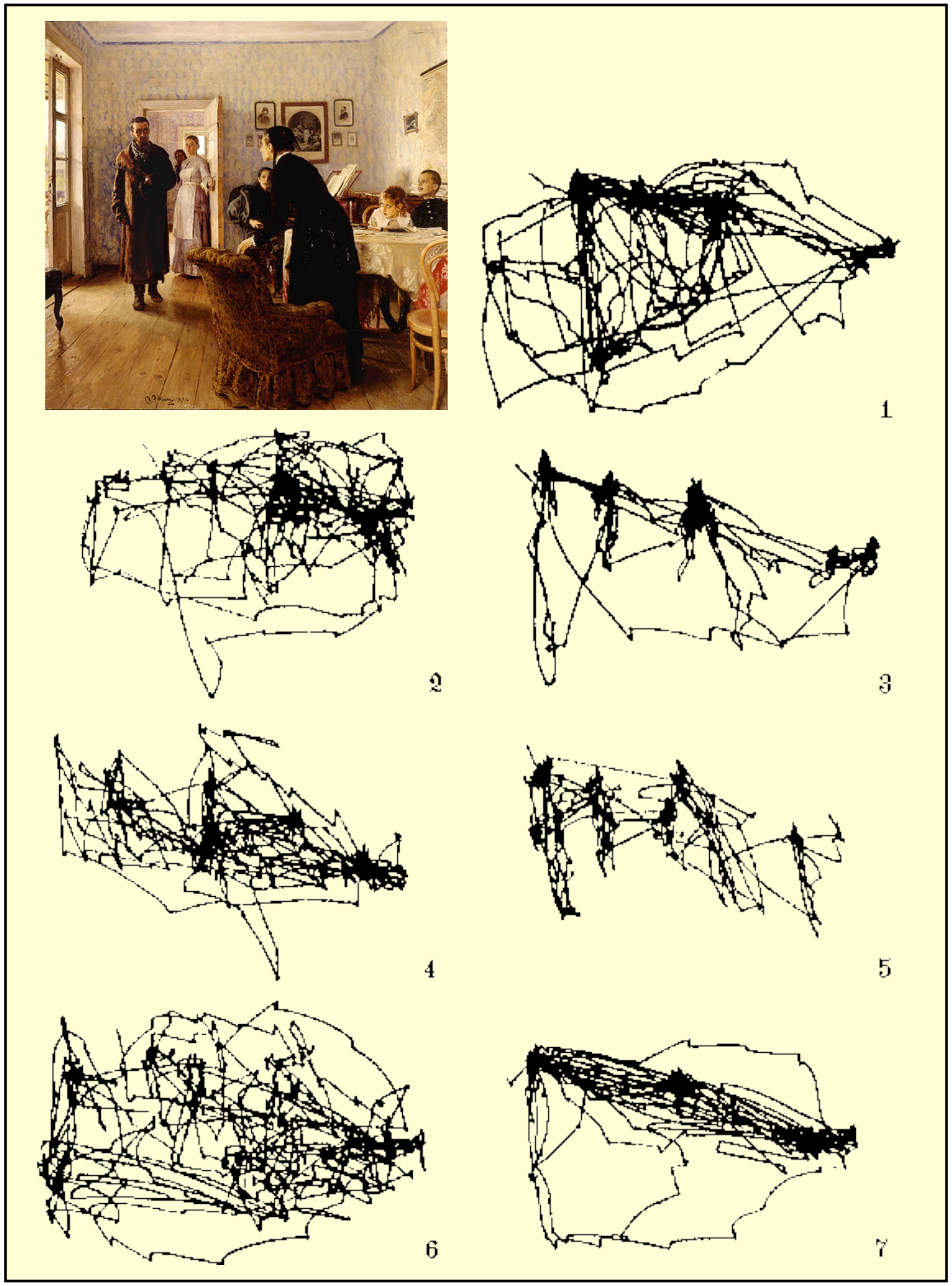
The learning process is in essence **sequential** and **adaptive/active**…

More Motivation – Visual Perception

Ilya Repin. *Unexpected Return* (1884)

Use previously collected data to guide the sampling process

(Eye tracking from Yarbus, 1967)



Seven records of eye movements by the same subject. Each record lasted 3 minutes. 1) Free examination. Before subsequent recordings, the subject was asked to: 2) estimate the material circumstances of the family; 3) give the ages of the people; 4) surmise what the family had been doing before the arrival of the "unexpected visitor;" 5) remember the clothes worn by the people; 6) remember the position of the people and objects in the room; 7) estimate how long the "unexpected visitor" had been away from the family (from Yarbus 1967).

How do we learn? - “Twenty Questions”

“Does the person have blue eyes ?”

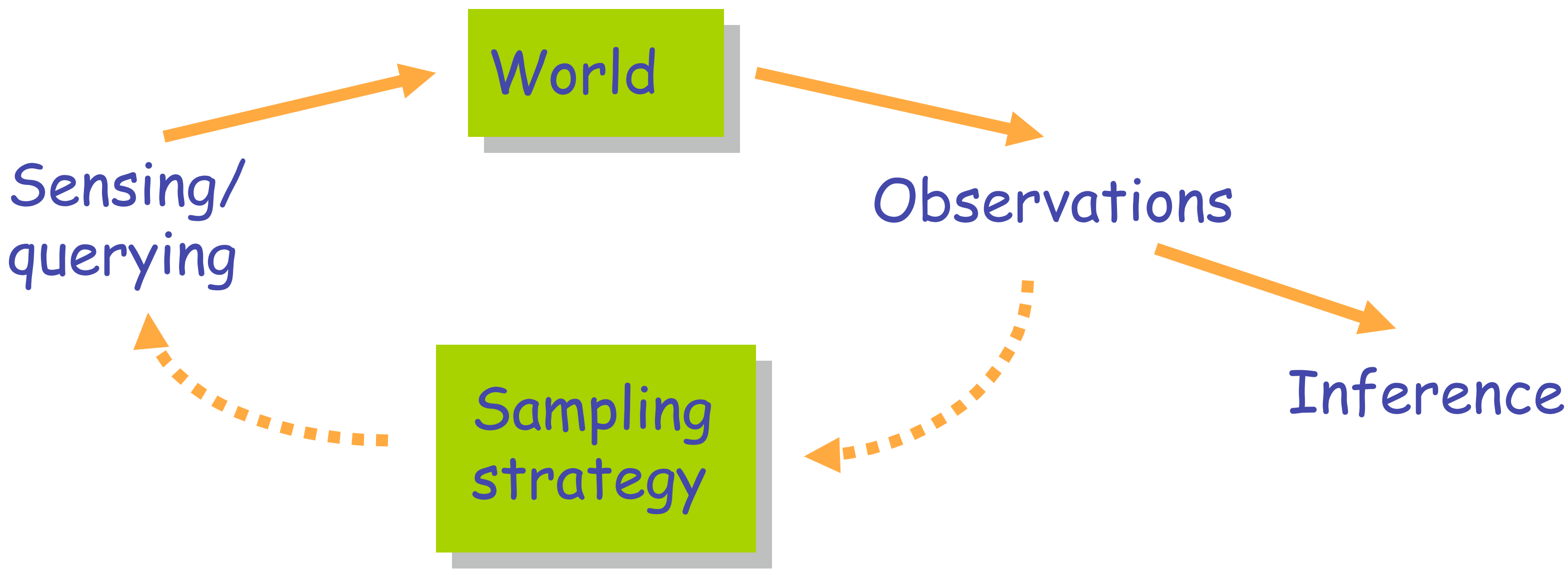
“Is the person wearing a hat ?”

“Active Learning” works very well in simple conditions

How about if the answers are not entirely reliable?

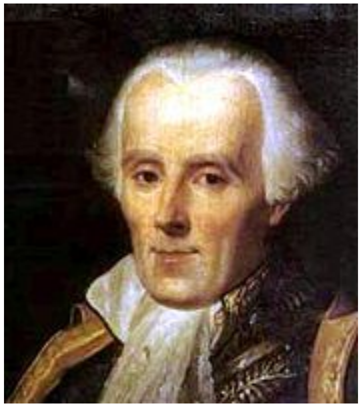
Learning to Learn

Sequential Sensing and Learning: learning using data collection procedures that use information gleaned from previous observations to guide the sensing process.

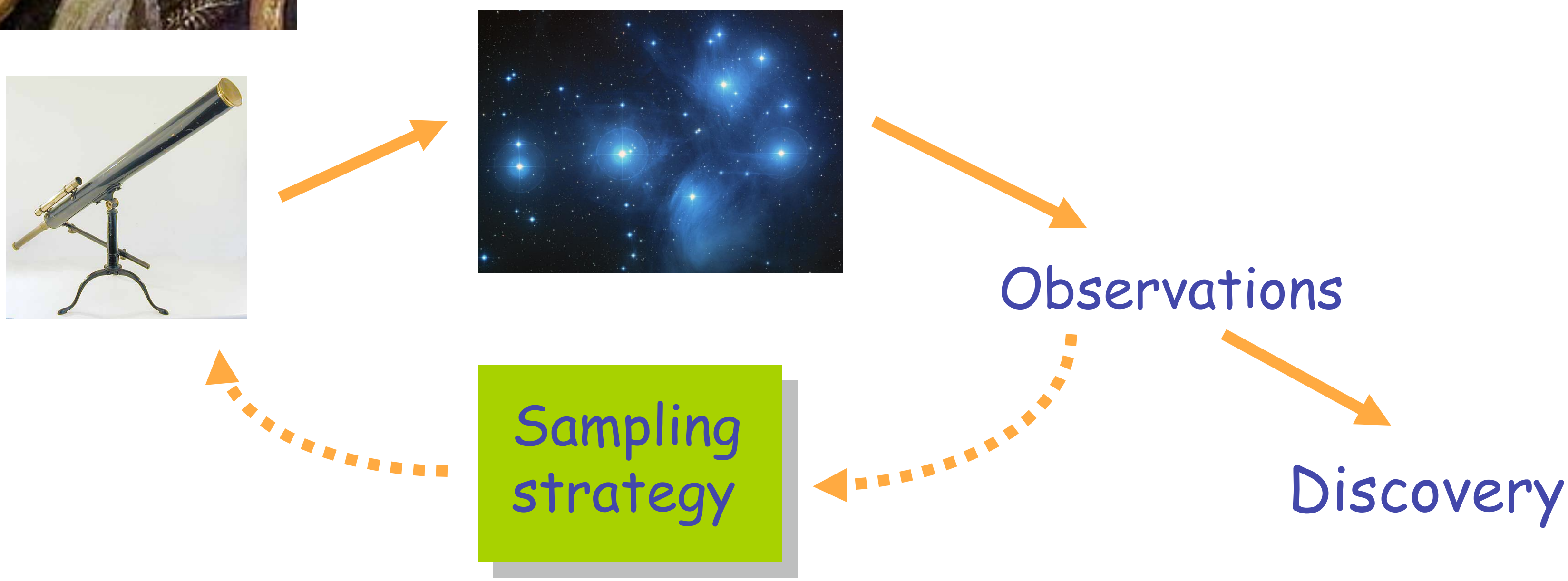


- How can we take advantage of the feedback?
- How much can be gained?
- Devise practical ways of using this feedback?

Laplace's Active Learning



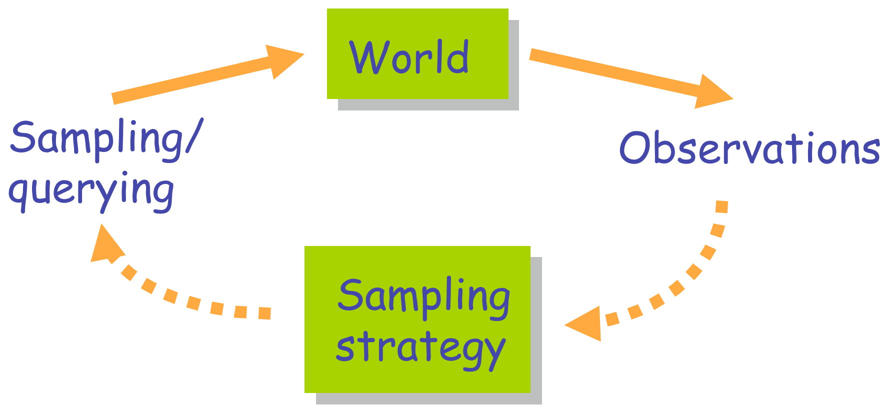
Decided to make new astronomical measurements when "the discrepancy between prediction and observation [was] large enough to give a high probability that there is something new to be found." Jaynes '86



Bayesian approach: select new samples/experiments that are predicted to be maximally informative in discriminating models; "sample where the uncertainty is greatest", Fedorov '72, Mackay '92

Challenges

With **feedback** comes great responsibility!!!



Strong dependencies among observations!!!

If an active learning algorithm is "too aggressive" it might start focusing on the wrong questions...

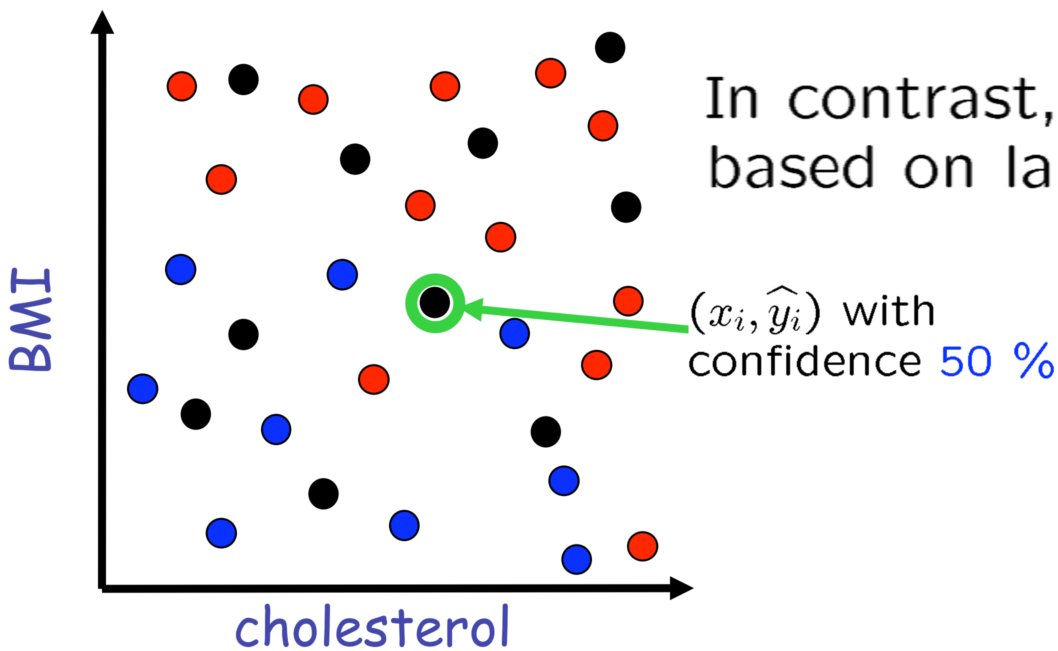
Curiosity can kill the cat!!!

Challenges - Classification

Examples come in pairs, a feature and a label, denoted (x, y).

Select unlabeled examples $(x, ?)$ for labeling if the predicted label $\widehat{y}$ is highly uncertain. These examples may be especially informative.

Heart disease?



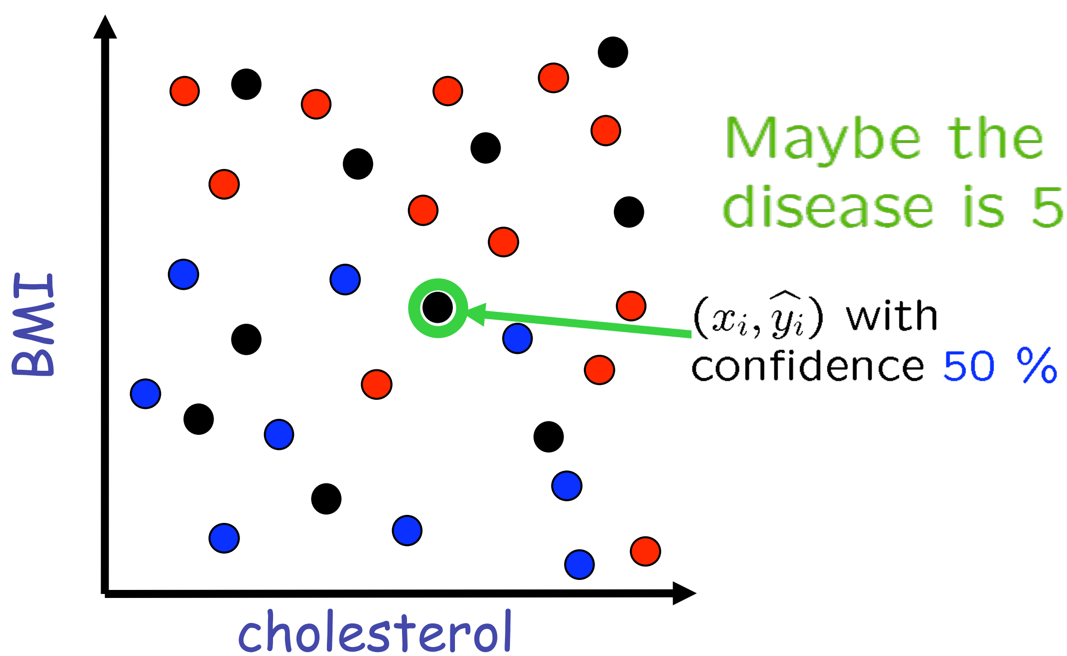
In contrast, passive learning is based on labeling random examples.

Does Active Learning **Always** Help?

Two problems:

1. active learning is greedy and usually myopic, and therefore can converge to a suboptimal hypothesis

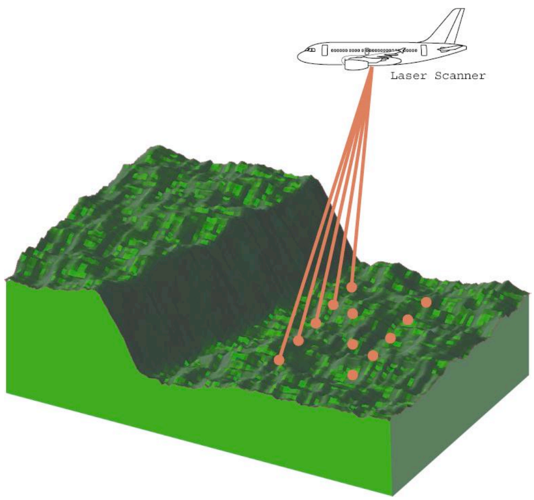
2. uncertainty sampling is 'noise-seeking', and thus may dwell unnecessarily long on highly noisy cases



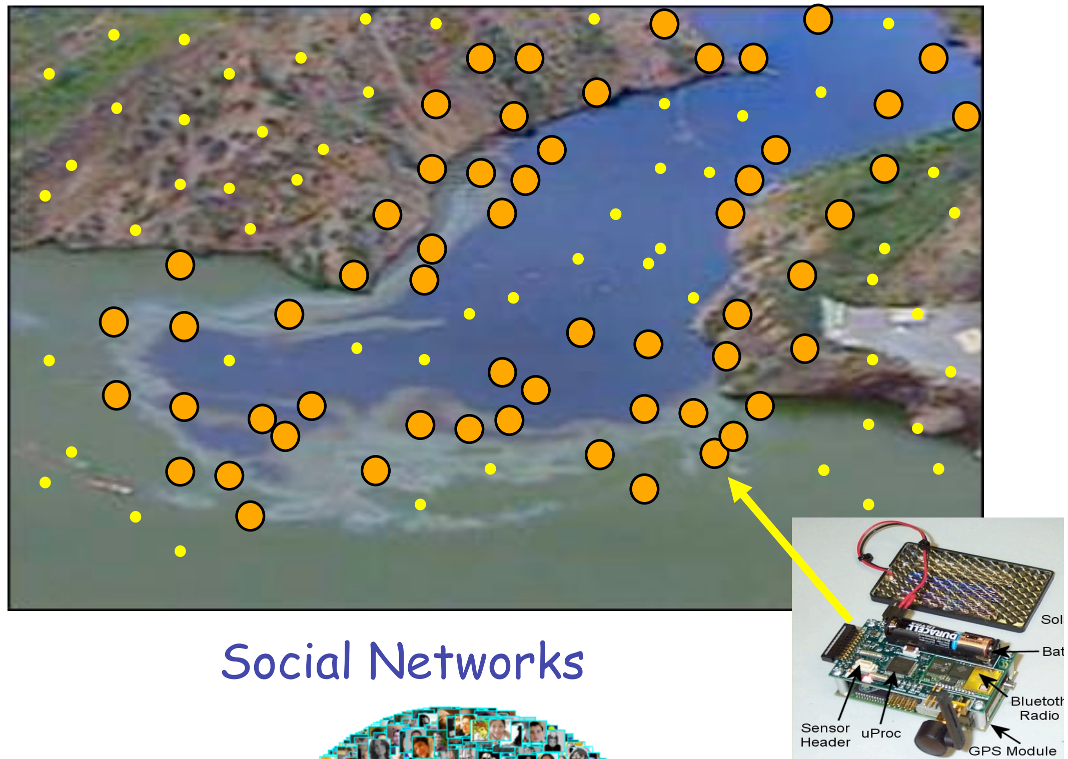
Maybe the chance of heart disease is 50/50 in this region.

