Human Active Learning, NIPS 2008

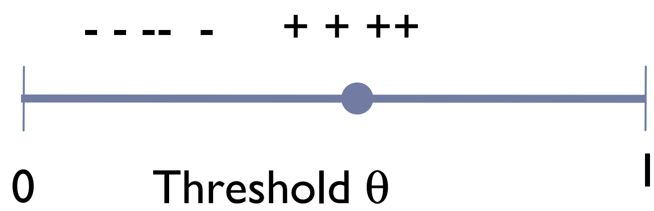
By R. Castro, C. Kalish, R. Nowak, R. Qian, T. Rogers, X. Zhu
Slides by Cheng William Hong

Active Learning

- Learner can pick examples for labeling
- For certain problems, has much better performance
- Paper focuses on application of active learning to classification
  - Both machines and humans
- No previous work attempting to quantify human active learning performance

Two category learning task

- 1D binary classification in [0,1]
- Data: (X_i, Y_i)
- Y_i is the category of X_i with probability $1 - \varepsilon$



No noise

- We have discussed this case extensively in the class
- Error from passive learning is $O(1/n)$
- Error from active learning is $O(2^{-(n+1)})$ via binary search
- What if there is noise?

In the presence of uncertainty

- Passive learning: Still polynomial error at least
- Active learning: Cannot use deterministic bisection
  - Still can do "binary search", from Bayesian estimation
  - Assume some prior distribution on θ: say it is uniformly distributed
  - Bayes' Rule:

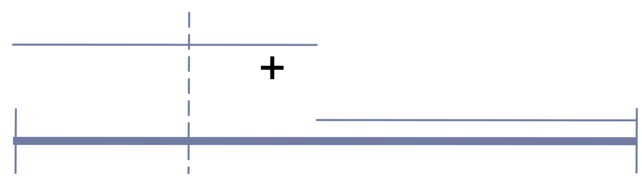
$$P(A|B) = \frac{P(B|A)\, P(A)}{P(B)}.$$

  - P(A): prior probability
  - P(A|B): posterior probability
  - P(B|A): conditional probability
  - P(B): marginal probability
- Idea: Pick point in the median of the distribution

Bayesian binary search

- Before any information:

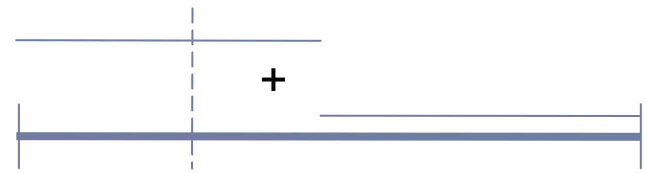
- The median of the CDF is at ½, so we pick that point, say we obtain 1
- $P(\theta > \frac{1}{2} | (X,Y) = (\frac{1}{2}, 1)) = \dfrac{P((X,Y) = (\frac{1}{2}, 1) | \theta > \frac{1}{2}) P(\theta > \frac{1}{2})}{P((X,Y) = (\frac{1}{2}, 1))} = \varepsilon$
- $P(\theta \leq \frac{1}{2} | (X,Y) = (\frac{1}{2}, 1)) = 1-\varepsilon$
- Update prior distribution:



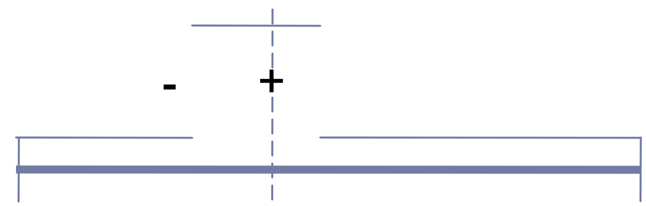
- Pick a new point in the median

Bayesian binary search

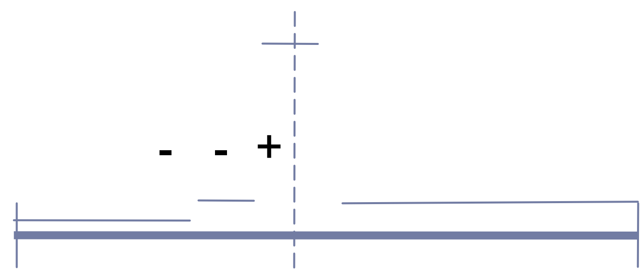
- We now have prior distribution:



