John Lipor and Laura Balzano, University of Michigan
lipor@umich.edu and girasole@umich.edu

Quantile Search for Minimum-Time Sampling
Active Learning

Images courtesy nih.gov, ras.org.uk, Center for Embedded Networked Sensing, wikipedia
Passive Learning

Scientist / Expert

Models / Hypotheses

Data / Experiment

Conclusion
Active Learning

Scientist / Expert

Models / Hypotheses

Data / Experiment

Conclusion
Active Learning

Scientist / Expert

Models / Hypotheses

Conclusion

Data / Experiment
Active Learning

We have a big lake (the Great Lakes specifically)

Inter-sample spacing matters

Problem formulation

Deterministic Quantile Search

Probabilistic Quantile Search
Active Classifier Learning

We seek a two class linear classifier.
We seek a two class linear classifier.

Suppose we are given a small number of labels.

The idea of active learning is that we may request further labels.
Active Classifier Learning

We seek a two class linear classifier.

Suppose we are given a small number of labels, and we can consider the remaining possible classifiers.

Where would we request a label?
“Is the person male or female?”

“Is she wearing a hat?”

We try as best as we can to cut the remaining possibilities in half.
The same principle applies here: We may choose the query that *throws away half* of the remaining possible linear classifiers.
Learn a 1D transition.

Figure from “Active Learning for Adaptive Mobile Sensing Networks” by Singh, Nowak, and Ramanathan. IPSN 2006

Learn a 2D boundary fragment.

Figure from “Active Learning and Sampling” by Castro and Nowak in Foundations and Applications of Sensor Management, 2008.
Active Function Estimation

Estimate a function.

Figure from “Backcasting: Adaptive Sampling for Sensor Networks” by Willett, Martin, and Nowak. IPSN 2004
Active Learning

We have a big lake (the Great Lakes specifically)
  Inter-sample spacing matters

Problem formulation

Deterministic Quantile Search

Probabilistic Quantile Search
Sampling the Great Lakes

Source: NYTimes 2012
Sampling the Great Lakes
The hypoxic region is at the lake bottom, where oxygen is scarce. We seek to estimate the spatial extent.

In black we see the largest observed hypoxic zone of Erie for a given year.

Estimates based on ~4 sampling cruises per year, each giving ~360 samples. Estimated ~2500 pixel values.

The sampling system

Images courtesy Branko Kerkez, senseplatypus.com, sontek.com, arduino.cc, forbes.com, amazon.com, arc-ts.umich.edu
Active Learning

We have a big lake (the Great Lakes specifically)

- Inter-sample spacing matters

Problem formulation

Deterministic Quantile Search

Probabilistic Quantile Search
Our function comes from the class of step functions:

\[ \mathcal{F} = \{ f : [0, 1] \rightarrow \mathbb{R} | f(x) = 1_{[0, \theta)}(x) =: f_\theta(x) \} \].

We wish to recover \( \theta \) from \( n \) noisy measurements \( Y_i, \ i = 1, \ldots, n \) where

\[ Y_i = \begin{cases} 
 f_\theta(X_i) & \text{with probability } 1 - p \\
 1 - f_\theta(X_i) & \text{with probability } p 
\end{cases} \]

and the \( X_i \) can be chosen based on previous pairs \( (X_j, Y_j), \ j = 1, \ldots, i - 1 \).

Denote our estimate of \( \theta \) after \( n \) samples as \( \hat{\theta}_n \). We consider either the worst-case or expected error,

\[
\sup_{\theta \in [0,1]} \left| \hat{\theta}_n - \theta \right| \quad \text{or} \quad \mathbb{E} \left[ \left| \hat{\theta}_n - \theta \right| \right].
\]
Sample Complexity $p=0$
Optimal Sample Complexity

The best adaptive sample placement uses binary bisection (see Castro and Nowak 2008), and the worst case error is bounded by

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

The expected error given a uniform distribution on $\theta$ will be

$$\mathbb{E} \left[ \left| \hat{\theta}_n - \theta \right| \right] = \frac{1}{2^{n+2}}$$

<table>
<thead>
<tr>
<th></th>
<th>worst case</th>
<th>expected (with uniform prior on $\theta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>adaptive (binary bisection optimal)</td>
<td>$\sup_{\theta \in [0,1]} \left</td>
<td>\hat{\theta}_n - \theta \right</td>
</tr>
<tr>
<td>non-adaptive (uniform grid optimal)</td>
<td>$\sup_{\theta \in [0,1]} \left</td>
<td>\hat{\theta}_n - \theta \right</td>
</tr>
</tbody>
</table>
... but worst case distance

| adaptive sample complexity (binary bisection optimal) | worst case \( \sup_{\theta \in [0,1]} |\hat{\theta}_n - \theta| \) | expected (with uniform prior on \( \theta \)) \( \mathbb{E} \left[ |\hat{\theta}_n - \theta| \right] \) |
|-------------------------------------------------------|-----------------------------------------------|-----------------------------------------------|
| \( \leq \frac{1}{2^{n+1}} \)                          | \( = \frac{1}{2^{n+2}} \)                     |                                               |

| adaptive distance traveled (fix desired error \( \epsilon \), start at \( X_i = 0 \)) | \( = 1 - \epsilon \) | \( = 1 - 2\epsilon \) |
Binary search is nothing other than taking a sample at the 2-quantile of the posterior distribution for $\theta$.

Quantile search generalizes this by instead taking a sample at the first $m$-quantile of the posterior distribution for $\theta$. 
Quantile search generalizes this by instead taking a sample at the first \( m \)-quantile of the posterior distribution for \( \theta \).