Why Do Active Learning?

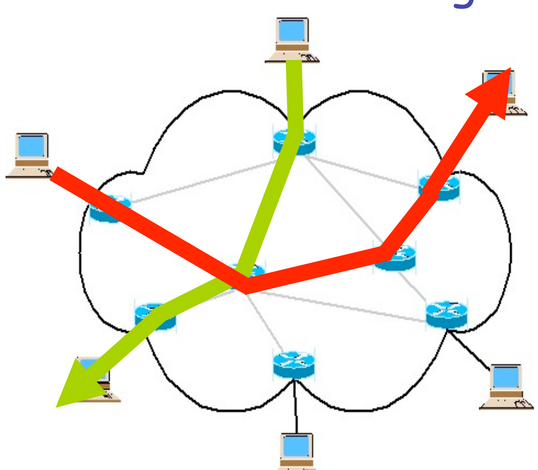
remote sensing



wireless sensor networks



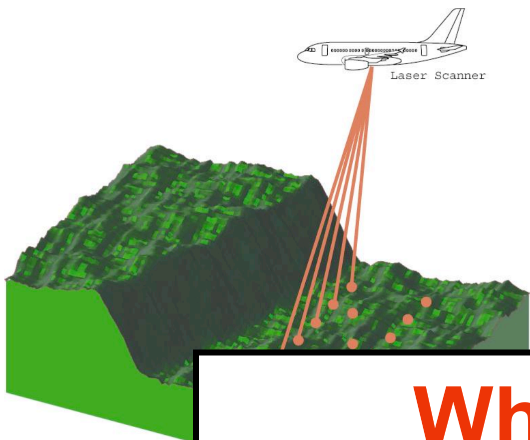
Internet Monitoring



Social Networks

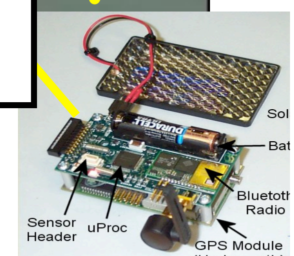
Why Do Active Learning?

remote sensing

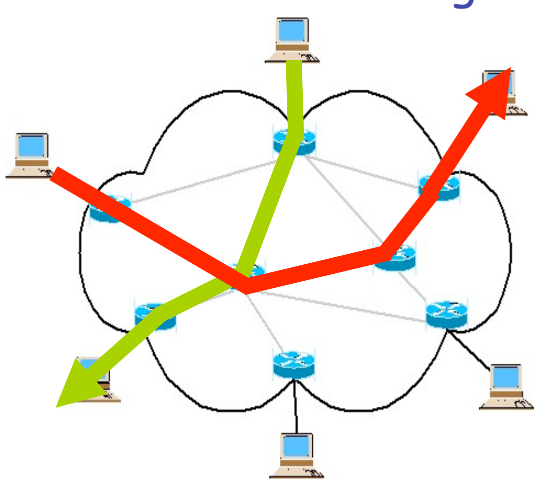


wireless sensor networks

Where, When and **How**
to collect information?



Internet Monitoring



Social Networks

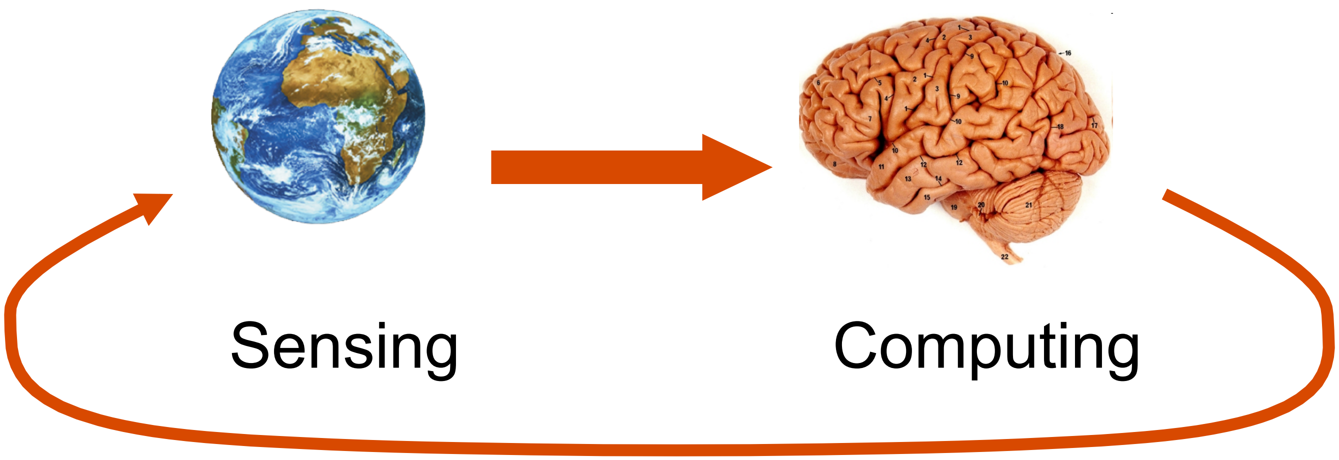
Why do **AL**? - Human Learning

The Theory of the Organism-Environment System: III. Role of Efferent Influences on Receptors in the Formation of Knowledge*

TIMO JARVILEHTO
Department of Behavioral Sciences, University of Oulu, Finland

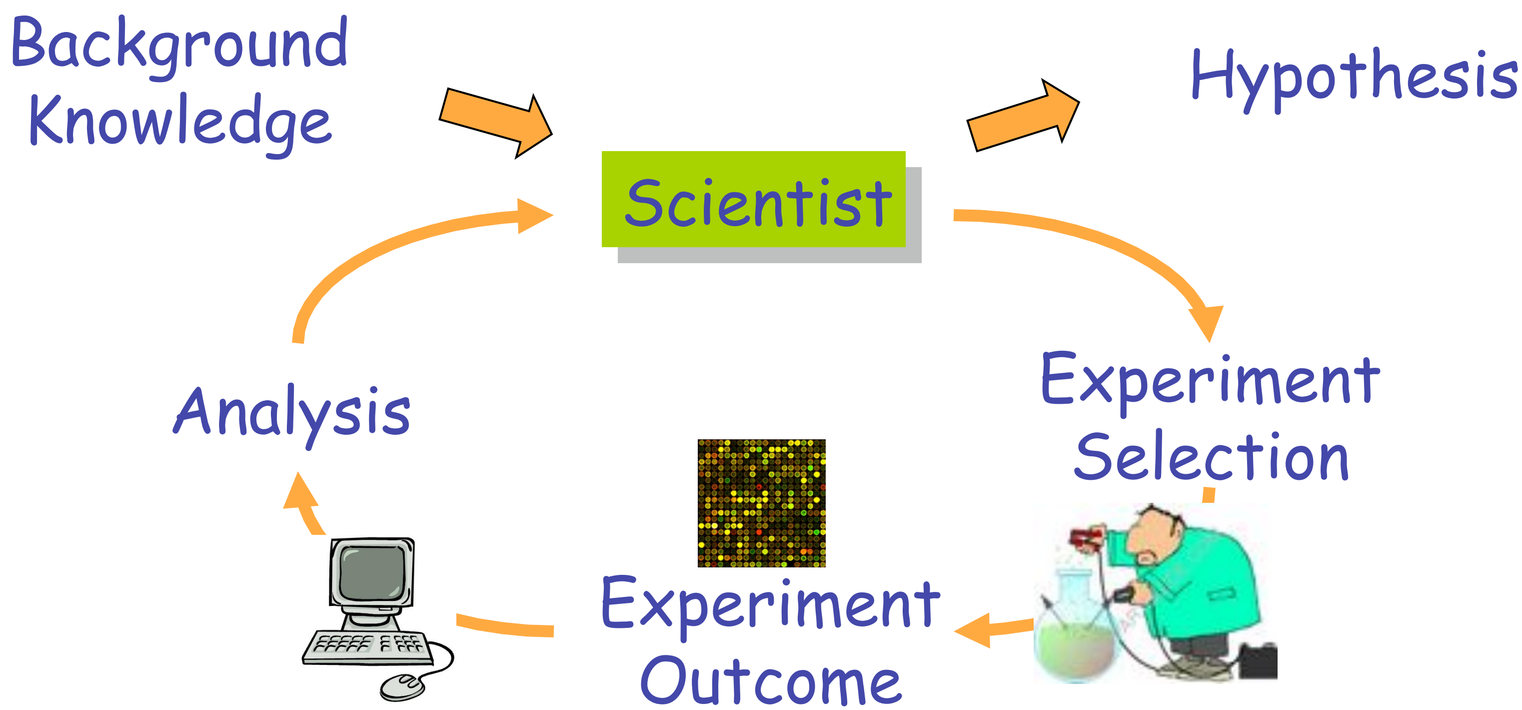
Abstract—The present article is an attempt to give—in the frame of the theory of the organism-environment system (Jarvilehto, 1998a)—a new interpretation to the role of efferent influences on receptor activity and to the functions of senses in the formation of knowledge. It is argued, on the basis of experimental evidence and theoretical considerations, that the senses are not transmitters of environmental information, but create a direct connection between the organism and the environment, which makes the development of a dynamic living system, the organism-environment system, possible. In this connection process, the efferent influences on receptor activity are of particular significance because, with their help, the receptors may be adjusted in relation to the parts of the environment that are most important in achieving behavioral results. Perception is the process of joining of new parts of the environment to the organism-environment system; thus, the formation of knowledge by perception is based on reorganization (widening and differentiation) of the organism-environment system, and not on transmission of information from the environment. With the help of the efferent influences on receptors, each organism creates its own peculiar world that is simultaneously subjective and objective. The present considerations have far-reaching influences as well on experimental work in neurophysiology and psychology of perception as on philosophical considerations of knowledge formation.

Why do **AL**? - Human Learning



Abstract—The present article is an attempt to give—in the frame of the theory of the organism-environment system (Jarvilehto, 1998a)—a new interpretation to the role of efferent influences on receptor activity and to the functions of senses in the formation of knowledge. It is argued, on the basis of experimental evidence and theoretical considerations, that the senses are not transmitters of environmental information, but create a direct connection between the organism and the environment, which makes the development of a dynamic living system, the organism-environment system, possible. In this connection process, the efferent influences on receptor activity are of particular significance because, with their help, the receptors may be adjusted in relation to the parts of the environment that are most important in achieving behavioral results. Perception is the process of joining of new parts of the environment to the organism-environment system; thus, the formation of knowledge by perception is based on reorganization (widening and differentiation) of the organism-environment system, and not on transmission of information from the environment. With the help of the efferent influences on receptors, each organism creates its own peculiar world that is simultaneously subjective and objective. The present considerations have far-reaching influences as well on experimental work in neurophysiology and psychology of perception as on philosophical considerations of knowledge formation.

Why do **AL**? - Automating Science



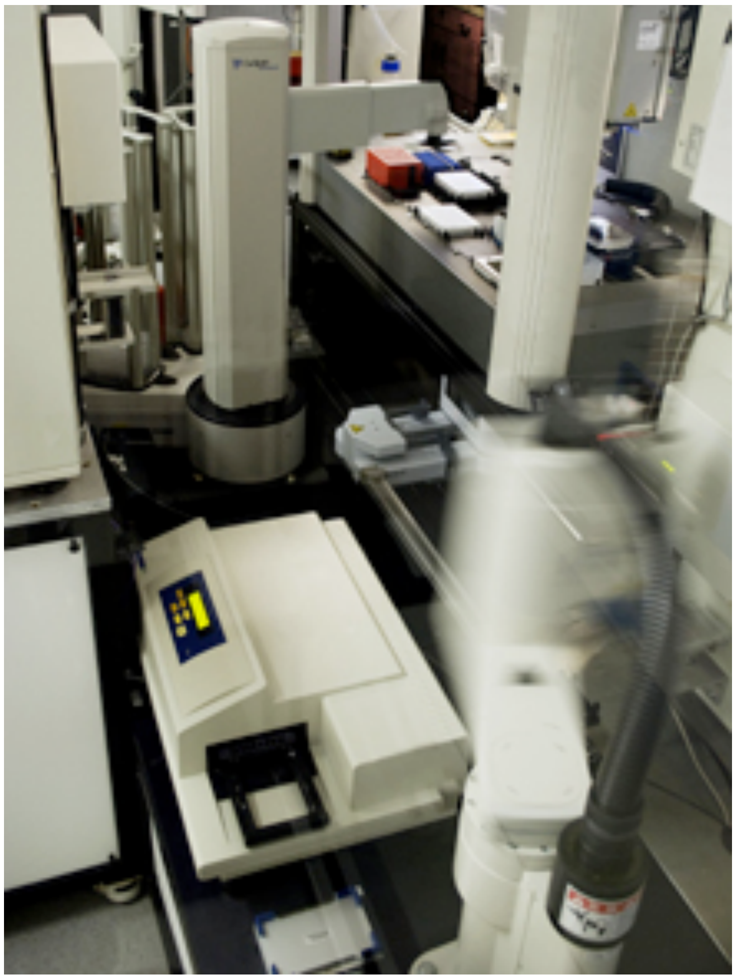
- Huge burden to the human in the loop
- Humans are unable to grasp the high-dimensional complexity of processes of interest

There is a need for "autonomous experimentation"

"Towards 2020 Science" – 40 eminent scientists' visions of the future of science

Robot Scientist

www.aber.ac.uk/compsci/Research/bio/robotsci/



Wired Magazine, April 2009:

For the first time, a robotic system has made a novel scientific discovery with virtually no human intellectual input.

Scientists designed "Adam" to carry out the entire scientific process on its own: formulating hypotheses, designing and running experiments, analyzing data, and deciding which experiments to run next. "It’s a major advance," says David Waltz of the Center for Computational Learning Systems at Columbia University. "Science is being done here in a way that incorporates artificial intelligence. It’s automating a part of the scientific process that hasn’t been automated in the past."

Adam is the first automated system to complete the cycle from hypothesis, to experiment, to reformulated hypothesis without human intervention.

Outline

Binary Classification and the fundamental limits of active learning

Algorithmic considerations, and active learning in practice

Probabilistic Framework for Classification

$\mathcal{X}$ - The **feature** space (e.g. $\mathcal{X} = [0,1]^d$)

$\mathcal{Y}$ - The **label** space (e.g. $\mathcal{Y} = \{0,1\}$)

$$(\underbrace{X}_{\text{features}}, \underbrace{Y}_{\text{label}}) \in \mathcal{X} \times Y \quad \sim \quad P_{XY}$$ (generally unknown)

Goal: Construct a classification rule $f : \mathcal{X} \to \mathcal{Y}$ minimizing the **risk**

$$R(f) = \underbrace{\Pr(f(X) \neq Y)}_{\text{probability of error}}$$

In words: given a feature vector X we want to predict the label Y as well as possible...

Bayes Classifier

What is the "best" classification rule?

$$f^* \;=\; \arg \min_{f \text{ measurable}} \Pr(f(X) \neq Y)$$

Since we are considering binary labels any reasonable classification rule has the form $f(x) = 1_G(x), \;\; G \subseteq \mathcal{X}$

$$G^* \;=\; \arg \min_{G \text{ measurable}} \Pr(1_G(X) \neq Y)$$

The **Bayes classifier** is defined by the level set

$$G^* = \{x : \eta(x) \geq 1/2\}$$

where $\eta(x) := P(Y = 1|X = x)$.

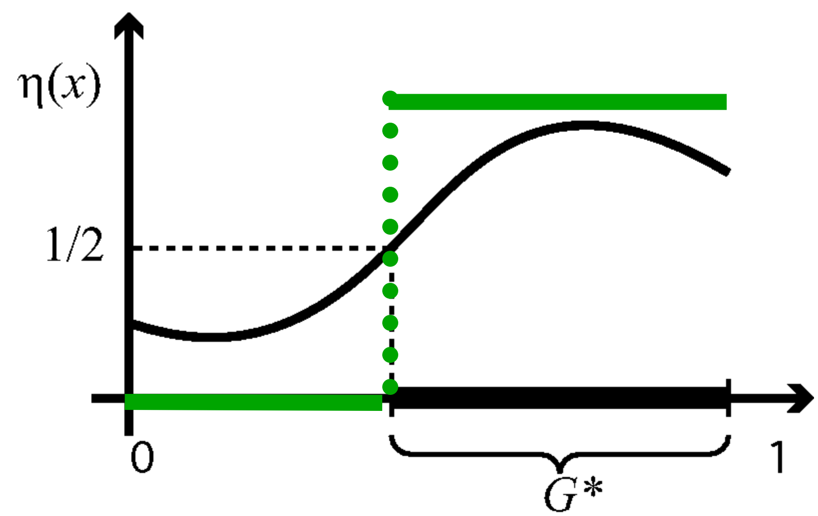
Bayes Classifier

The Bayes classifier says 1 if, given a feature X, it is more likely that the corresponding label is 1

$$G^* = \{x : \underbrace{P(Y=1|X=x)}_{\eta(x)} \geq 1/2\}$$

G^* is the $\frac{1}{2}$ level set of $\eta(\cdot)$

requires knowledge of P_{XY}



Classification is just a level-set estimation problem

Learning from Examples

In most problems $P_{Y|X}$ is unknown. We have to rely on data

$$\{(X_i, Y_i)\}_{i=1}^n \qquad\qquad Y_i | X_i \;\sim\; P_{Y|X}$$

Goal: Construct a classifier $\widehat{G}_n \equiv \widehat{G}(X_1, Y_1, \ldots, X_n, Y_n)$ minimizing the **excess risk**

$$\begin{aligned} \mathcal{E}(\widehat{G}_n) &= R(\widehat{G}_n) - R(G^*) \\ &= \Pr(\mathbf{1}_{\widehat{G}_n}(X) \neq Y) - \Pr(\mathbf{1}_{G^*}(X) \neq Y) \end{aligned}$$

We want to find a classifier "close" to G^*!

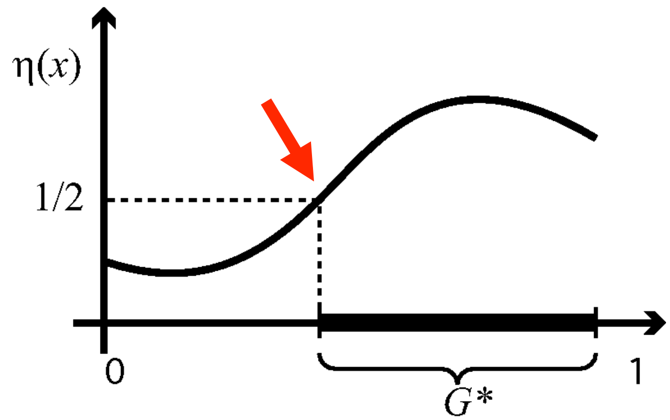
Excess Risk

$$\eta(x) := P(Y = 1 | X = x).$$

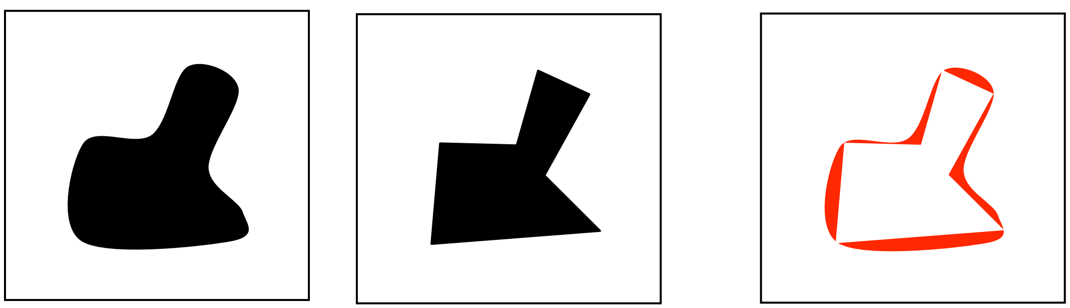
$$\begin{aligned} \mathcal{E}(G) &= R(G) - R(G^*) \\ &= P(1_G(x) \neq Y) - P(1_{G^*}(x) \neq Y) \\ &= \int_{\underbrace{G \Delta G^*}} \underbrace{|2\eta(x) - 1|} \, \mathrm{d}P_X(x) \end{aligned}$$

How smooth is η near ∂G^*

"noise" characterization



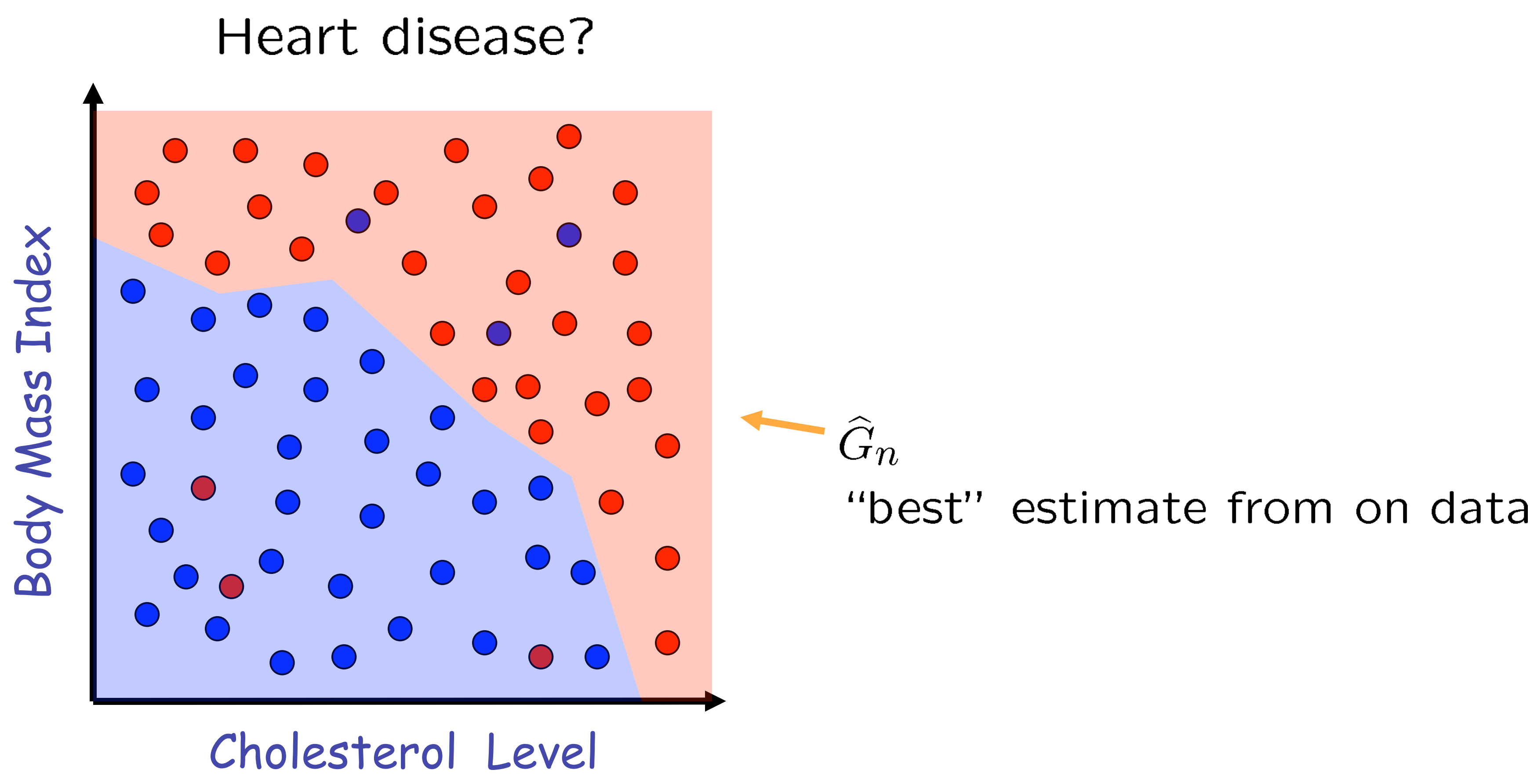
How easy is to approximate G^*



G^* $\qquad$ G $\qquad$ $G \Delta G^* = G \cap \bar{G}^* \bigcup \bar{G} \cap G^*$

Passive Learning $\{(X_i, Y_i)\}_{i=1}^n \overset{\text{i.i.d.}}{\sim} P_{XY}$

Given n randomly selected examples how well can we do?