- Say next label is 0

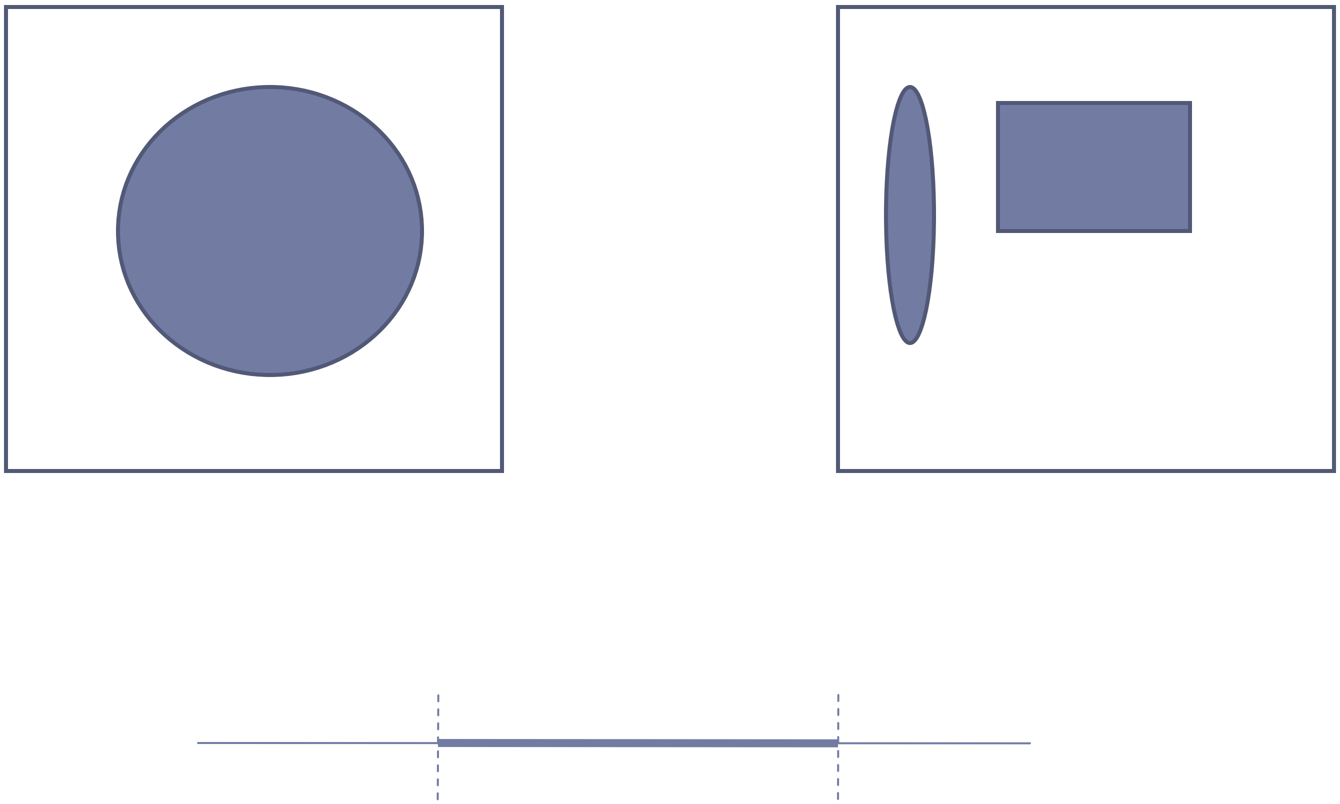


- Next label is 0



- This method works well in practice
- Can be applied to more complicated scenarios

More complicated boundaries



Mathematical bounds

- Analysis of a slightly different method with discrete query locations gives:

$$\sup_{\theta \in [0,1]} \mathbb{E}[|\hat{\theta}_n - \theta|] \leq 2 \left(\sqrt{\frac{1}{2} + \sqrt{\epsilon(1-\epsilon)}} \right)^n$$

- The performance of any passive learning algorithm is bounded by:

$$\inf_{\hat{\theta}_n} \sup_{\theta \in [0,1]} \mathbb{E}[|\hat{\theta}_n - \theta|] \geq \frac{1}{4} \left(\frac{1+2\epsilon}{1-2\epsilon} \right)^{2\epsilon} \frac{1}{n+1}$$

- Still an exponential advantage from active learning!

Minimax bounds for active learning by R. Castro and R. Nowak

- Bounded error: $\forall x \; |P(Y=1, X=x) - ½| > c, \; c > 0$
- Discusses the case of unbounded error
- Complexity of the boundary characterized by $\rho = (d-1)/\kappa$
  - d are the dimensions of the feature space
  - ρ is the Hölder regularity of the boundary
    - A function is Hölder smooth if it has continuous partial derivatives up to order k = Floor(α)
  - Behavior of $P(Y=1, X=x)$ around characterized by κ
    - $\kappa = 1$ for bounded error, >1 for unbounded

Error bounds for unbounded error

- Idea is to reduce problem to deciding among a finite collection of representative distributions
- Fastest error decay for active learning:

$$n^{-\frac{\kappa}{2\kappa+\rho-2}}$$

- Fastest error decay for passive learning:

$$n^{-\frac{\kappa}{2\kappa+\rho-1}}$$

- Active learning always superior to passive learning (fallback guarantee)
- Upper bounds for learning are similar to a logarithmic factor

How do humans learn?

- Passive learning: observe some object and its category label
- Active learning: can also ask questions

Are people good at picking examples?

- Rich literature of conflicting claims regarding people's ability to pick optimal examples
- Classic example: to assess $p \Rightarrow q$
  - People examine q instances to see if p holds, ignoring $\neg q$ instances
- Is that based on analyzing the task wrongly?
- Much of the debate in psychological literature is on task analysis and assessing performance
- Opportunity for applying the formal descriptions from machine learning
- How good are they in comparison to computers?

What are we looking for in active learning?

- Consistency
  - Generalization error should go to 0
- Fallback guarantee
  - At least as good as passive learning
- What we really want:
  - Error decreases much faster than passive learning

Questions

- Do humans perform better when they can select their own examples?
- Do they achieve the full benefit?
- Can machine learning be used to help them?
- Do the answers to the above depend on the difficulty of the problem?

Experimental Setup

- “Random”
  - Passive learning condition, subject is presented with uniformly sampled examples
- “Human-Active”
  - Active learning condition, subject selects queries and receives labels
- “Machine-Yoked”
  - Active learning with machine learning, human observes labels for queries selected by the machine learning algorithm