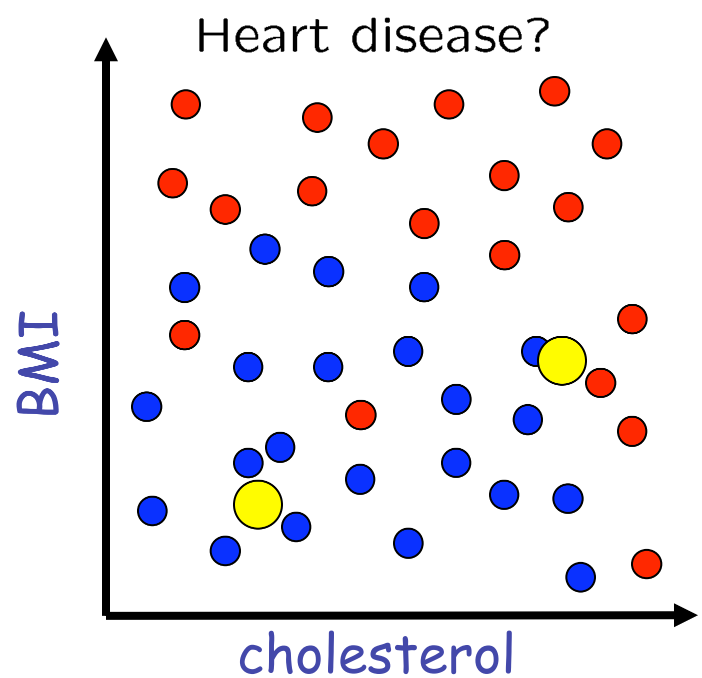
Active Learning

$$\underbrace{\{X_j\}_{i=1}^{\infty}}_{\text{Large pool of } \textbf{unlabeled examples}} \xrightarrow{\text{select}} X \longrightarrow \text{request } Y|X \sim P_{Y|X}$$

many unlabeled examples
(e.g., people, documents)

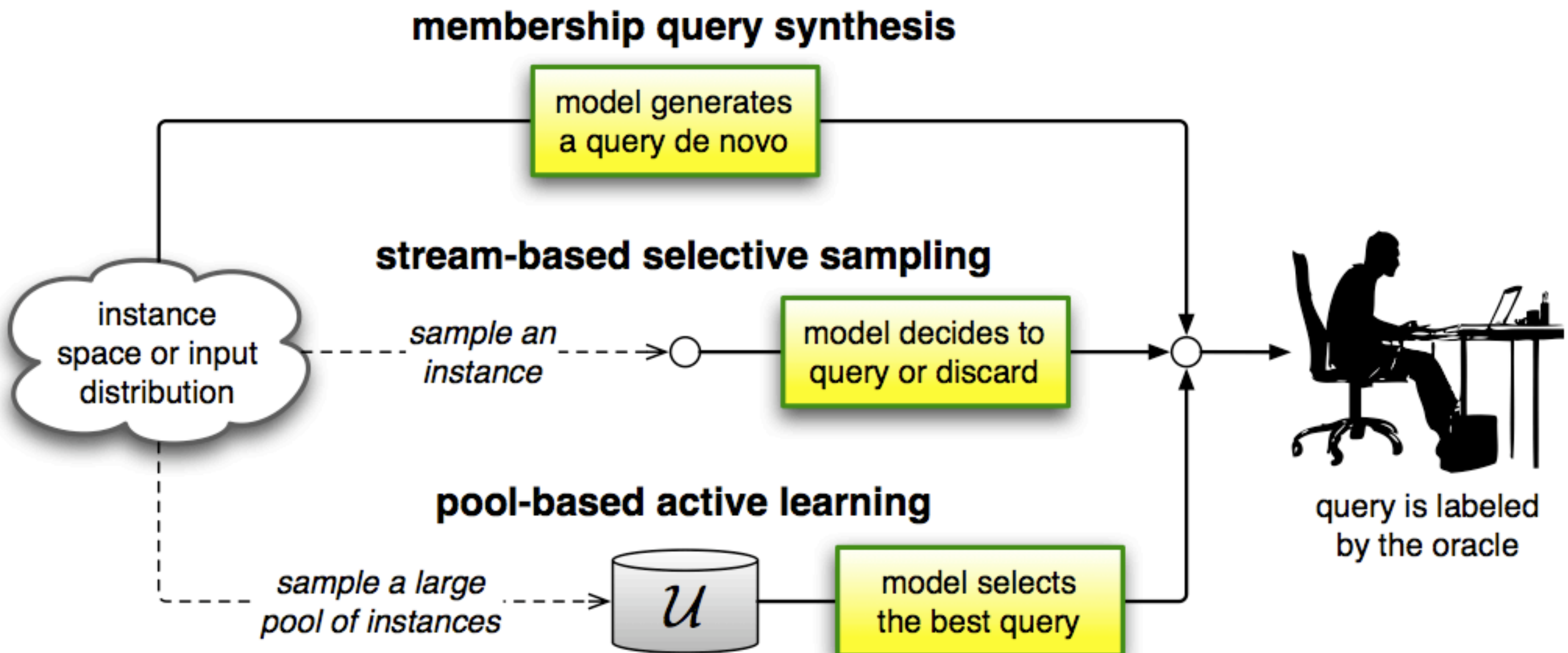
labeling examples is expensive

some examples are more
informative than others



Given *n* **selectively** chosen training examples,
how well can we do?

Three Active Learning Paradigms



Passive vs. Active Sampling

Passive Sampling:

Features $X_i \in [0,1]^d$ are independent of $\{Y_j\}_{j \neq i}$. That is, you can select all the features $\{X_i\}$ prior to collecting the labels $\{Y_i\}$.

Active Sampling:

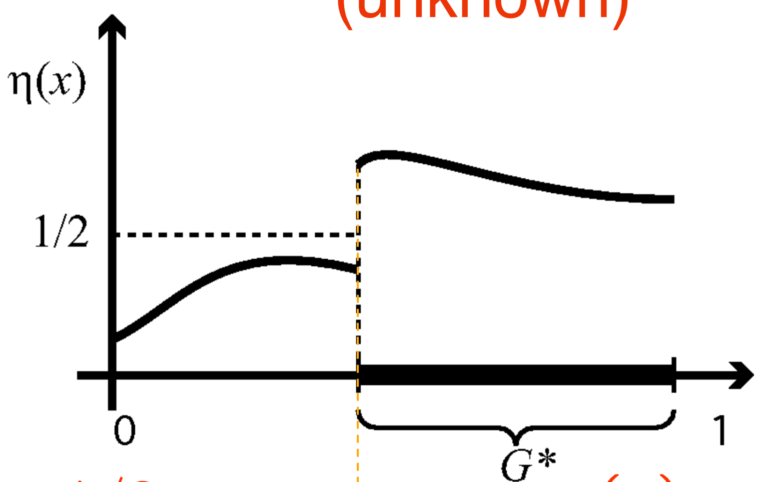
Features X_i are random and depend only on past observations $\{X_j, Y_j\}_{j=1}^{i-1}$. That is, X_i is completely defined by

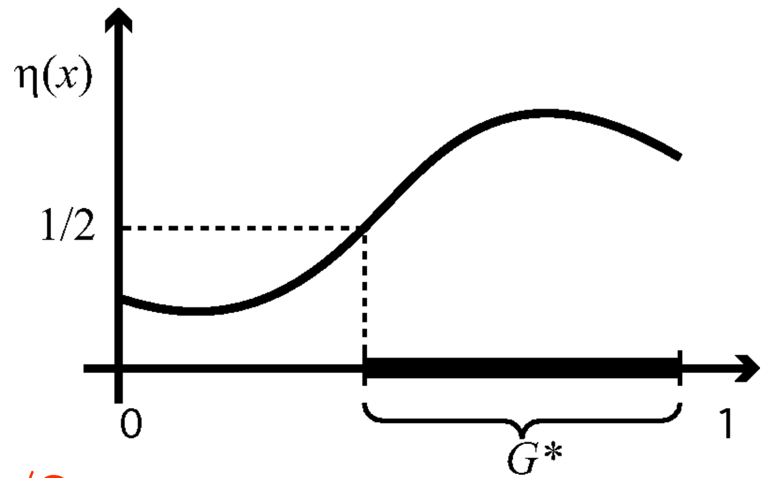
$$X_i | (X_{i-1}, Y_{i-1}), \ldots, (X_1, Y_1)$$

The One Dimensional Threshold Problem

$\mathcal{X} = [0,1]$ and $G^* = [\theta^*, 1]$

(unknown)





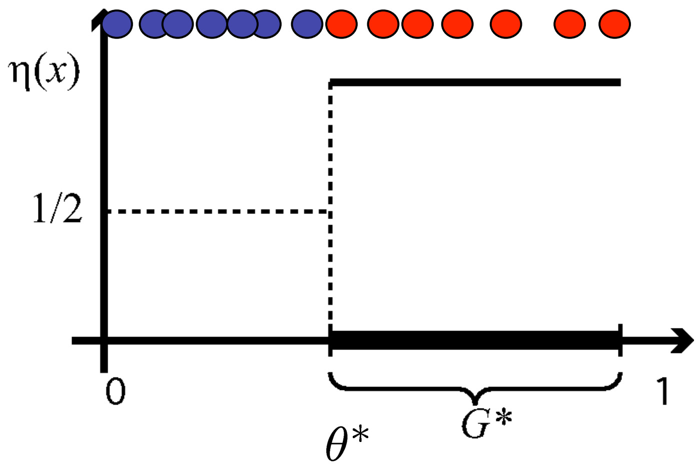
$\eta(x) < 1/2$ ⟵ ⟶ $\eta(x) > 1/2$

Assume also $X \sim \underbrace{\text{Unif}([0,1])}_{\text{This can be made more general (bounded density)}}$

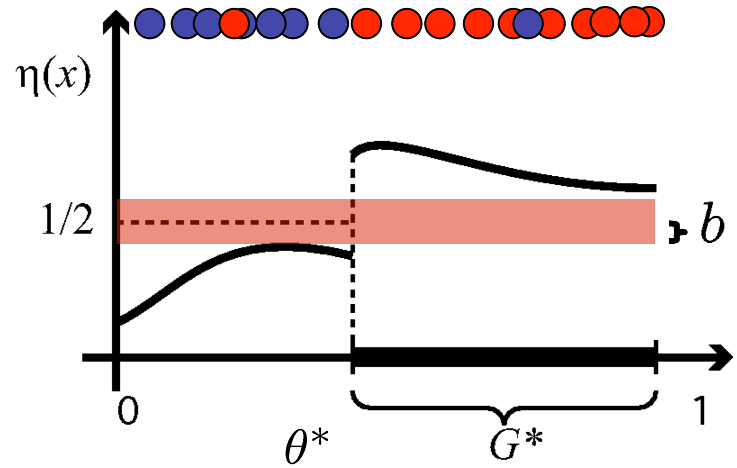
Goal: Minimizing the excess risk boils down to constructing a good estimate $\widehat{\theta}_n$ of θ^*

Various Scenarios

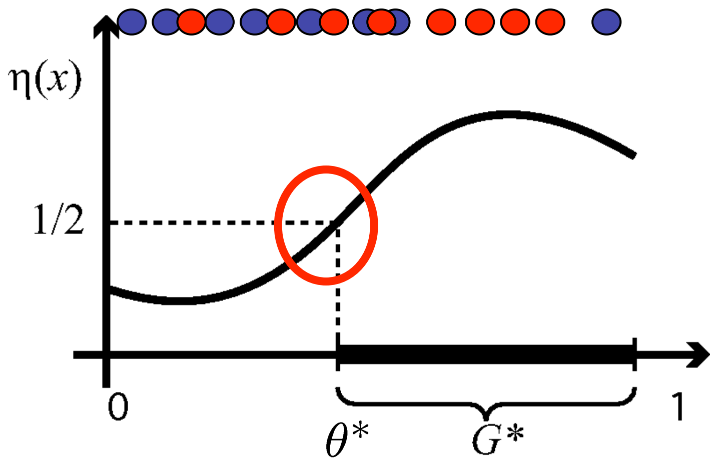
noiseless



bounded noise



unbounded noise

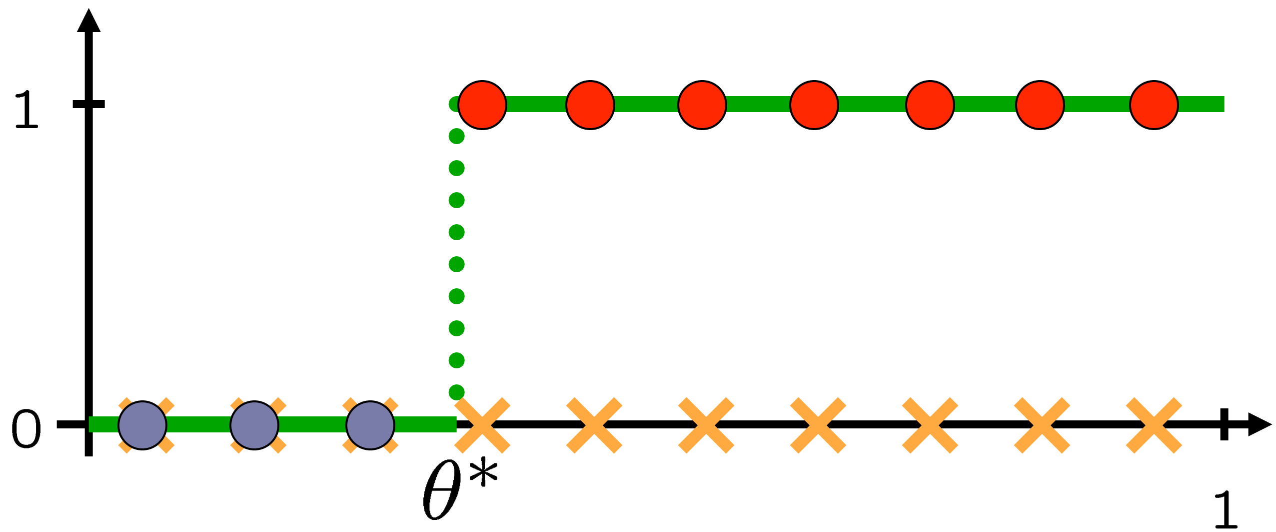


No strong cue about the location of the boundary

How much does active learning help in each case?

Passive Learning

Sample locations must be chosen before any observations are made

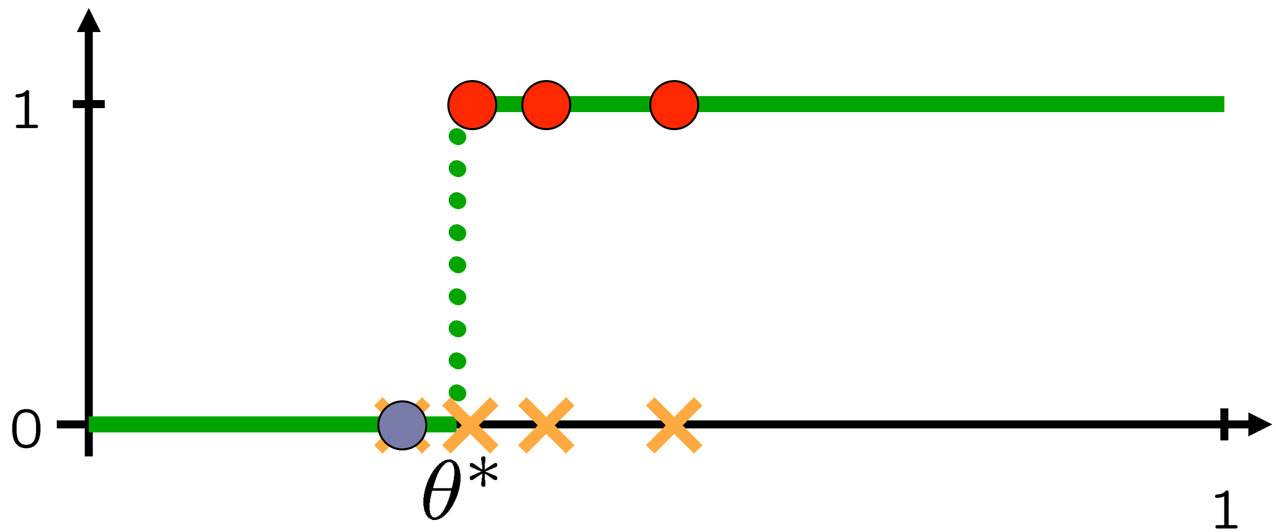


$$|\widehat{\theta}_n - \theta^*| \sim \frac{1}{n}$$

Too many wasted samples. Learning is limited by **sampling resolution**

Active Learning

Sample locations are chosen as a function of previous observations

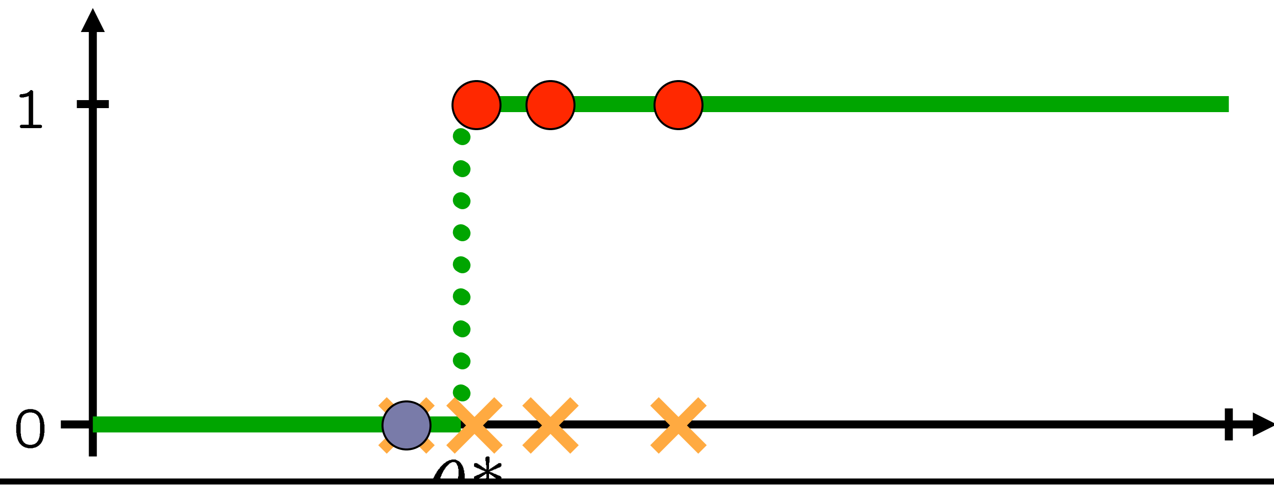


$$|\widehat{\theta}_n - \theta^*| \sim 2^{-n}$$

The error decays **much faster** than in the passive scenario. No wasted samples…