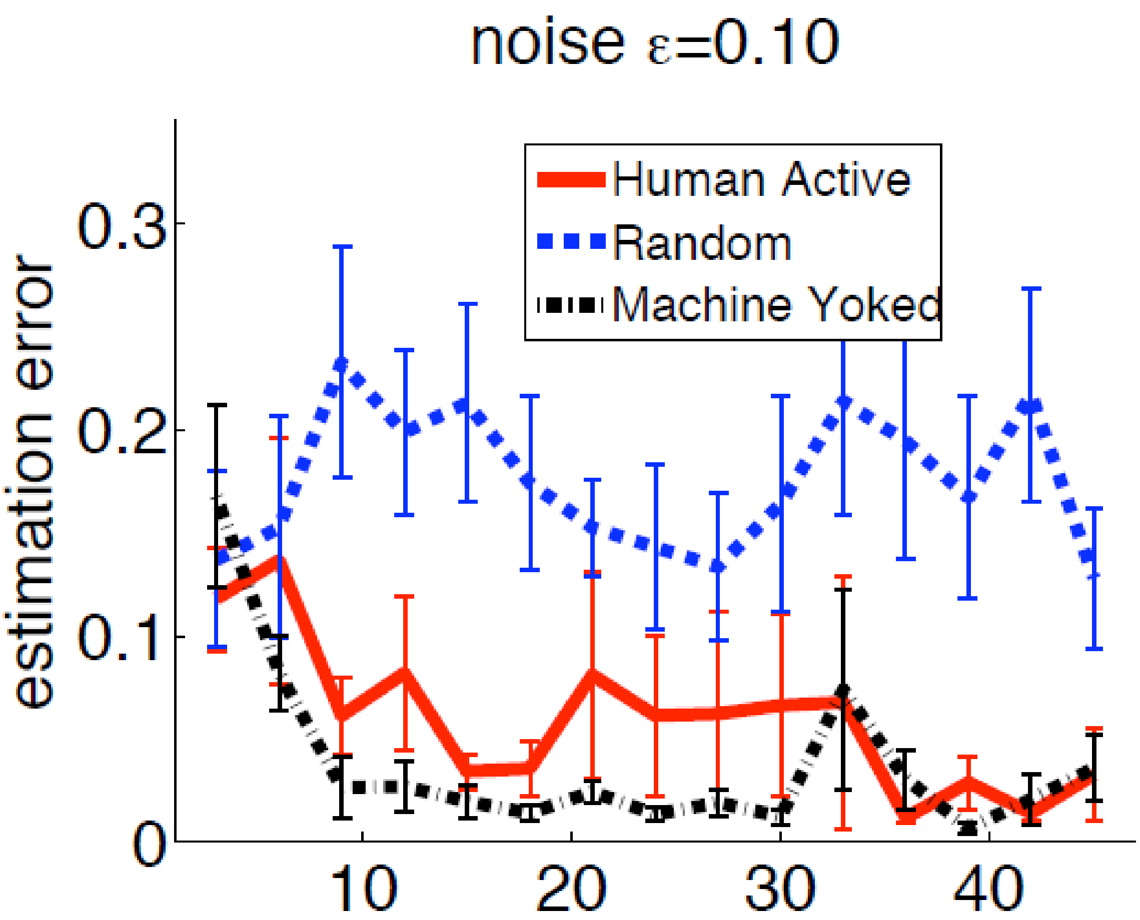
Conditions

- 33 participants assigned: 13, 14, 6 to the three conditions
- Short practice session followed by 5 x 45 iterations
- ε = 0, 0.05, 0.1, 0.2, 0.4, random order
- θ in [1/16, 15/16]
- Participants asked to guess θ after every 3 iterations
- Compute mean $| \theta_n - \theta |$

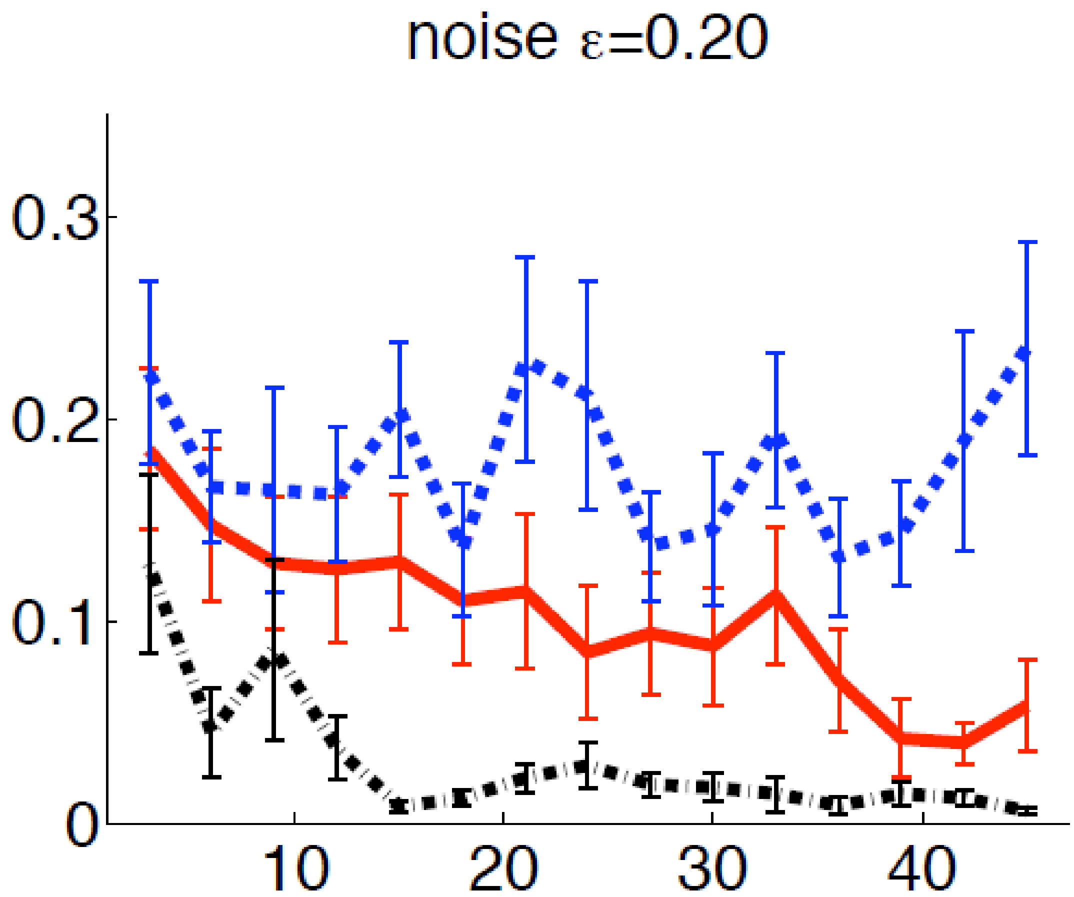
Q1. Do humans perform better when they can actively select samples for labeling?