Active Learning

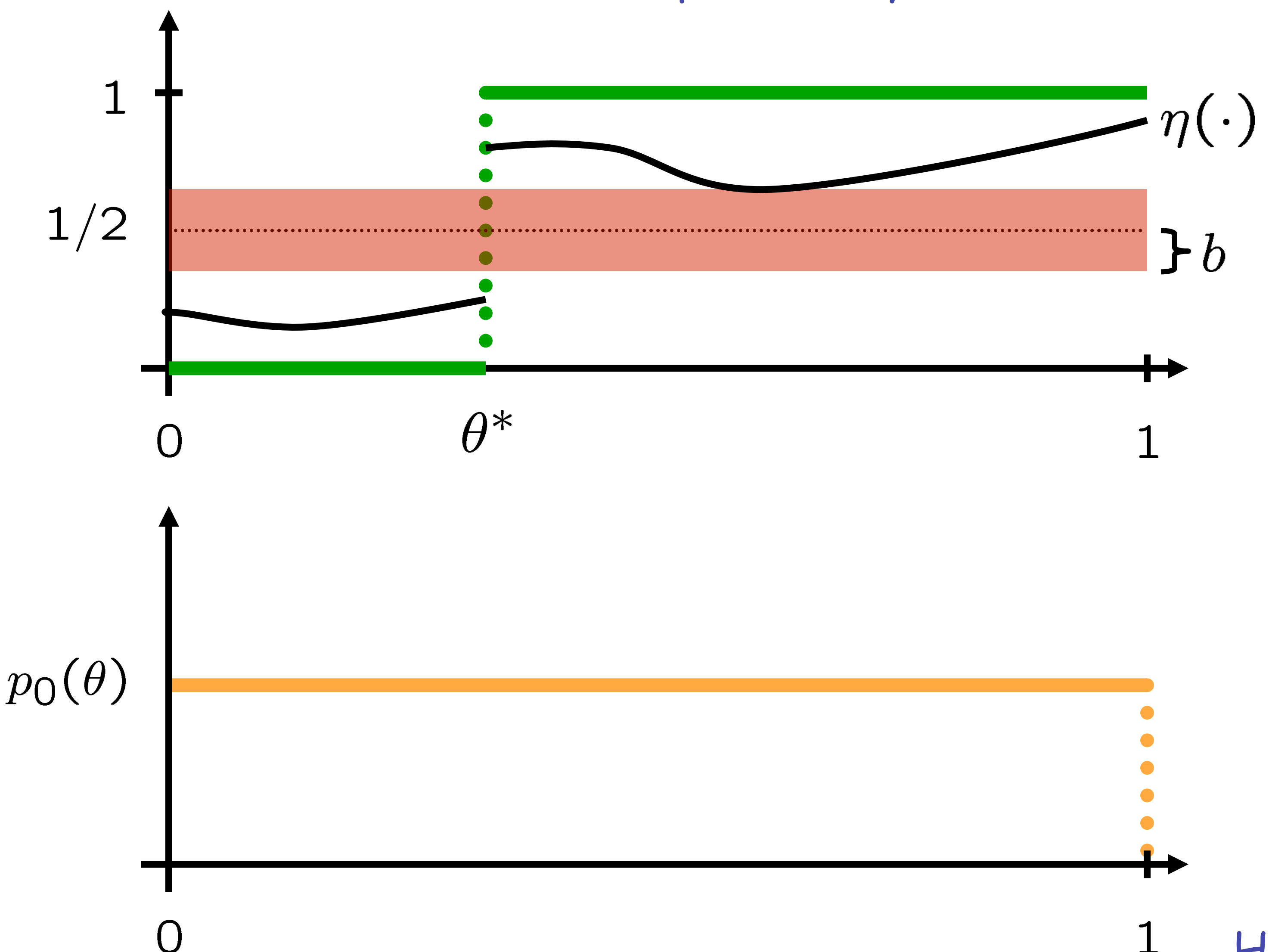
Sample locations are chosen as a function of previous observations



What if there is uncertainty?

Active Learning – Bounded Noise

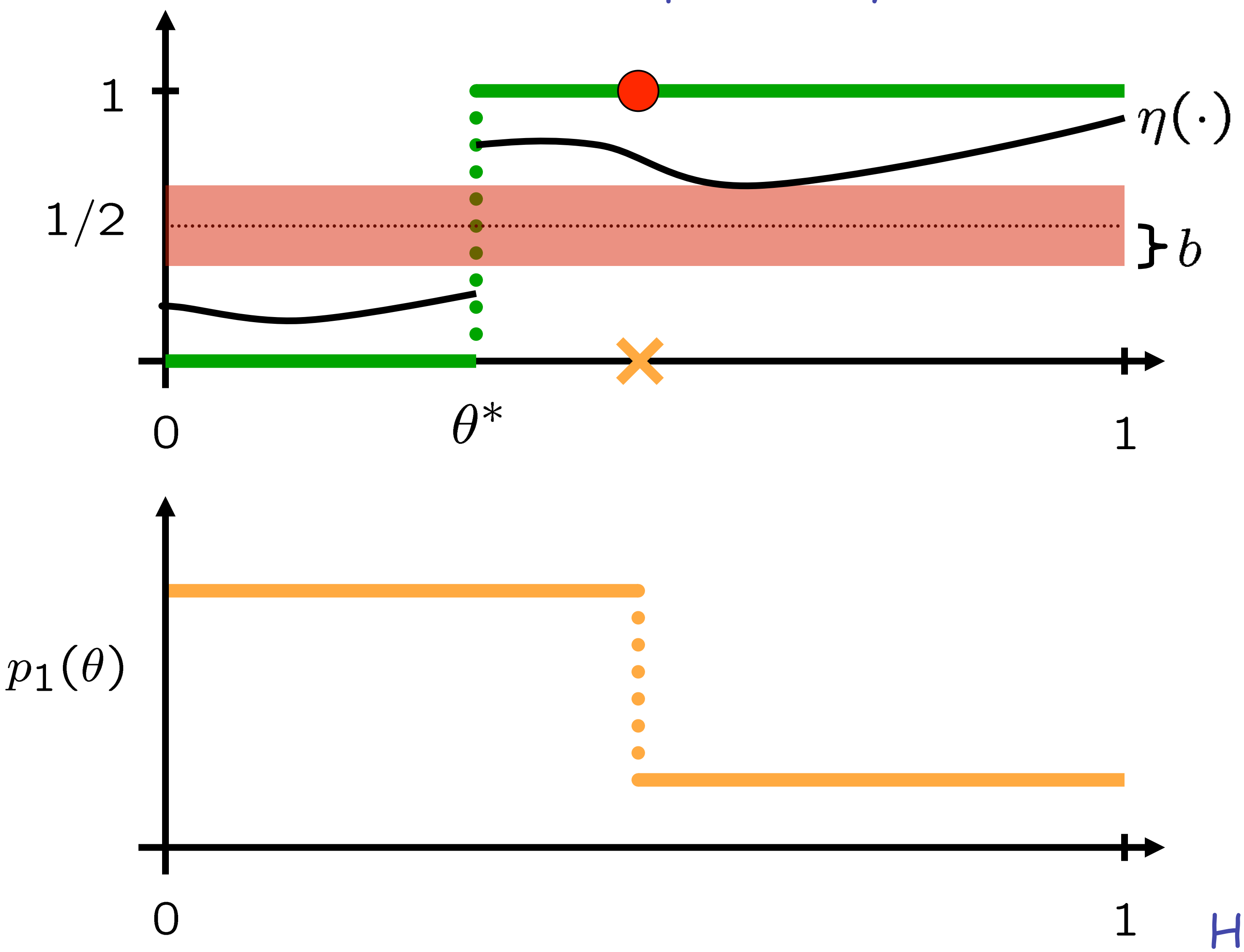
Collect an erroneous label with probability $\leq 1/2 - b$



Horstein, '63

Active Learning – Bounded Noise

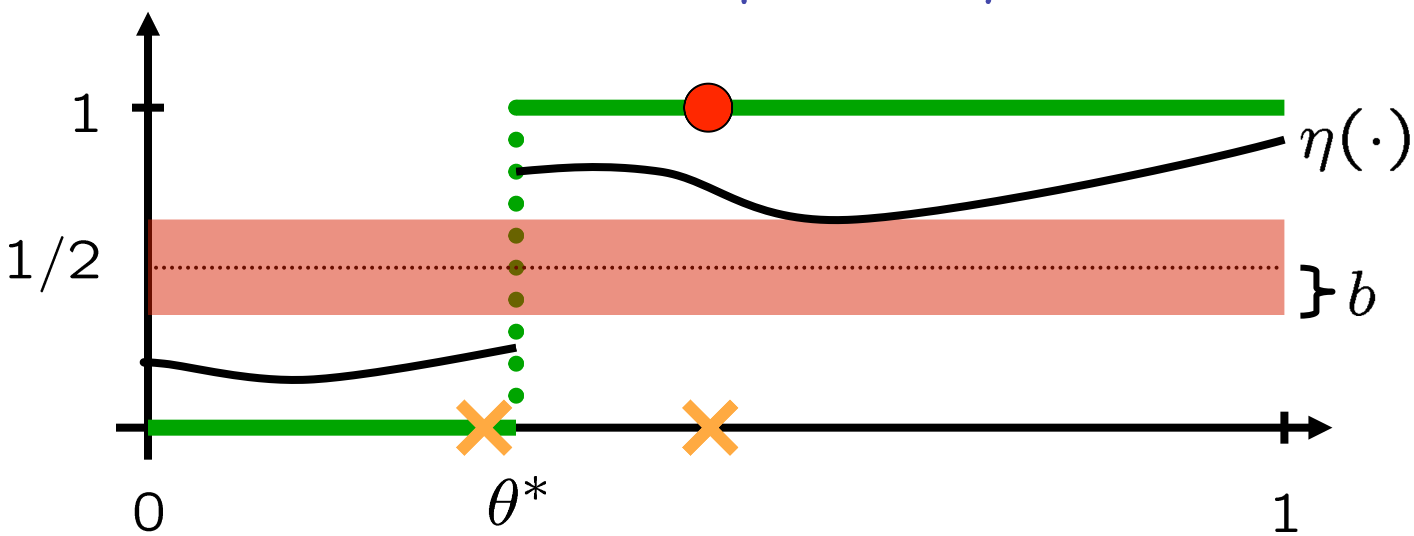
Collect an erroneous label with probability $\leq 1/2 - b$

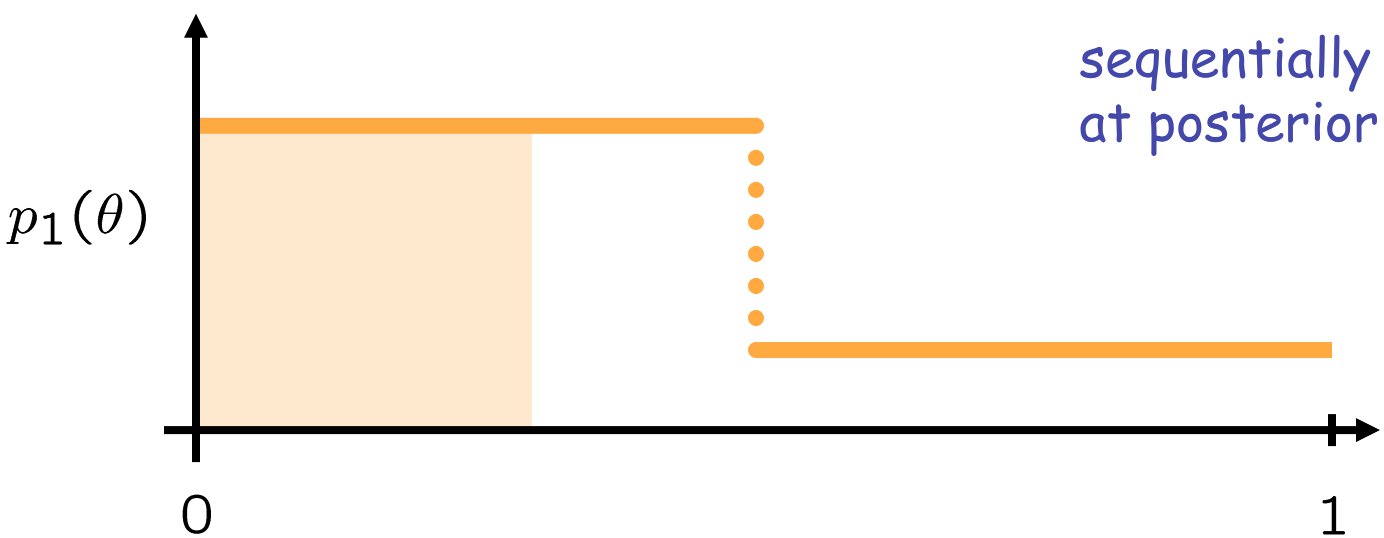


Horstein, '63

Active Learning – Bounded Noise

Collect an erroneous label with probability $\leq 1/2 - b$



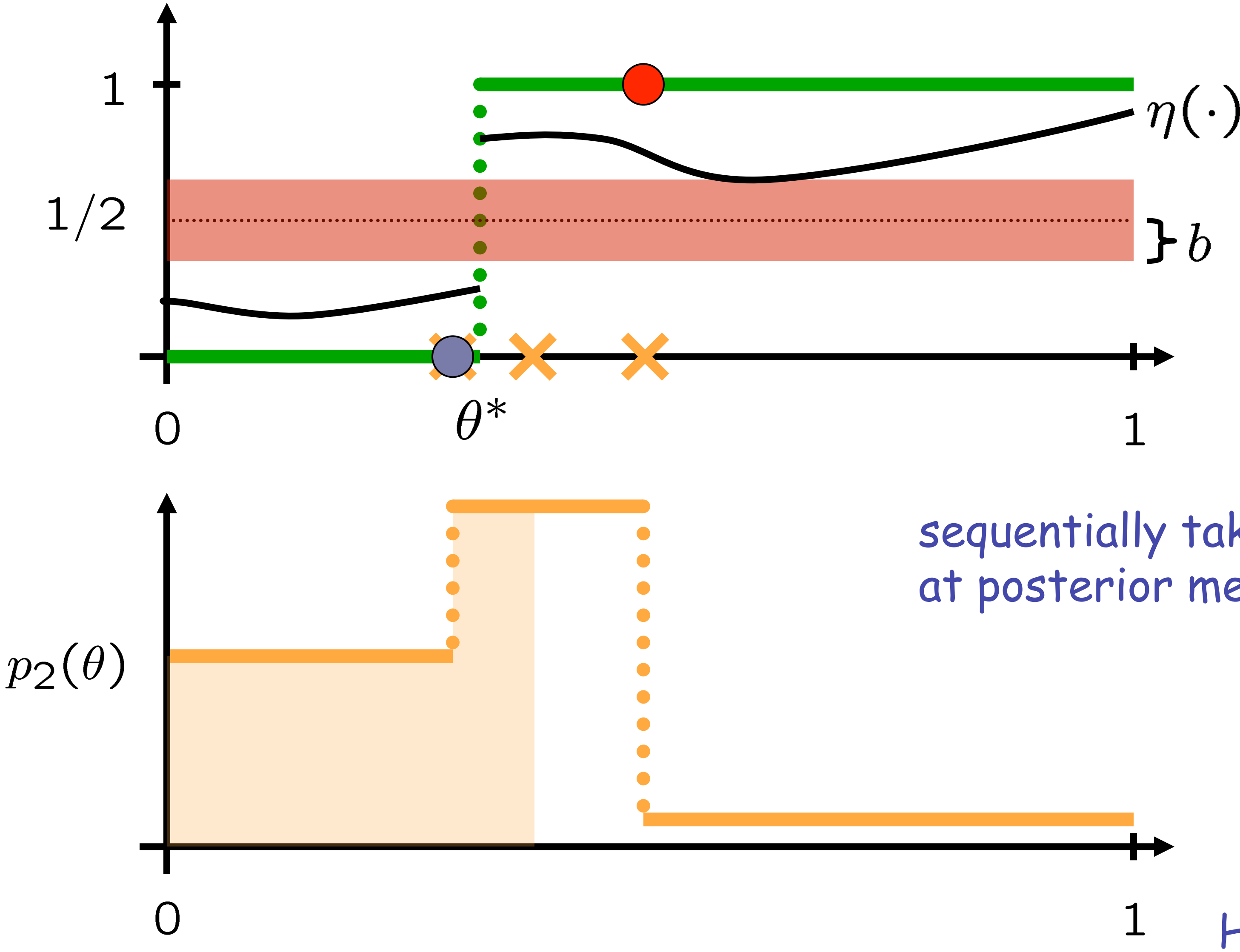


sequentially take samples at posterior median

Horstein, '63

Active Learning – Bounded Noise

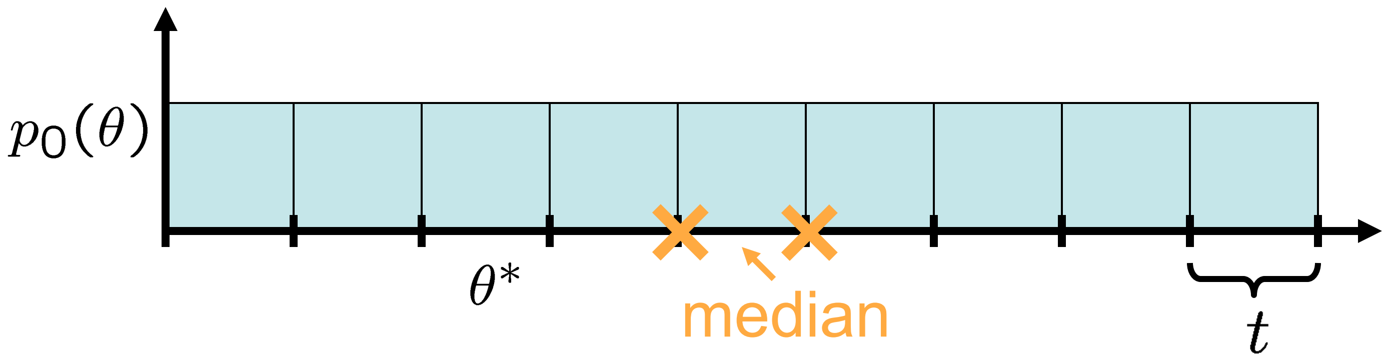
Collect an erroneous label with probability $\leq 1/2 - b$



sequentially take samples at posterior median

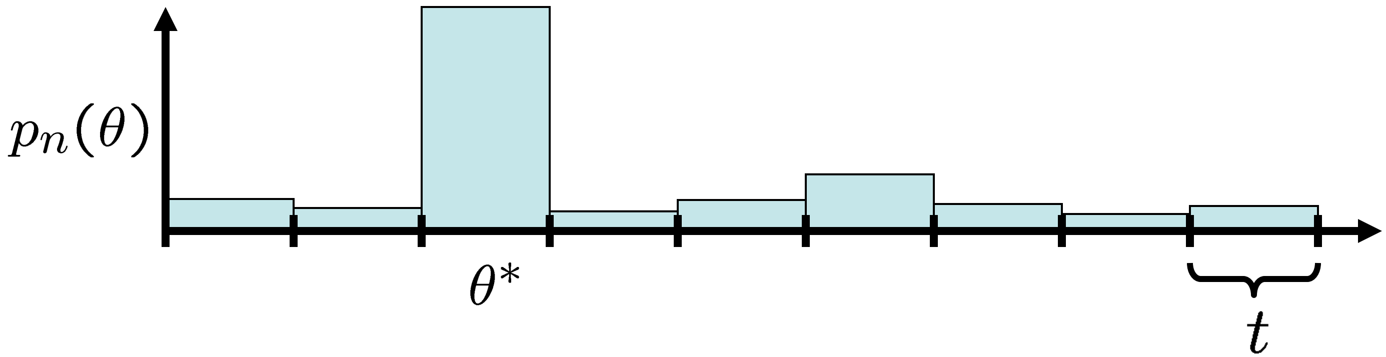
Horstein, '63

Burnashev-Zigangirov (BZ) Algorithm ‘73



search over a discrete grid

Burnashev-Zigangirov (BZ) Algorithm '73



search over a discrete grid

$$\Pr(\theta^* \text{ not in heaviest bin}) \leq \frac{1}{t}\exp(-nb^2)$$

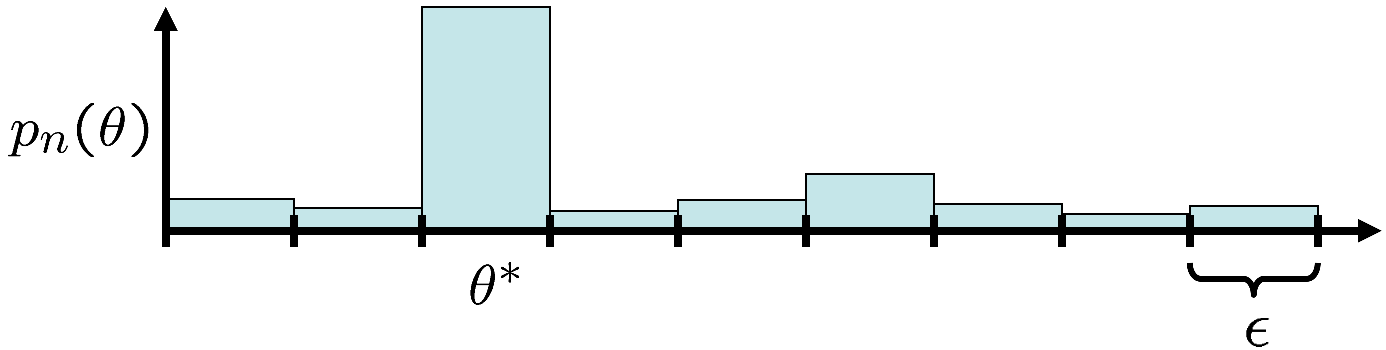
$$\mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \leq \mathbb{E}[|\widehat{\theta}_n - \theta^*|] \leq \underbrace{t}_{\text{approximation error}} + \underbrace{\tfrac{1}{t}\exp(-nb^2)}_{\text{estimation error}}$$

balancing the two terms

$$t = \exp(-\tfrac{b^2}{2}n)$$

$$\Rightarrow \quad \mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \leq 2\exp(-\tfrac{b^2}{2}n)$$

Burnashev-Zigangirov (BZ) Algorithm ‘73



The previous analysis implies also that

$$\mathbb{P}\left(\mathcal{E}(\widehat{G}_n) > \epsilon\right) \ < \delta \ ,$$

if the number of samples n is greater than

$$S(\epsilon, \delta, G^*) = \frac{1}{b^2}\left(\log\left(\frac{1}{\epsilon\delta}\right)\right) = \frac{1}{b^2}\left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)$$

Active vs. Passive – Bounded Noise

Theorem:

Under the **active sampling** scenario

$$\sup_{P_{XY} \in \text{Bounded\_Noise}} \mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \leq 2\exp(-\frac{b^2}{2}n)$$

Compare with the lower bounds for passive learning

$$\inf_{G_n} \sup_{P_{XY} \in \text{Bounded\_Noise}} \mathbb{E}\left[\mathcal{E}(G_n)\right] \succeq 1/n$$

Even with measurement uncertainty the active learning gains are HUGE!!!