- Yes, at least for low noise levels. At higher noise levels, the performance is similar.
- Human estimation error is smaller in Human-Active than in Random
  - Very significant at low noise
  - Deteriorates and becomes similar in performance at high noise levels

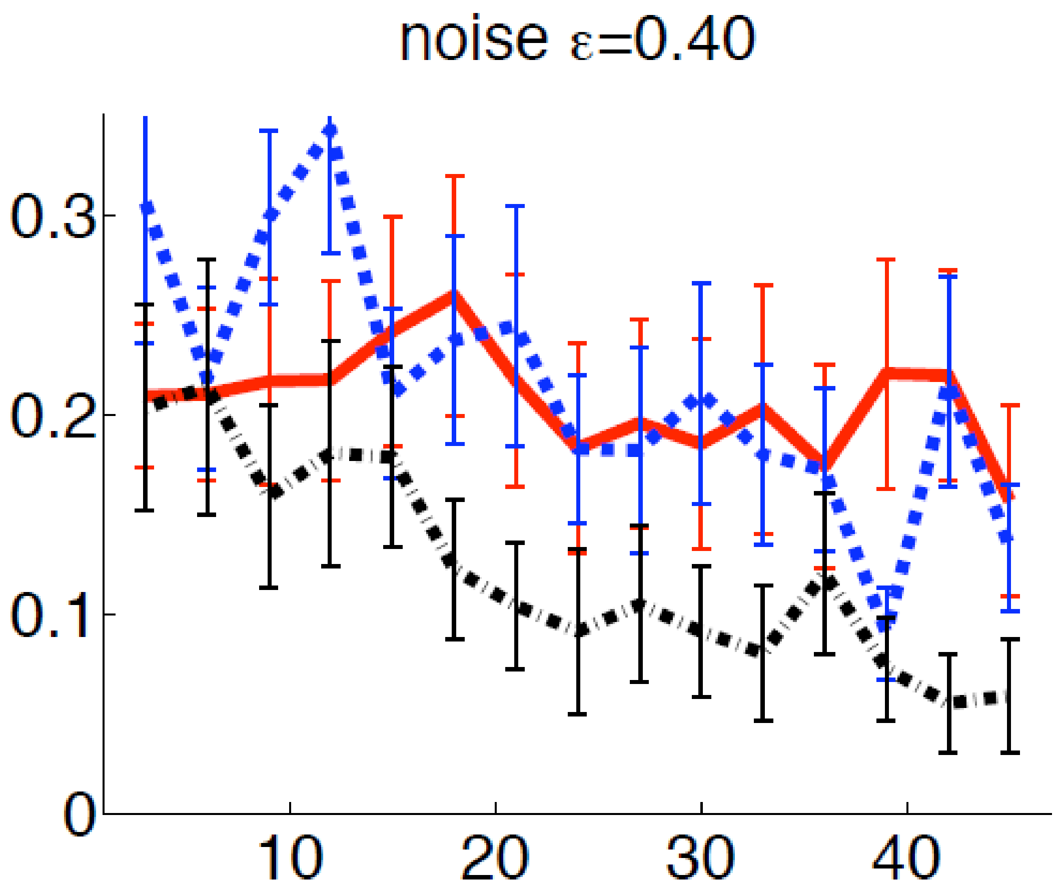
Error trends for ε = 0.10



Error trends for ε = 0.20



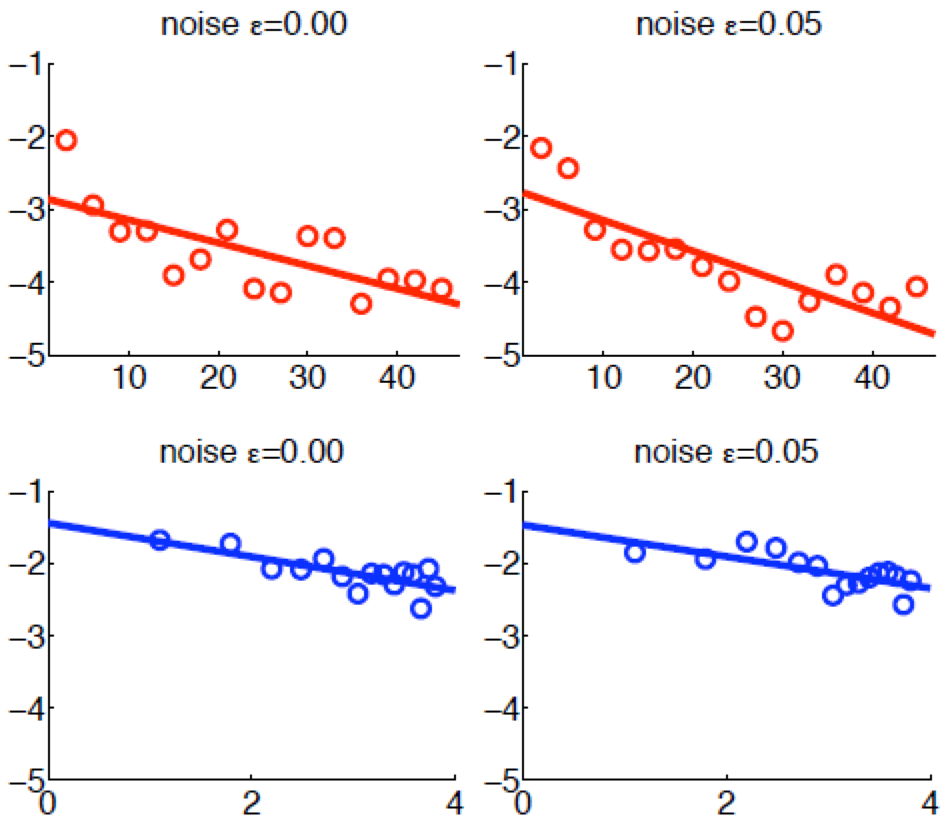
Error trends for ε = 0.40



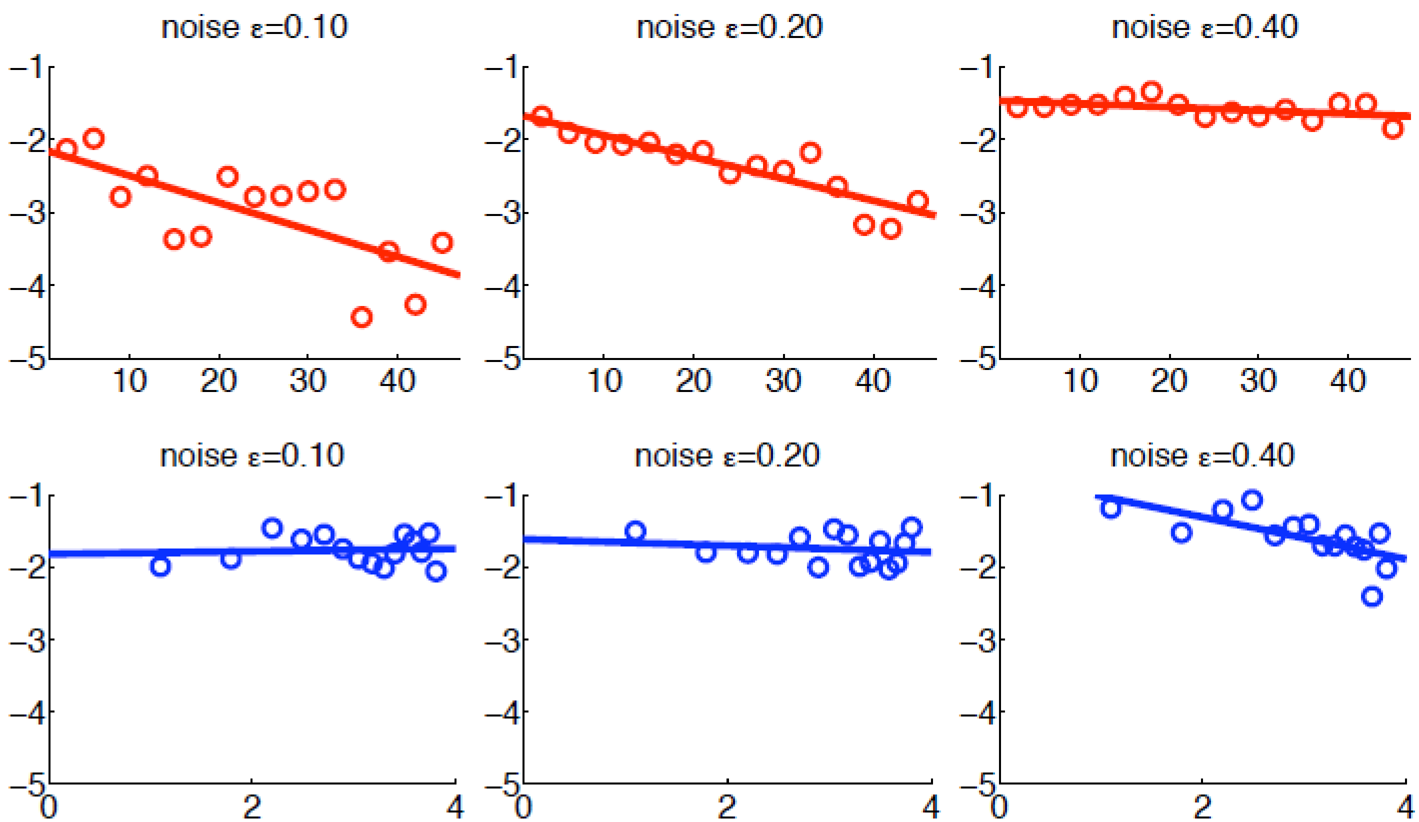
Q2. Can humans achieve the full benefit of active learning?

- Human active learning does have exponential convergence
  - Slower decay constants
- Human passive learning
  - Occasionally does not achieve even polynomial convergence
  - Does not approach optimal performance

Rate of error decrease (low noise)



Rate of error decrease (high noise)



Analysis of error decrease

	$\epsilon = 0$	0.05	0.1	0.2	0.4
Human-Active	0.031	0.042	0.037	0.030	0.005
bound (2)	0.347	0.166	0.112	0.053	0.005

Table 1: The exponential decay constants of human active learning is slower than predicted by statistical learning theory for lower noise levels.

Q3. Can machine learning be used to enhance human learning?