Active vs. Passive – Bounded Noise

In terms of sample complexity:

Active learning:

$$\sup_{P_{XY} \in \text{Bounded\_Noise}} S(\epsilon, \delta, \widehat{G}_n) \sim \frac{1}{b^2}\left(\log\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)$$

Passive learning:

$$\inf_{\widehat{G}_n} \sup_{P_{XY} \in \text{Bounded\_Noise}} S(\epsilon, \delta, \widehat{G}_n) \sim \frac{1}{\epsilon}\log\left(\frac{1}{\delta}\right)$$

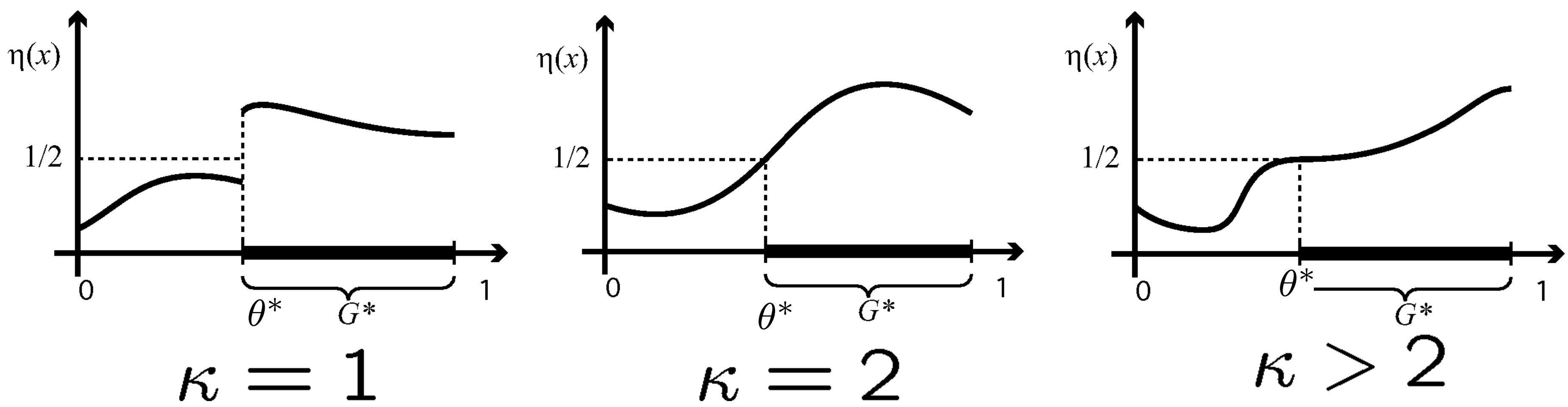
Significantly fewer samples are needed to achieve the same accuracy…

Characterizing the Noise Level

"Noise" characterization near boundary:

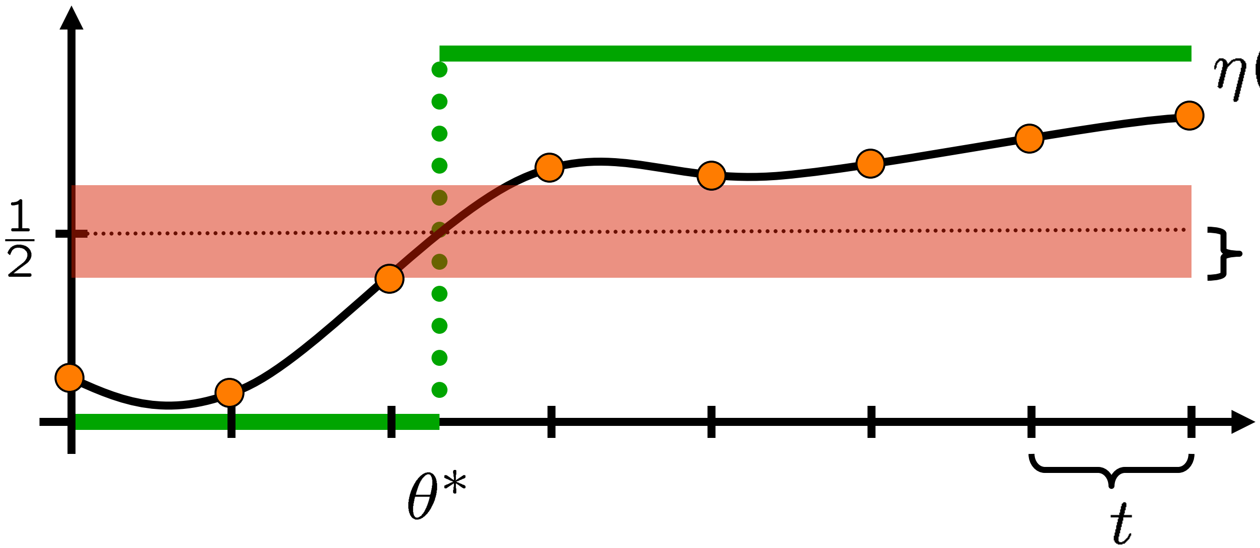
Let $\kappa \geq 1$ and assume there exist constants $c, C, \delta > 0$ so that $\forall x$ such that $|\eta(x) - 1/2| \leq \delta$

$$c|x - \theta^*|^{\kappa-1} \leq |\eta(x) - 1/2| \leq C|x - \theta^*|^{\kappa-1}$$



recall $\eta(x) = P(Y = 1|X = x)$.

Unbounded Noise $(\kappa > 1)$



(assume θ^* is not too close to bin edges)

$\} \sim c\ t^{\kappa-1}$, since $|\eta(x) - 1/2| \geq c\ |x - \theta^*|^{\kappa-1}$

very similar to the bounded noise case replacing b by $c\ t^{\kappa-1}$

$$\mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] = \mathbb{E}\left[\int_{\widehat{G}_n \Delta G^*} |2\eta(x) - 1| \mathrm{d}x\right]$$

$$\preceq \mathbb{E}\left[\int_{\widehat{G}_n \Delta G^*} |x - \theta^*|^{\kappa-1} \mathrm{d}x\right]$$, since $|\eta(x) - 1/2| \leq C|x - \theta^*|^{\kappa-1}$

$$\preceq \mathbb{E}[|\widehat{\theta}_n - \theta^*|^\kappa]$$

Unbounded Noise $(\kappa > 1)$

$$\mathbb{E}[|\widehat{\theta}_n - \theta^*|^\kappa] \;\le\; \underbrace{t^\kappa}_{\textbf{approximation error}} + \underbrace{\tfrac{1}{t}\exp(-nc^2t^{2\kappa-2})}_{\textbf{estimation error}}$$

balancing the two terms $\Rightarrow$ $t \sim \left(\dfrac{\log(n)}{n}\right)^{\frac{1}{2\kappa-2}}$

$$\boxed{\mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \;\preceq\; \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}}}$$

A practical modification of the BZ algorithm can be devised achieving the above bound without the alignment assumption.

Active vs. Passive – Unbounded noise

Theorem:

Under the **active sampling** scenario

$$\sup_{P_{XY} \in \text{Thresh}(\kappa)} \mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \preceq \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}}$$

Compare with the lower bounds for passive learning

$$\inf_{G_n} \sup_{P_{XY} \in \text{Thresh}(\kappa)} \mathbb{E}\left[\mathcal{E}(G_n)\right] \succeq n^{-\frac{\kappa}{2\kappa-1}}$$

Active learning has much faster error decay, especially when κ is small

Example: $\kappa = 2$

active $\Rightarrow$ n^{-1}

passive $\Rightarrow$ $n^{-2/3}$

Active vs. Passive – Unbounded noise

Theorem:

Under the **active sampling** scenario

$$\sup_{P_{XY} \in \text{Thresh}(\kappa)} \mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \preceq \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}}$$

Compare with the lower bounds for passive learning

$$\inf_{G_n} \sup_{P_{XY} \in \text{Thresh}(\kappa)} \mathbb{E}\left[\mathcal{E}(G_n)\right] \succeq n^{-\frac{\kappa}{2\kappa-1}}$$

Active learning has much faster error decay, especially when κ is small

Example: $\kappa \to 1$

active ⇨ n^{-p} , $p \to \infty$

passive ⇨ $\to n^{-1}$

Active vs. Passive – Unbounded noise

Theorem:

Under the **active sampling** scenario

$$\sup_{P_{XY} \in \text{Thresh}(\kappa)} \mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \preceq \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}}$$

Compare with the lower bounds for passive learning

$$\inf_{G_n} \sup_{P_{XY} \in \text{Thresh}(\kappa)} \mathbb{E}\left[\mathcal{E}(G_n)\right] \succeq n^{-\frac{\kappa}{2\kappa-1}}$$

Active learning has much faster error decay, especially when κ is small

Can we do even better with active sampling ?

Lower Bound – Active Learning

Theorem:

Under the **active sampling** scenario

$$\inf_{G_n, S_n} \sup_{P_{XY} \in \mathrm{Thresh}(\kappa)} \mathbb{E}\left[\mathcal{E}(G_n)\right] \succeq n^{-\frac{\kappa}{2\kappa-2}}$$

sampling strategy

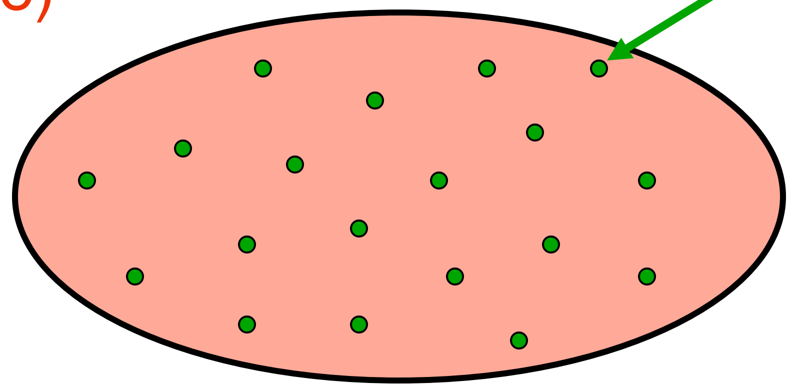
The modified BZ algorithm nearly achieves this bound

$$\sup_{P_{XY} \in \mathrm{Thresh}(\kappa)} \mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \preceq \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa-2}}$$

Lower Bound Proof Technique

Reduce the original problem to a multiple hypotheses test

$$\inf_{G_n, S_n} \sup_{P_{XY} \in \underbrace{\text{Thresh}(\kappa)}_{\text{big (infinite) class}}} \mathbb{E}\left[\mathcal{E}(G_n)\right] \geq \inf_{G_n, S_n} \sup_{P_{XY} \in \underbrace{\Psi}_{\text{finite subclass}}} \mathbb{E}\left[\mathcal{E}(G_n)\right]$$



Key fact: A sufficiently challenging subclass Ψ can be chosen independently of the classification rule and sampling strategy

Lower Bound Proof Technique

$$\inf_{G_n, S_n} \sup_{P_{XY} \in \Psi} \mathcal{E}(G_n)$$

$$\Psi = \left\{ P^{(1)}_{XY}, P^{(2)}_{XY}, \ldots, P^{(1)}_{XY} \right\}$$

Two conflicting goals: elements of Ψ must be such that:

- Hard to distinguish from data:

$$\Rightarrow \quad P^{(i)}_{XY} \text{ and } P^{(j)}_{XY} \text{ are "close"}$$

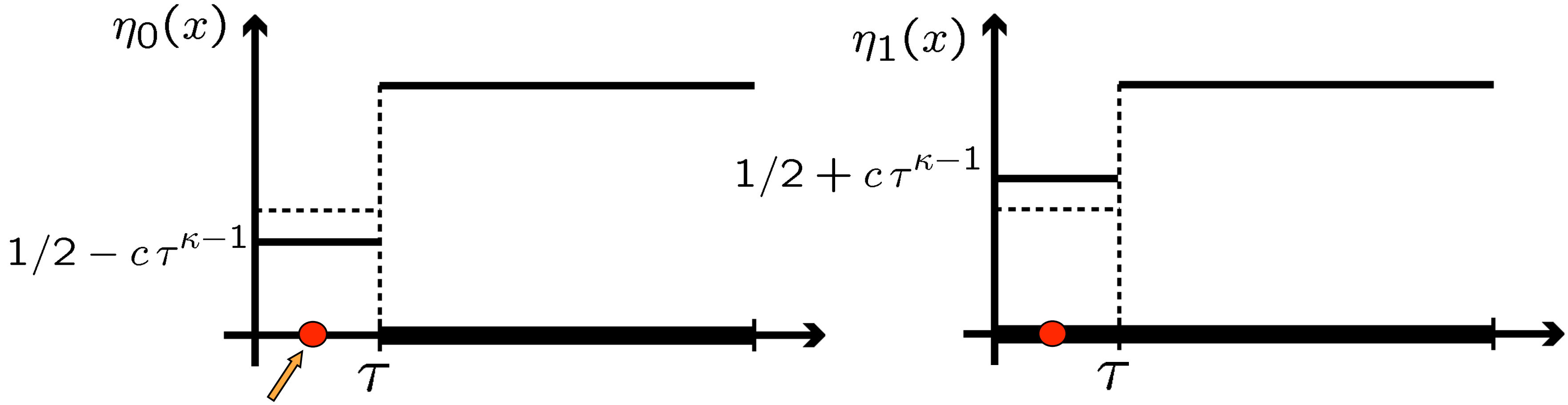
- If an estimator infers the wrong distribution then we incur a significant error

$$\Rightarrow \quad R(G^{*(i)}) - R(G^{*(j)}) \text{ is large if } i \neq j$$

Proof Sketch

special case: consider only lower regularity constraint

$$c|x-\theta^*|^{\kappa-1} \;\leq\; |\eta(x)-1/2| \;\leq\; \cancel{C|x-\theta^*|^{\kappa-1}}$$



best possible sampling location

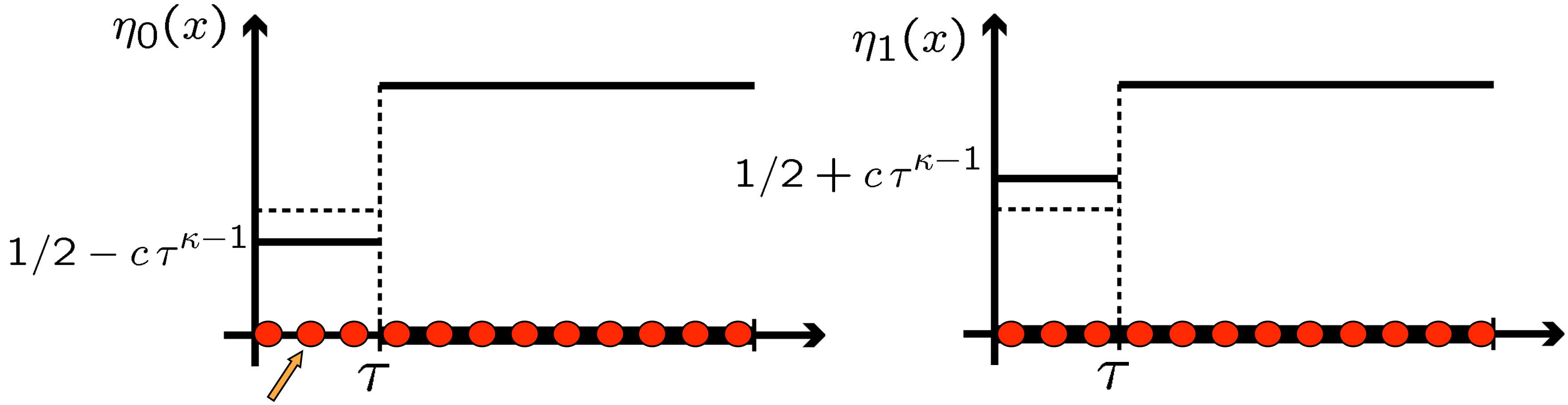
$$\Pr(\text{choosing wrong hypothesis}) \;\geq\; \text{fnc}\left[\mathrm{KL}(P_{1,n}\|P_{0,n})\right]$$

"cost" of being wrong: $\left|R(G^{*(0)}) - R(G^{*(1)})\right| \;= 2c\tau^{\kappa}$

$$\mathrm{KL}(P_{1,n}\|P_{0,n}) \sim 8c^2 n\tau^{2\kappa-2} \quad \Rightarrow \quad \tau \sim n^{-1/(2\kappa-2)}$$

$$\inf_{S_n,G_n}\max_{\theta\in\{0,1\}} \Pr_\theta\left(\mathcal{E}(G_n) \geq c n^{-\kappa/(2\kappa-2)}\right) \;\geq\; \text{const} > 0$$

Lower Bound Proof – Passive Sampling

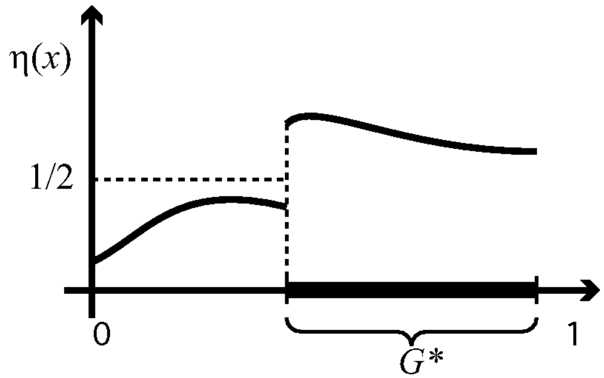


Only a fraction τ of the samples are informative

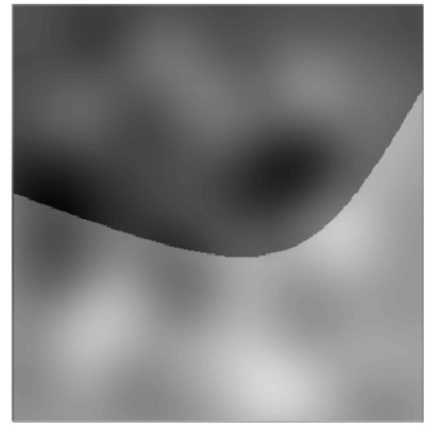
$$\mathrm{KL}(P_{1,n} \| P_{0,n}) \sim 8c^2\; n\tau^{2\kappa-2} \cdot \tau \quad \Rightarrow \quad \tau \sim n^{-1/(2\kappa-1)}$$

$$\inf_{S_n, G_n} \max_{\theta \in \{0,1\}} \Pr_\theta\left(\mathcal{E}(G_n) \geq c n^{-\kappa/(2\kappa-1)}\right) \;\geq\; \text{const} > 0$$

From 1D to Multiple Dimensions



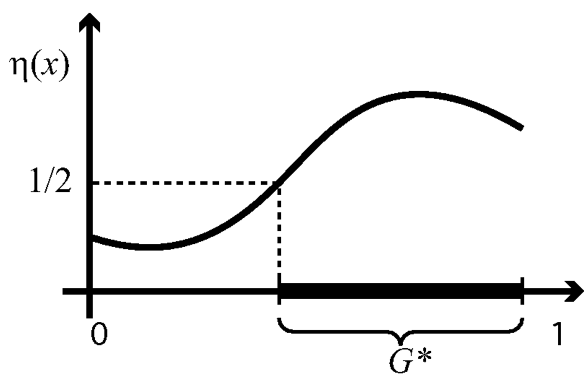
$\eta(x_1, \ldots, x_d)$



$\kappa = 1$

One-dimensional threshold

Multidimensional "threshold"



$\eta(x_1, \ldots, x_d)$

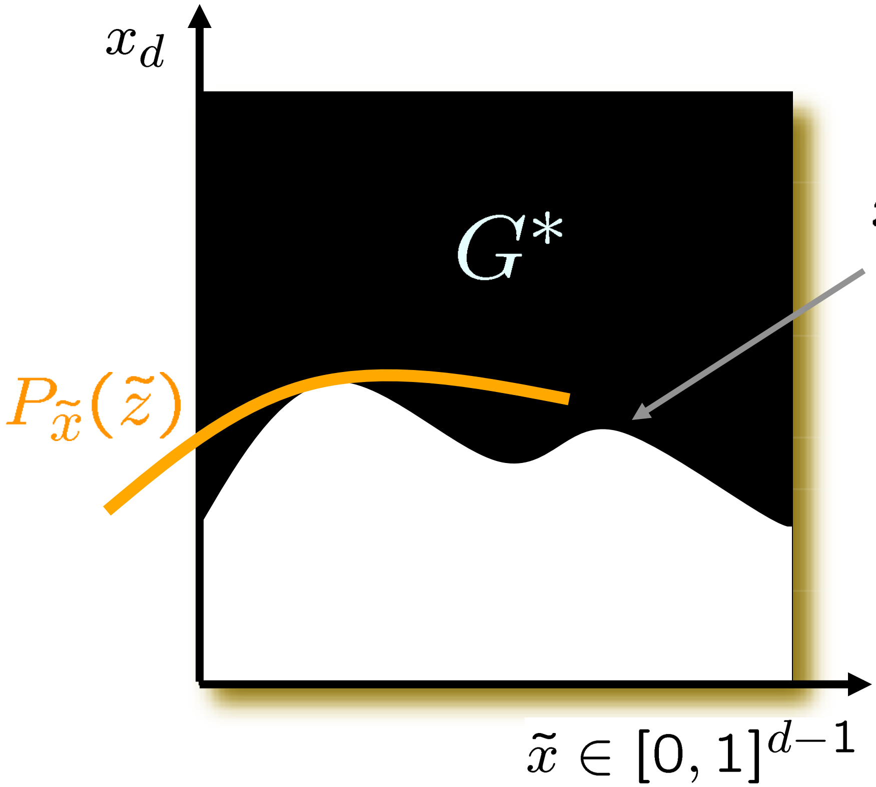
$\kappa > 1$

$X \sim \mathsf{Unif}([0,1]^d)$

Multidimensional Settings

Consider the class of "boundary fragment" sets

(Korostelev & Tsybakov '93, Donoho '97, '99)



Hölder smooth function

$x_d = g^*(\tilde{x})$

$$|g^*(\tilde{z}) - P_{\tilde{x}}(\tilde{z})| \leq L\|\tilde{z} - \tilde{x}\|^{\alpha}$$

where $L, \alpha > 0$, and $P_{\tilde{x}}(\cdot)$ denotes the degree $\lfloor \alpha \rfloor$ Taylor polynomial of g^* expanded around $\tilde{x}$

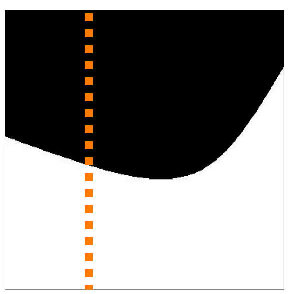
$$G^* \in \mathcal{G}_{\text{BF}} := \{\text{the sets defined above}\}$$

Noise Condition – Transition Smoothness

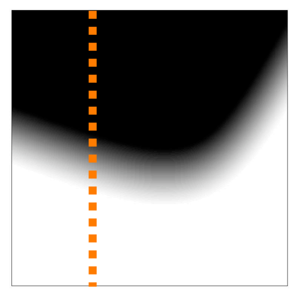
$$x = (\tilde{x}, x_d) \in [0,1]^d \text{ and } G^* \in \mathcal{G}_{\mathsf{BF}}$$

Let $\kappa \geq 1$ and assume there exist constants $c, C, \delta > 0$ so that $\forall x$ such that $|\eta(x) - 1/2| \leq \delta$

$$c|x_d - g^*(\tilde{x})|^{\kappa-1} \leq |\eta(x) - 1/2| \leq C|x_d - g^*(\tilde{x})|^{\kappa-1}$$

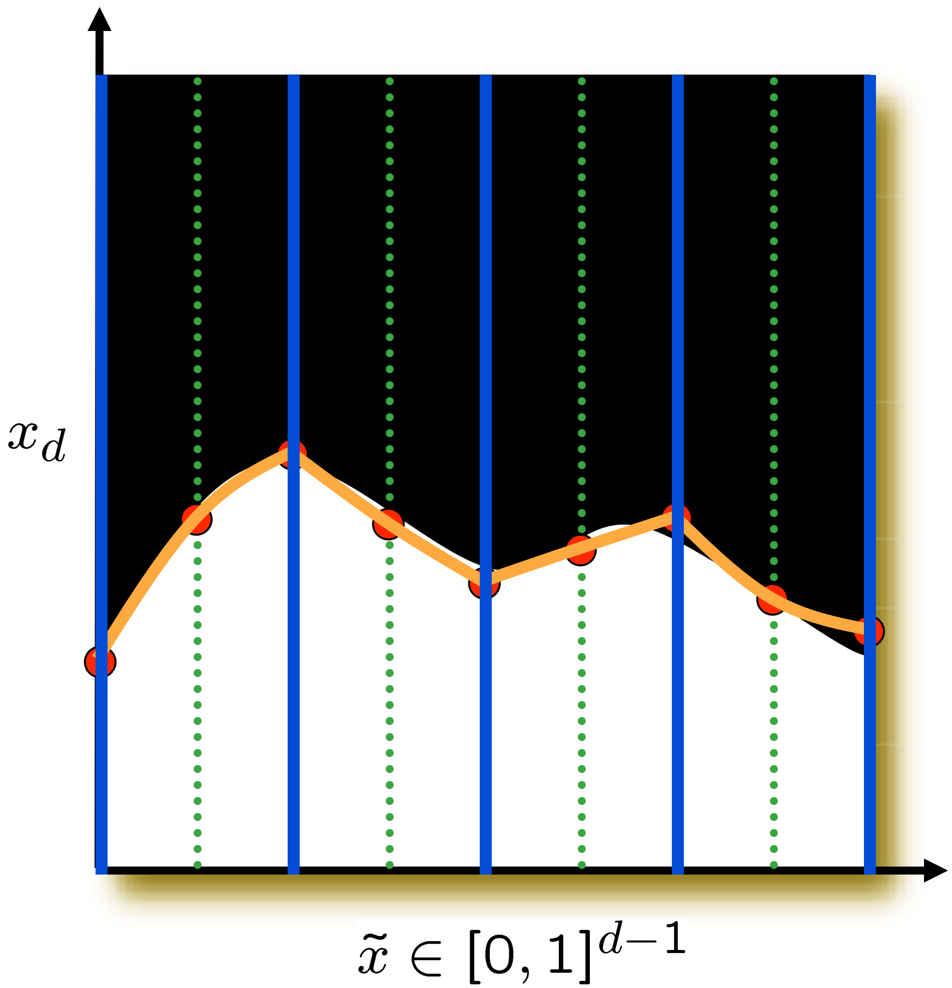


sharp transition
$\kappa = 1$



smooth transition
$\kappa > 1$

Active Learning for Boundary Fragments



1. Take M^{d-1} uniformly spaced lines in $\tilde{x}$ coordinates

2. Estimate change-pts at each location via BZ with N samples

3. Partition into $M^{d-1}/\lceil\alpha\rceil^{d-1}$ bins and poly-interpolate change-pt estimates

Estimating Boundary Fragments

$\bar{g}$ - Best poly. interpolant (best model in our class)

$$\mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \preceq \mathbb{E}\left[\int |\widehat{g} - g^*|^\kappa\right]$$
$$= \mathbb{E}\left[\int |\underbrace{(\bar{g} - g^*)}_{\text{approximation error}} + \underbrace{(\widehat{g} - \bar{g})}_{\text{estimation error}}|^\kappa\right]$$

spacing between interpolation points $= M^{-1}$

$$|\widehat{g} - \bar{g}| \sim \max_{\tilde{x} \in \text{Grid}} |\widehat{g}(\tilde{x}) - \bar{g}(\tilde{x})|$$

BZ $\Rightarrow$ with very high probability

$$|\widehat{g}(\tilde{x}) - \bar{g}(\tilde{x})| \preceq \left(\frac{\log N}{N}\right)^{\frac{1}{2\kappa-2}}, \quad \forall \tilde{x}$$

$\Rightarrow$ $|\bar{g} - g^*| \preceq M^{-\alpha}$

$\Rightarrow$ $|\widehat{g} - \bar{g}| \preceq \left(\frac{\log N}{N}\right)^{\frac{1}{2\kappa-2}}$

Estimating Boundary Fragments

$$\mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \preceq \mathbb{E}\left[\int |(\bar{g} - g^*) + (\widehat{g} - \bar{g})|^\kappa\right]$$
$$\preceq \left(M^\alpha + \left(\frac{\log N}{N}\right)^{1/(2\kappa-2)}\right)^\kappa$$

We have the constraint $M^{d-1} N \leq n =$ total # samples

Take $$M = \left\lfloor n^{\frac{1}{\alpha(2\kappa-2)+d-1}} \right\rfloor$$
$$N = \lfloor n/M^{d-1} \rfloor$$

$\Rightarrow$ $$\mathbb{E}\left[\mathcal{E}(\widehat{G}_n)\right] \preceq \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa+\rho-2}}, \quad \rho = (d-1)/\alpha$$

Upper and Lower Bounds

Theorem:

$$\left(\frac{1}{n}\right)^{\frac{\kappa}{2\kappa+\rho-2}} \preceq \inf_{S_n,G_n} \sup_{P_{XY}\in \mathsf{BF}(\alpha,\kappa)} \mathbb{E}\left[\mathcal{E}(G_n)\right] \preceq \left(\frac{\log n}{n}\right)^{\frac{\kappa}{2\kappa+\rho-2}}$$

$$, \quad \rho = (d-1)/\alpha$$

Note: The constructive estimation strategy is near optimal

Compare with passive sampling (similar to Tsybakov '04)

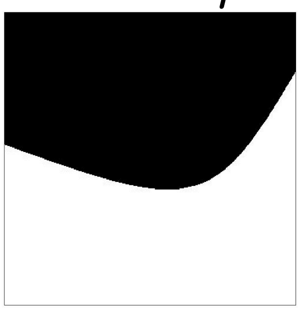
$$\inf_{G_n} \sup_{P_{XY}\in \mathsf{BF}(\alpha,\kappa)} \mathbb{E}[R(G_n)] - R(G^*) \asymp \left(\frac{1}{n}\right)^{\frac{\kappa}{2\kappa+\rho-1}}$$

Implication: General Classes

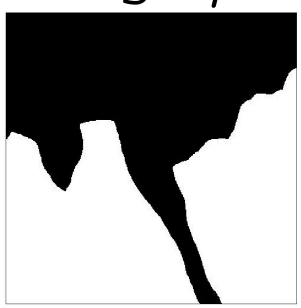
Active learning lower bounds for general classes

$$\inf_{G_n,S_n} \sup_{P_{XY}\in \mathrm{Class}(\rho,\kappa)} \mathbb{E}\left[\mathcal{E}(G_n)\right] \succeq \begin{cases} n^{-\frac{\kappa}{2\kappa+\rho-2}} & \text{active} \\ n^{-\frac{\kappa}{2\kappa+\rho-1}} & \text{passive} \end{cases}$$

small ρ

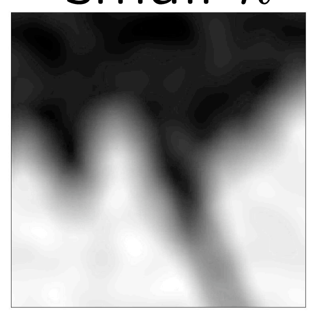


large ρ

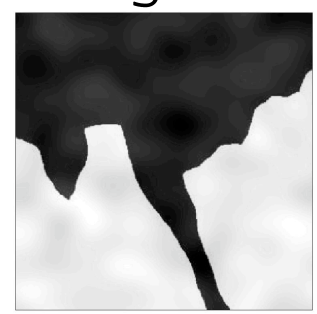


$\rho -$ Complexity of decision boundary (metric entropy of Bayes class)

small κ



large κ



$\kappa -$ Smoothness of transition

These results can be generalized for estimation of level sets and functions

Why are these Results Important?

Indicate **when** active learning can be beneficial, and **quantify** the gain.

> **Active Learning helps when problem complexity is spatially concentrated (e.g., locating a boundary or threshold)**

The threshold and boundary fragment classes provide benchmark problems for the design and assessment of practical general-purpose algorithms

Practical problems:

multiple change-points, arbitrary boundary sets, etc...

Outline

Binary Classification and the fundamental limits of active learning

Algorithmic considerations and Active Learning in practice…

Hypothesis and Query/Feature Spaces

$$\begin{aligned} \mathcal{H} &= \text{space of hypotheses or models} \\ \mathcal{X} &= \text{space of queries or unlabeled features} \end{aligned}$$

h^* is the true model (might not belong to $\mathcal{H}$).

Noiseless Learning : $x \in \mathcal{X} \;\rightarrow\; y = h^*(x)$

Noisy Learning : $x \in \mathcal{X} \;\rightarrow\; y = h^*(x) + \text{noise}$

Active Learning: Sequentially select *most informative* queries/examples based on past queries/examples and responses.

A Simple Algorithm for Separable Case

Cohn, Atlas and Ladner '92 $h : \mathcal{X} \to \{-1, +1\},\ h^* \in \mathcal{H}$

initialize: $i = 1,\ \mathcal{H}_1 = \mathcal{H}$

Region of Disagreement

while $|\mathcal{H}_i| > 1$

1. Select $x_i \in \{\text{any } x \in \mathcal{X} \text{ where } h \in \mathcal{H}_i \text{ disagree}\}$
2. Query with x_i to obtain $y_i = h^*(x_i)$
3. Set $\mathcal{H}_{i+1} = \{h \in \mathcal{H}_i : h(x_i) = y_i\},\ i = i + 1$

Version Space

CAL algorithm may also be operated in an online fashion

Flavors of Active Learning Analysis

How many queries or labeled examples are required ?

Extended Teaching Dimension a combinatorial parameter of $\mathcal{H}$ and $\mathcal{X}$ (Hegedüs '95, Hellerstein et al '96)

Disagreement Coefficient a measure of the growth of the region of disagreement (Hanneke '07)

Neighborly Condition geometric relationship between $\mathcal{X}$ and $\mathcal{H}$ (Nowak '08)

Unfortunately theoretically sound methods that have been developed are for the most part either computational intractable, or empirically not so good…

What if there is Noise or Mismatch?

Noise-tolerance:

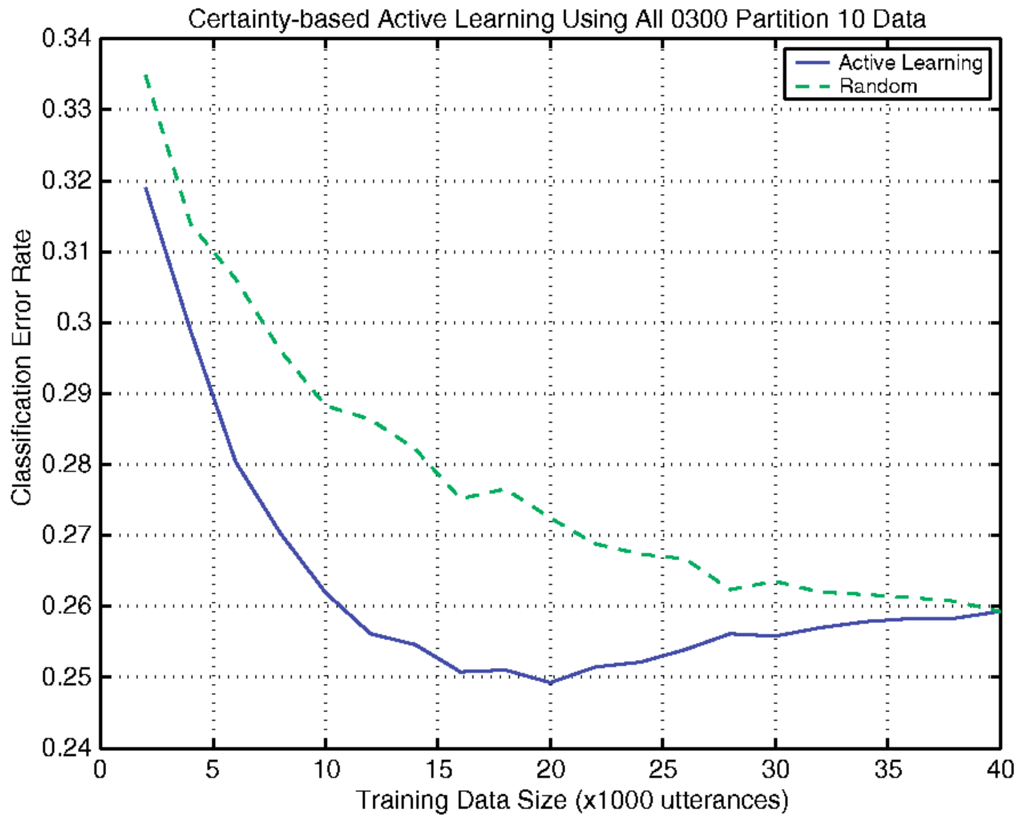
1. stochastic version space (all hypotheses with errors that could be explained by noise alone)
2. repeated querying (collect several labels for uncertain examples until highly confident in probably correct labeling)
3. hypothesis weighting (weight each hypothesis according to its prediction performance)

Agnostic active learning: If h^* is not in $\mathcal{H}$, then can we at least guarantee performance equal to that of passive learning? Yes

Split sample budget into three equal parts:
- active learning with $1/3$ of sample budget $\rightarrow \widehat{h}_n$
- passive learning with $1/3$ of sample budget $\rightarrow \widetilde{h}_n$
- remaining $1/3$ of samples are collected from region of disagreement between $\widehat{h}_n$ and $\widetilde{h}_n$, best hypothesis wins!

Active Learning in Practice

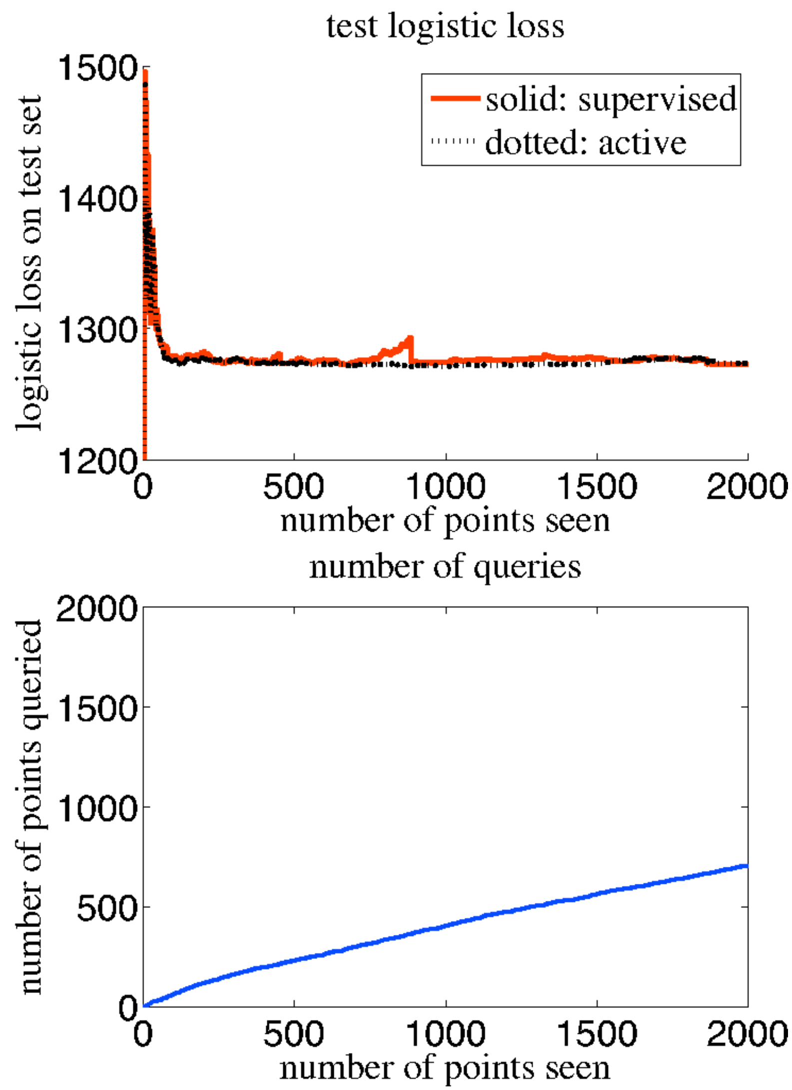
The most successful active learning methods are based on empirical ideas, and are not guaranteed to always work. Generally their performance is reported only in the settings where these succeed.