- Looks like it at high noise levels
- Machine-Yoked is similar to Human-Active in low noise but a lot better at high noise

Human estimate error

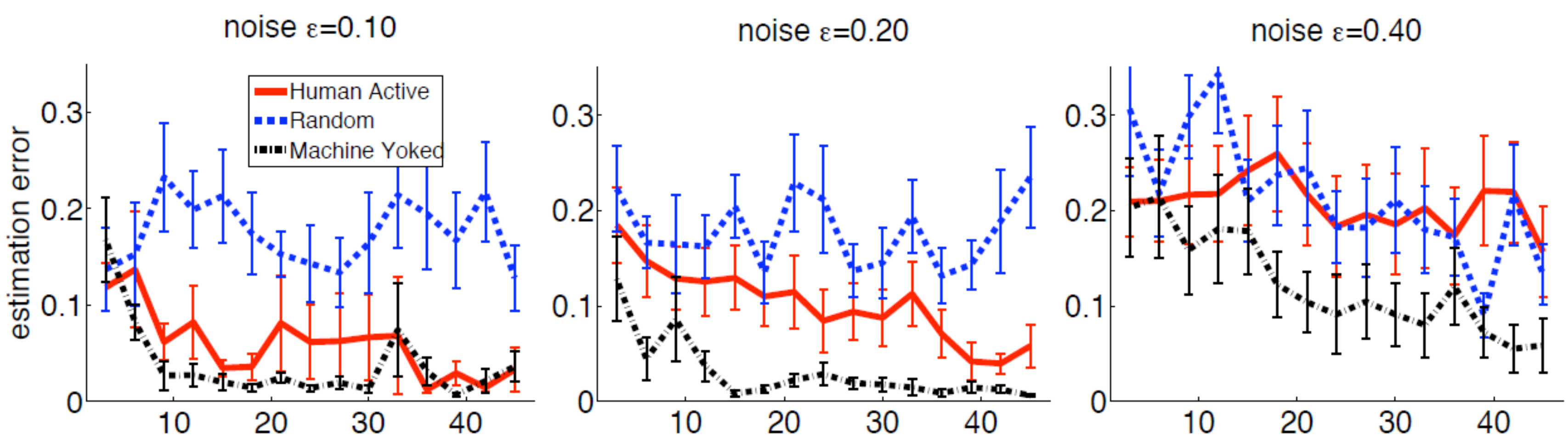


Figure 5: Human estimate error $|\hat{\theta}_n - \theta|$ under different conditions and noise levels. The x-axis is iteration n. The error bars are ± 1 standard error. Human-Active is better than Random when noise is low; Machine-Yoked is better than Human-Active when noise is high.

Q3. Can machine learning be used to enhance human learning?

- Upon inspection:
  - Subjects noticed that the computer was generating examples that converge to the true boundary
  - Simply placed their guess near the last training example
  - They are probably not actually "learning"
- Inconclusive!

Q4. Do the above answers depend on the difficulty of the task?

- Noise level affects human learning significantly
- At high noise the advantage of active learning over passive learning seems to disappear

Revisit our wishlist

- Consistency:
  - Holds except for a few cases where the slope is almost horizontal
- Fallback guarantee:
  - Holds, active learning's advantage may diminish or disappear but it never becomes worse
- Rate improvement
  - Seems to be only true at low noise levels

Conclusions

- Humans are able to actively select queries and use them to learn faster
  - Ability to do this diminishes with high noise
  - Do not approach theoretic bounds
- Passive learning alone is not a good model for human learning
- The task is not especially natural
  - Perhaps we will obtain different results for a task which is more intuitive and where people have more experience

My comments

- Interesting premise and experiment
- Very small sample size (only 33)
  - Are the results reproducible?
- One or two people performing particularly badly affected the graph a lot
- At high noise, still exponential advantage from active learning, but graphs are really similar
- General trend is believable
- Comment about the failure to learn at $\varepsilon = 0.10$ and 0.20 but not 0.40 is insufficiently supported
- Seems like the differential of the decay constant is smaller for higher noise

More comments

- When extrapolating linear relationships, would have been nice if R^2 values were provided
  - A few of them don't seem to fit well at all
- A side idea about the Machine-Yoked
  - "Memorizing" strategy by the human
  - Perhaps we could generate the labeled examples using active learning, but provide them to the human in random order
  - If people are memorizing, then it would greatly affect convergence in later rounds

Thanks! Questions?