Tur, Tur and Shapire, “Combining active and semi-supervised learning for spoken language understanding” 2005

Active Learning in Practice

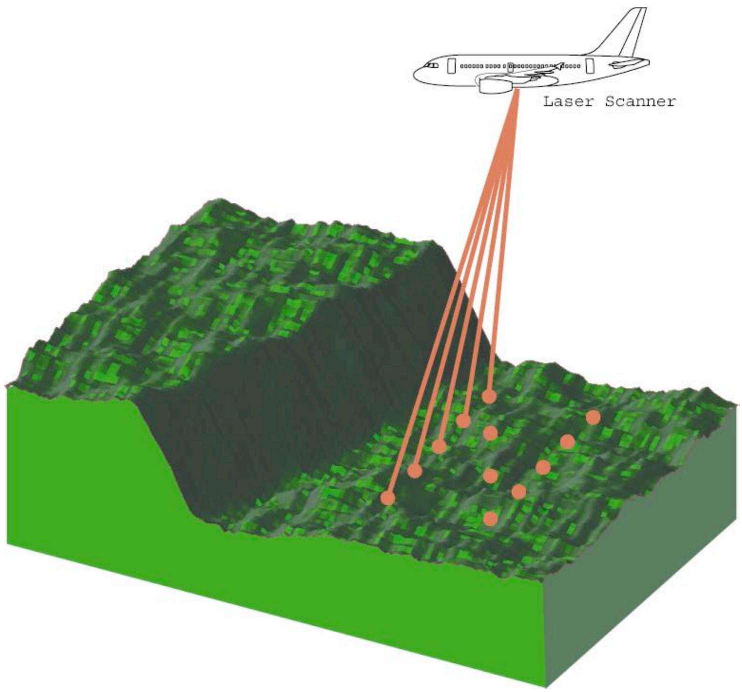
A mostly practical general purpose algorithm for the classification setting with provable performance.



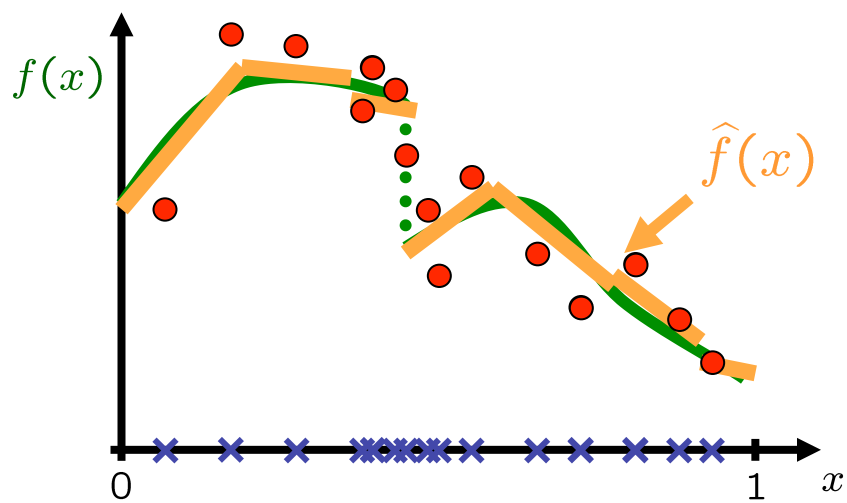
Beygelzimer, Dasgupta & Langford, “Importance Weighted Active Learning”, ICML 2009

Active Learning in Regression

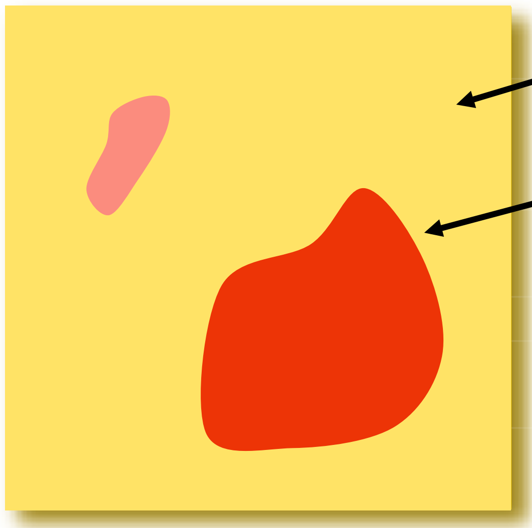
Goal: Accurately "learn" a function/set, as fast as possible, by strategically focusing in regions of interest



Function Estimation



Regression of Piecewise Constant Functions



constant away from boundary

$d-1$ non-fractal boundaries separating constant regions

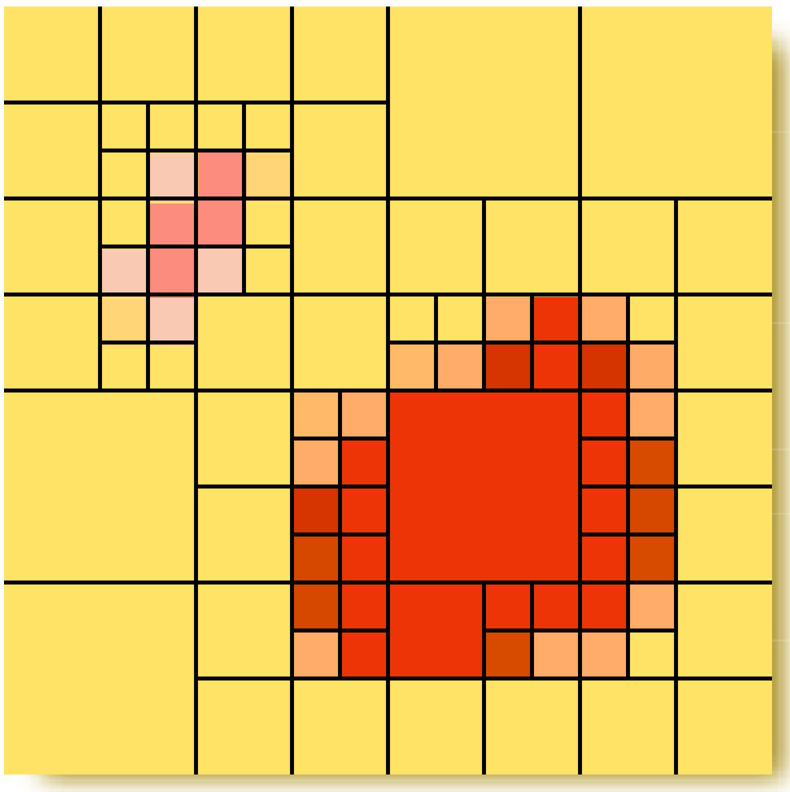
Goal: Construct an estimator $\widehat{f}_n : [0,1]^d \to \mathbb{R}$ based on point samples $\{\boldsymbol{X}_i, Y_i\}_{i=1}^n$ minimizing

$$\mathbb{E}\left[\|\widehat{f}_n - f\|^2\right]$$

Observation Model: $Y_i = f(\boldsymbol{X}_i) + W_i, \quad W_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

Passive Learning in the PC Class

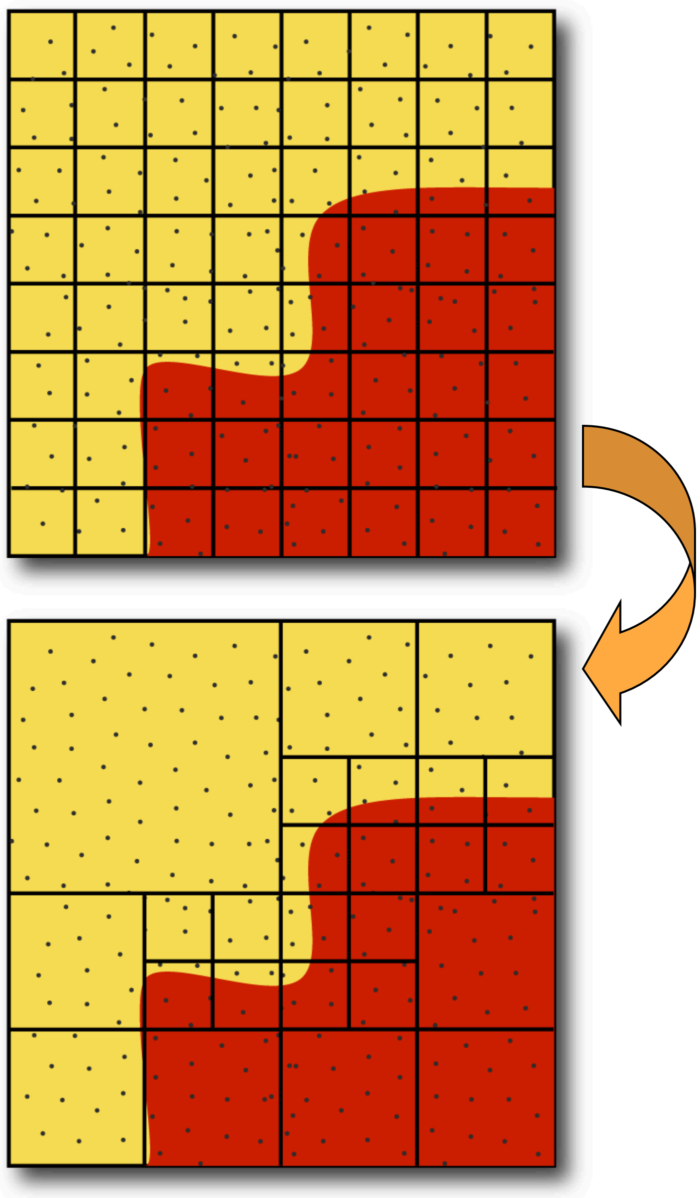
A multiscale approach (the "wavelet" idea):



- Distribute sample points uniformly over $[0,1]^d$
- Recursively divide the domain into hypercubes
- Prune the partition, adapting to the data
- Fit a model in each partition set

Idea: Use Recursive Dyadic Partitions to **find** the boundary

Active Learning in the PC Class



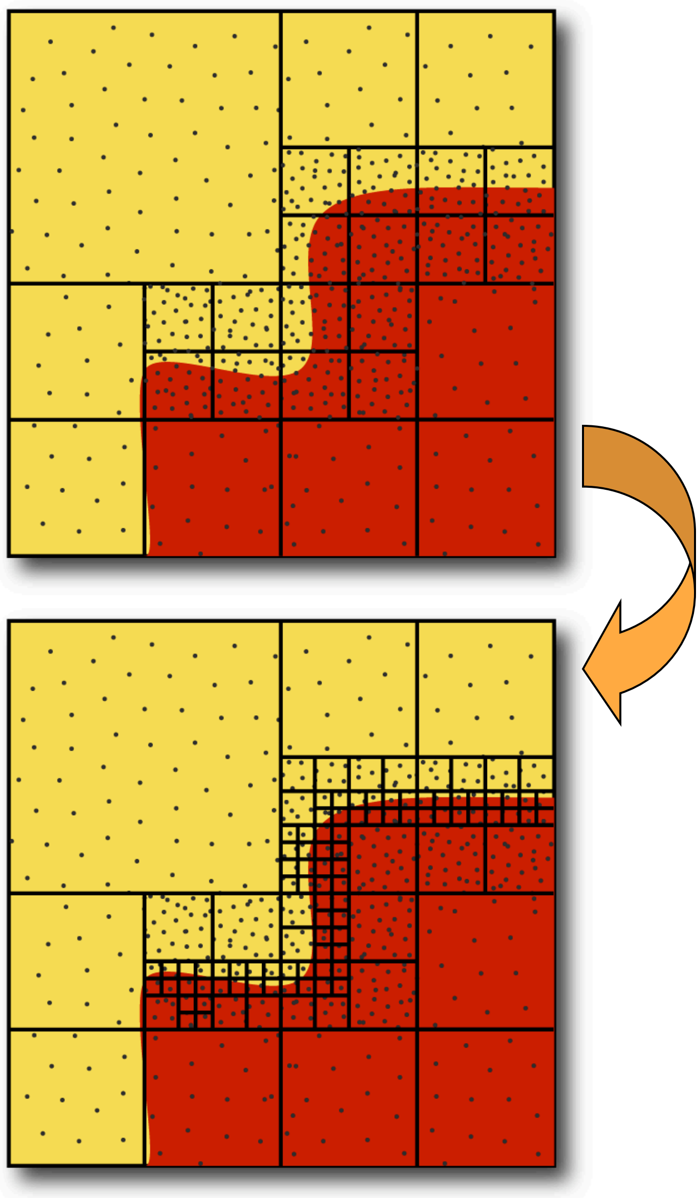
Stage 1: "<u>Oversample</u>" at coarse resolution

- n/2 samples uniformly distributed
- **<u>Limit the resolution</u>**: many more samples than cells
- biased, but very low variance result (high approximation error, but low estimation error)

"boundary zone" is reliably detected

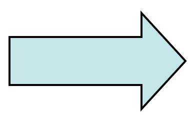
Some delicate issues relating alignment of partition and boundaries

Active Learning in the PC Class



Stage 2: Critically sample in boundary zone

- n/2 samples uniformly distributed within boundary zone
- construct fine partition around boundary
- prune partition according to standard multiscale methods



high resolution estimate of boundary

How to choose the right balance between detection of the boundary and refinement ???

Performance Bounds

Theorem (Castro, Willett & Nowak '05):

Let f be a piecewise constant function whose boundaries separating constant regions are locally Lipschitz. Then

$$\mathbb{E}[\|\widehat{f}_n - f\|^2] \preceq \left(\frac{\log n}{n}\right)^{1/(d-1+1/d)} .$$

Moreover, for every $\epsilon > 0$ there is a multi-stage estimator $\widehat{f}_n$ satisfying

$$\mathbb{E}[\|\widehat{f}_n - f\|^2] \preceq n^{-1/(d-1+\epsilon)}$$

Best possible error rates:

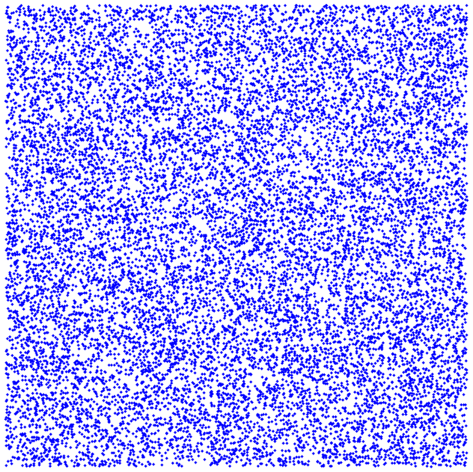
active $\Rightarrow$ $n^{-\frac{1}{d-1}} = 1/n$

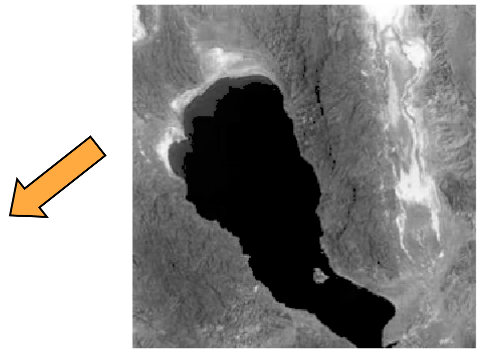
passive $\Rightarrow$ $n^{-\frac{1}{d}} = 1/\sqrt{n}$

$, d = 2$

Function Estimation

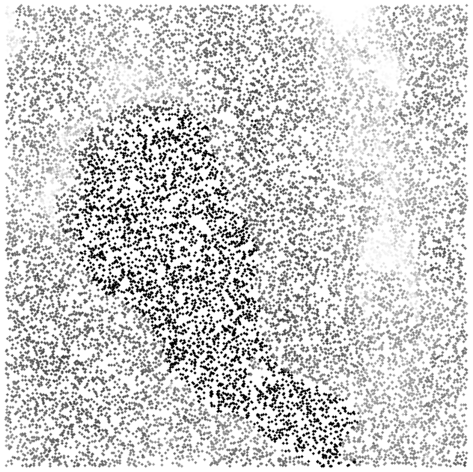
16384 non-adaptive samples

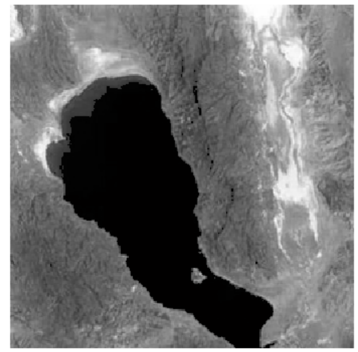


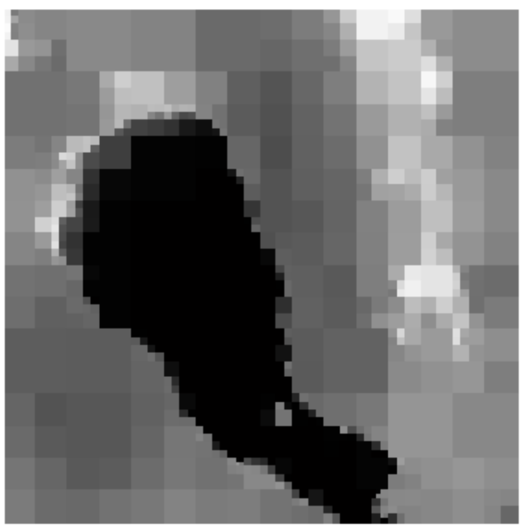


Function Estimation

16384 non-adaptive samples

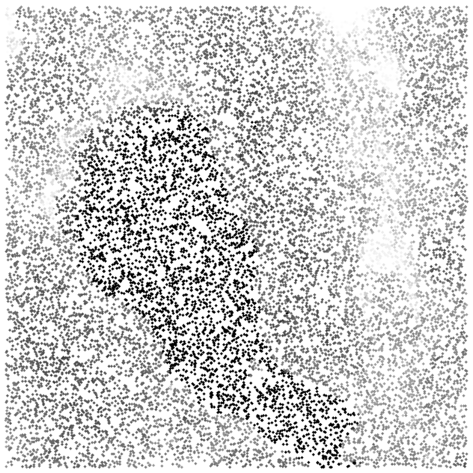


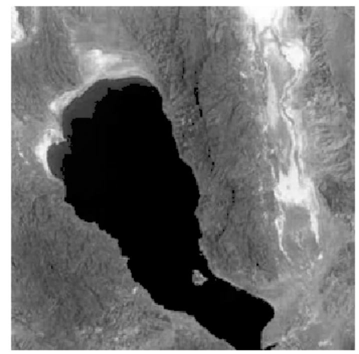




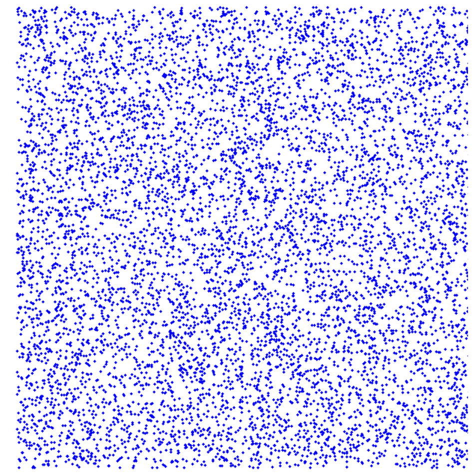
Function Estimation

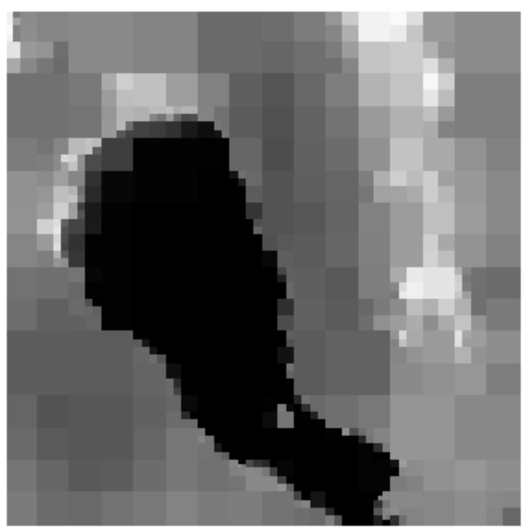
16384 non-adaptive samples





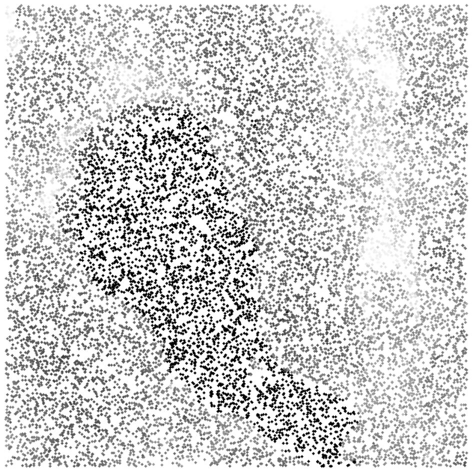
8192 non-adaptive samples

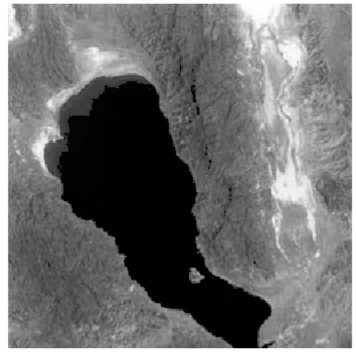




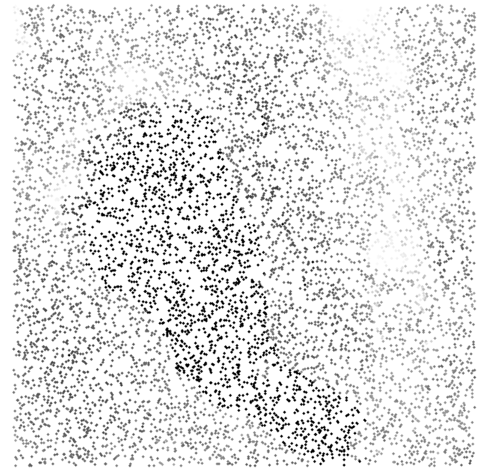
Function Estimation

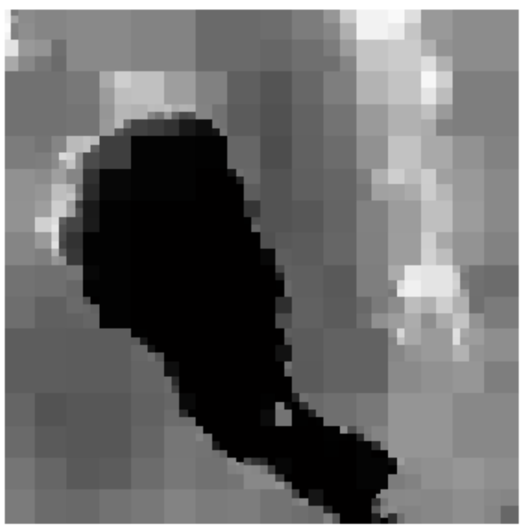
16384 non-adaptive samples





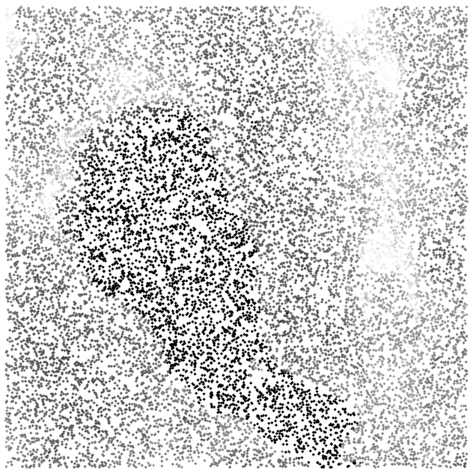
8192 non-adaptive samples

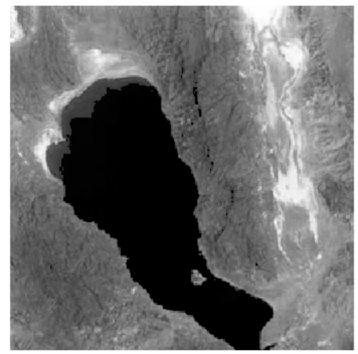




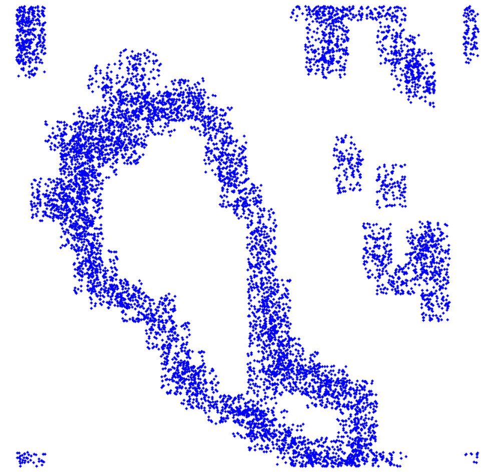
Function Estimation

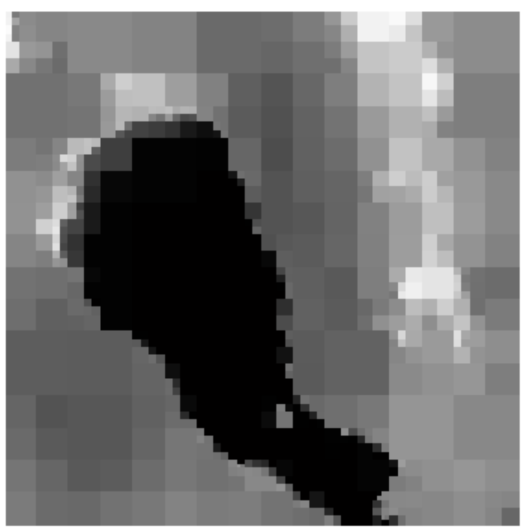
16384 non-adaptive samples



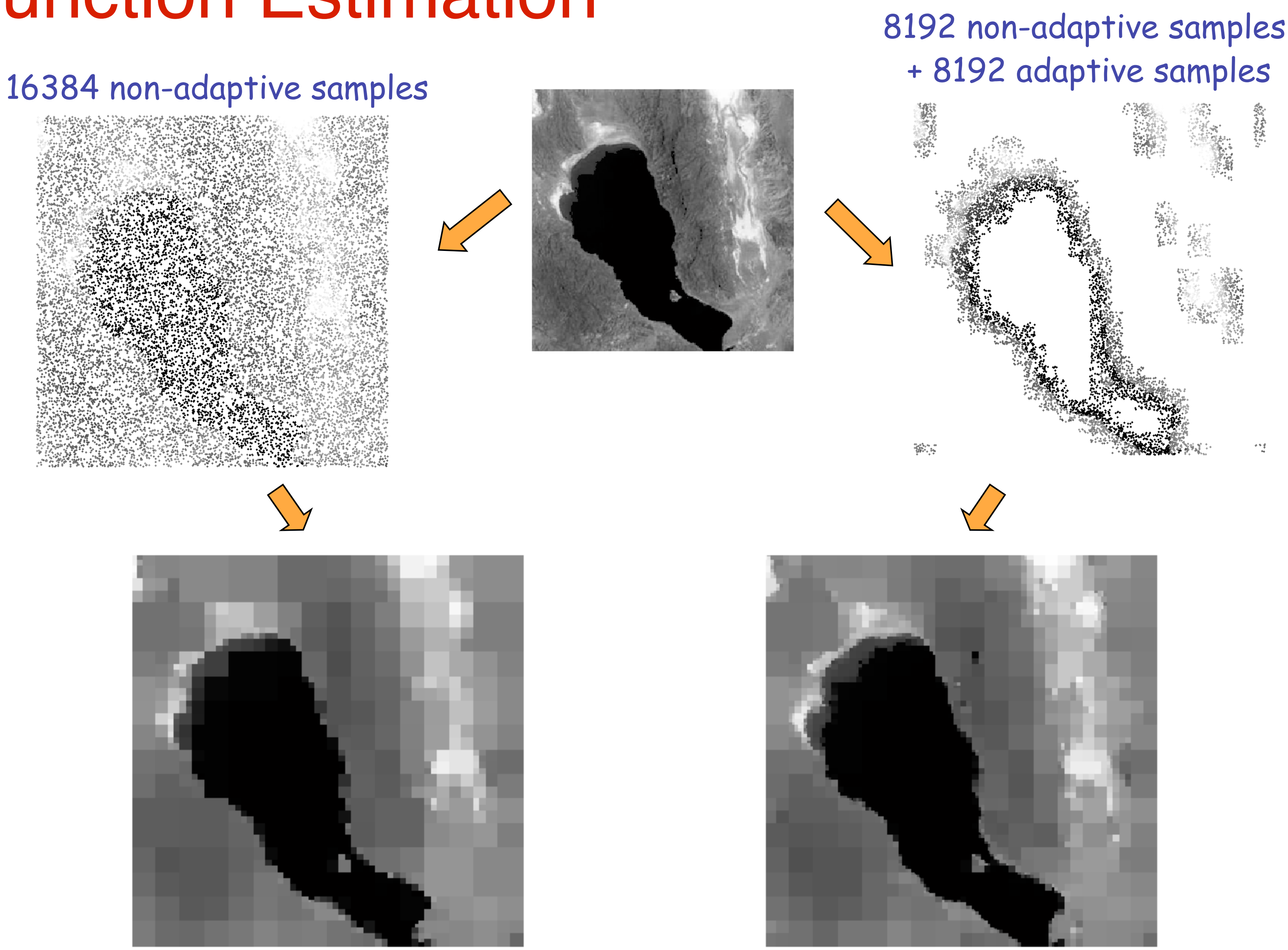


8192 non-adaptive samples
+ 8192 adaptive samples

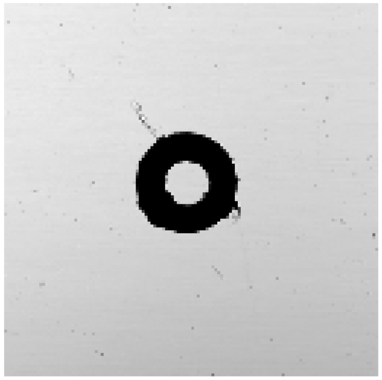


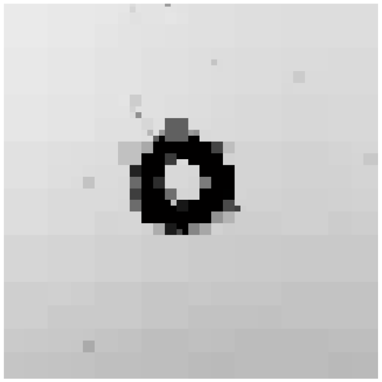


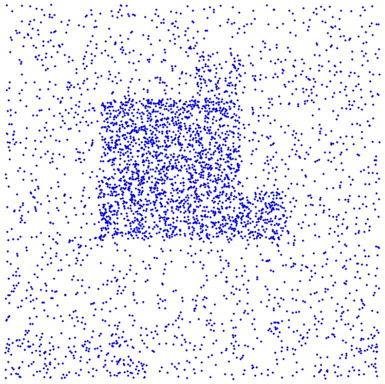
Function Estimation

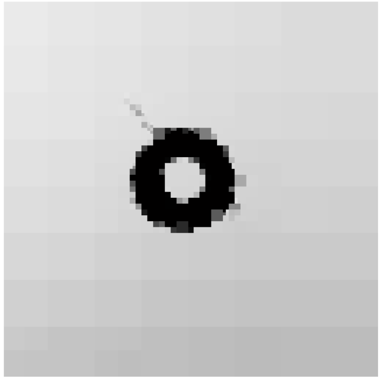


Real-World Application – Ballistic Laser Imaging


65536 Passive Samples


4096 Passive samples


Active Sample Locations

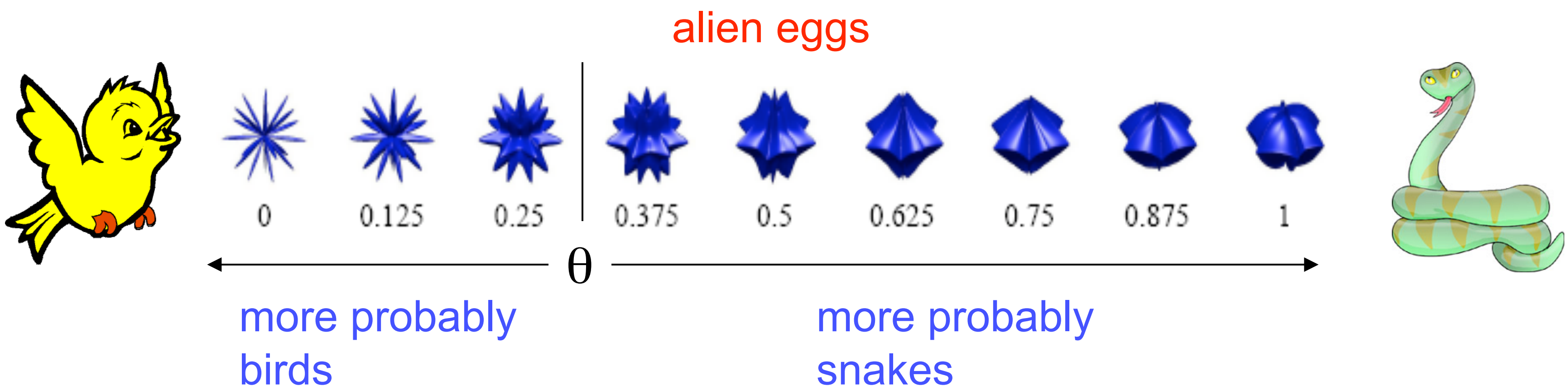

4096 active samples

Data kindly provided by Sina Farsiu (Duke)

HAL: Are you a good active learner?

Castro, Kalish, Nowak, Qian, Rogers & Zhu (NIPS 2008)

Investigate human active learning in task analogous to 1-d threshold problem



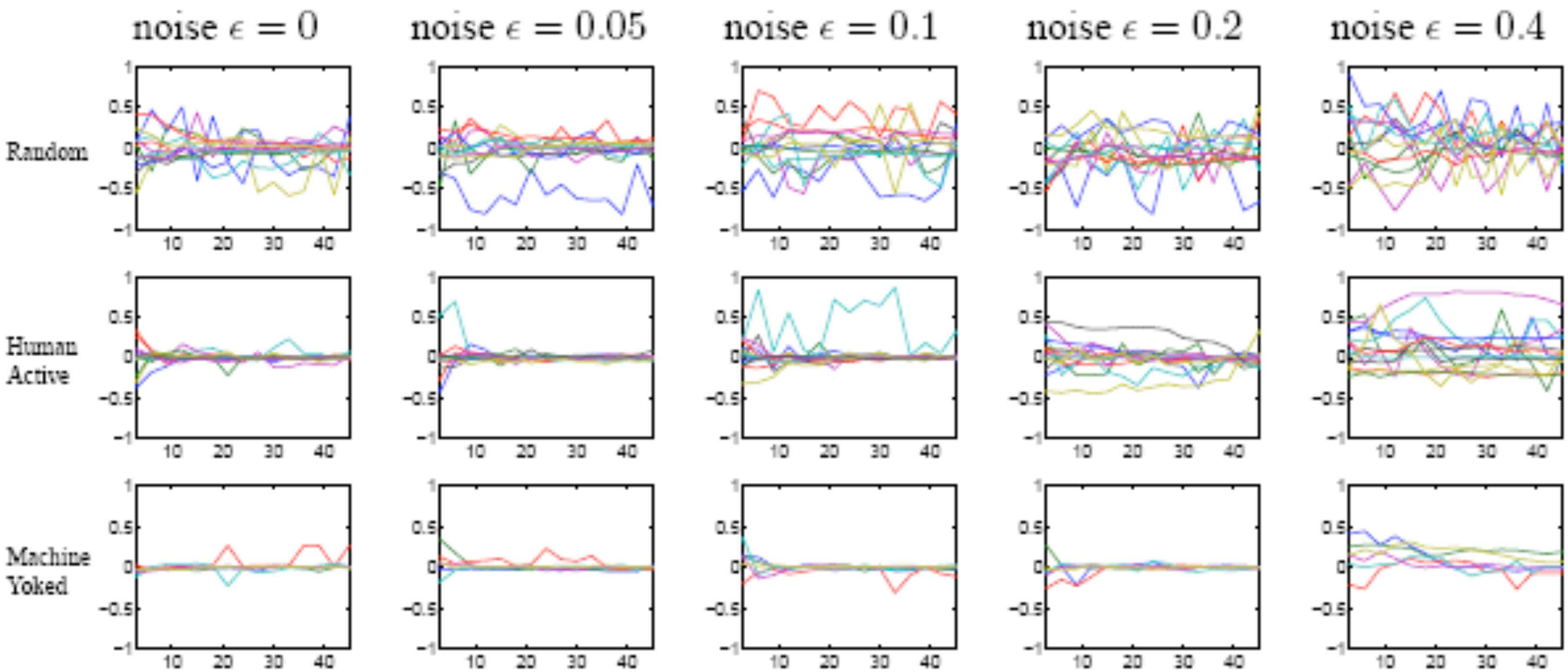
Subjects observe random egg hatchings (passive learning) or they can select eggs to hatch (active learning).

They are asked to determine the egg shape where snakes become more probable than birds.

Results: Human learning rates agree with theory, $1/n$ in passive mode and $\exp(-cn)$ in active mode.

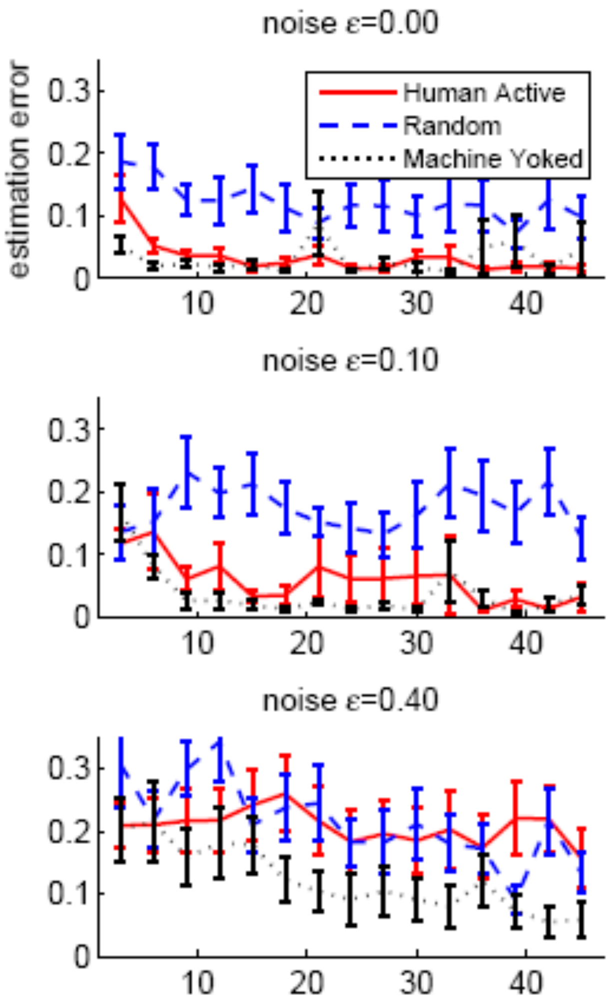
HAL: The Data

33 subjects split up among various conditions



Error vs. number of samples

HAL: Man vs. Man, Man vs. Machine



Conclusions:

1. Human learning benefits significantly from selective sampling/querying.
2. Machines may assist human learning by providing informative samples or suggesting experiments

Active Learning is an Active Area of Research

Channel Coding with Feedback

• Horstein, “Sequential decoding using noiseless feedback,” IEEE Trans. Info. Theory, vol. 9, no. 3, 1963

• Burnashev & Zigangirov, “An interval estimation problem for controlled observations,” Problems in Information Transmission, vol. 10, 1974

Active Learning and Sequential Experimental Design

• Cohn, Atlas, and Ladner, “Improving generalization with active learning,” Machine Learning, 15(2), 1994

• Fedorov, “Theory of Optimal Experiments,”. New York: Academic Press” 1972

• Freund, Seung, Shamir, and Tishby, “Selective sampling using the query by committee algorithm,” Machine Learning, vol. 28, no. 2-3, 1997

• Mackay, “Information-based objective functions for active data selection,” Neural Computation, vol. 4,, 1991

• Cohn, Ghahramani, & Jordan, “Active learning with statistical models,” Journal of Artificial Intelligence Research, 1996

• Cesa-Bianchi, Conconi, & Gentile, “Learning probabilistic linear threshold classifiers via selective sampling,” COLT 2003

Active Learning is an Active Area of Research

Active Learning and Sequential Experimental Design (cont.)

• Korostelev, “On minimax rates of convergence in image models under sequential design,” Statistics & Probability Letters, vol. 43, 1999

• Korostelev & Kim, “Rates of convergence for the sup-norm risk in image models under sequential designs,” Statistics & probability Letters, vol. 46, 2000

• Hall & Molchanov, “Sequential methods for design-adaptive estimation of discontinuities in regression curves and surfaces,” The Annals of Statistics, vol. 31, no. 3, 2003

• Castro, Willett, & Nowak, “Faster rates in regression via active learning,” NIPS 2005

• Dasgupta, “Analysis of a greedy active learning strategy,” NIPS 2004

•Dasgupta, Hsu & Monteleoni, “A general agnostic active learning algorithm,”, NIPS 2007

• Balcan, Beygelzimer & Langford, “Agnostic active learning,” ICML 2006

• Hanneke, “Teaching dimension and the complexity of active learning,”, COLT 2007

• Hanneke, “A bound on the label complexity of agnostic active learning,” ICML 2007

• Kaariainen, “Active learning in the non-realizable case,” ALT 2006

Active Learning is an Active Area of Research

Active Learning and Sequential Experimental Design (cont.)

• Castro & Nowak, "Minimax Bounds for Active Learning", IEEE Transactions on Information Theory, vol. 54, no. 5, 2008

• Hanneke, "Adaptive Rates of Convergence in Active Learning", 2009

Learning with Queries

• Hegedus, "Generalized teaching dimensions and the query complexity of learning," COLT 1995

• Nowak, "Generalized binary search", In Proceedings of the Allerton Conference 2008

• Kulkarni, Mitter, & Tsitsiklis, "Active learning using arbitrary binary valued queries," Machine Learning, 1993

•Karp and Kleinberg, "Noisy binary search and its applications. In Proceedings of the 18th ACM-SIAM Symposium on Discrete Algorithms (SODA 2007), pages 881– 890, 2007

•Angluin, "Queries revisited," Springer Lecture Notes in Computer Science: Algorithmic Learning Theory, pages 12–31, 2001.

• Hellerstein, Pillaipakkamnatt, Raghavan, & Wilkins, "How many queries are needed to learn? J. ACM, 43(5), 1996

Active Learning is an Active Area of Research

Learning with Queries (cont.)

• Garey and Graham, “Performance bounds on the splitting algorithm for binary testing,” Acta Inf., 3, 1974

• Hyafil & Rivest, “Constructing optimal binary decision trees is NP-complete,” Inf. Process. Lett., 5, 1976