E.g. \( m=3 \):

Choosing \( m \) allows for a tradeoff between number of samples and distance traveled.
## Quantile Search Cost

<table>
<thead>
<tr>
<th></th>
<th>worst case</th>
<th>expected</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\sup_{\theta \in [0,1]}</td>
<td>\hat{\theta}_n - \theta</td>
</tr>
<tr>
<td>bisection sample complexity</td>
<td>$\leq \frac{1}{2^{n+1}}$</td>
<td>$= \frac{1}{2^{n+2}}$</td>
</tr>
<tr>
<td>$m$-quantile search sample complexity</td>
<td>$\rho = \frac{m-1}{m}$</td>
<td>$\leq \frac{1}{2} \rho^n$</td>
</tr>
<tr>
<td>bisection distance traveled (as desired error $\epsilon \to 0$)</td>
<td>$= 1$</td>
<td>$= 1$</td>
</tr>
<tr>
<td>$m$-quantile search distance traveled (as desired error $\epsilon \to 0$)</td>
<td>$= 1$</td>
<td>$= \frac{m}{2m-2}$</td>
</tr>
</tbody>
</table>
Goal: Minimize the total sampling time subject to a given reconstruction error.

\[
\text{min} \quad T = \gamma N + \eta D \\
\text{subject to} \quad |\theta - \hat{\theta}_n| \leq \epsilon
\]

\[
\begin{align*}
\gamma &= \text{time/sample} \\
N &= \text{number of samples} \\
\eta &= \text{time/distance} \\
D &= \text{distance traveled}
\end{align*}
\]
Let $N$ be a random variable denoting the samples required to achieve an error $\varepsilon$, and $D$ the distance. Rearranging the expected sample complexity, we have

$$\mathbb{E}[N] = \frac{\log(4\varepsilon)}{\log \left( \left( \frac{m-1}{m} \right)^2 + \frac{1}{m^2} \right)} \equiv n'$$

Denote the sampling time $T$. Then

$$\mathbb{E}[T] = \gamma \mathbb{E}[N] + \eta \mathbb{E}[D]$$

$$= \gamma n' + \eta \left( \frac{m}{2m-2} - \frac{1}{(2m-2)(2m-1)^n'} \right)$$

$$\approx \gamma n' + \eta \frac{m}{2m-2}$$
(Top left) Simulated and theoretical expected error after 20 samples as we vary $m$,

(Top right) Simulated and theoretical distance traveled as we vary $m$, and

(Bottom left) optimal $m$ for $\gamma = 60 \frac{s}{samp}$ and $\eta = \frac{1}{4} \frac{s}{m}$. 

$\hat{\theta}_{20} - \theta$

$\|m\| - \|m\|$

average distance

$\sup_{\|m\| \in [0,1]} b_{\|m\|}$

expected sampling time (s)

$E_{b_{\|m\|}}{\|m\|}$

Theoretical

Simulated
To simulate sampling all of Lake Erie, we split it first in half and then into 16 strips and perform DQS. In most cases we can sample the entire boundary in 2-3 days; fast enough to assume a stationary hypoxic zone.

(Bottom left) optimal $m$ for $\gamma = 60 \frac{s}{samp}$ and $\eta = \frac{1}{4} \frac{s}{m}$. 
To simulate sampling all of Lake Erie, we split it first in half and then into 16 strips and perform DQS. In most cases we can sample the entire boundary in 2-3 days; fast enough to assume a stationary hypoxic zone.

(Bottom left) optimal $m$ for $\gamma = 60 \frac{s}{samp}$ and $\eta = \frac{1}{4} \frac{s}{m}$.
Active Learning

We have a big lake (the Great Lakes specifically)

- Inter-sample spacing matters

Problem formulation

Deterministic Quantile Search

Probabilistic Quantile Search
Our function comes from the class of step functions:

$$\mathcal{F} = \{ f : [0, 1] \rightarrow \mathbb{R} | f(x) = 1_{[0, \theta]}(x) =: f_{\theta}(x) \}.$$ 

We wish to recover \( \theta \) from \( n \) noisy measurements \( Y_i, i = 1, \ldots, n \) where

$$Y_i = \begin{cases} 
  f_{\theta}(X_i) & \text{with probability } 1 - p \\
  1 - f_{\theta}(X_i) & \text{with probability } p 
\end{cases}$$

and the \( X_i \) can be chosen based on previous pairs \((X_j, Y_j), j = 1, \ldots, i - 1\).

You may think of this noise model as modeling a detector where the probability of false alarm or false detection are the same, \( e.g. \) detection after additive Gaussian noise.
Noisy case $p > 0$

If we were to update our posterior according to the measurements without taking the noise probability into consideration, we’d get lost:
Rather than sampling at $1/m$ into the feasible interval, we sample $1/m$ into the posterior distribution on $\theta$ (See Burnashev and Zigangirov, “An interval estimation problem for controlled observations,” Problems in Information Transmission, 1974 for special case bisection algorithm).

**Algorithm 1** Probabilistic Quantile Search (PQS)

1: initialize prior density $\pi_0(\theta) = 1$ for $\theta \in [0, 1]$
2: while not converged do
3: choose $X_n$ such that $\int_0^{X_n} \pi_n(x)dx = 1/m$
4: $Y_n \leftarrow f(X_n)$
5: perform Bayesian update to obtain $\pi_{n+1}(x)$
6: end while
7: return $\hat{\theta}_n$ such that $\int_0^{\hat{\theta}_n} \pi_{n+1}(x)dx = 1/2$
Algorithm 1 Probabilistic Quantile Search (PQS)

1: initialize prior density $\pi_0(\theta) = 1$ for $\theta \in [0, 1]$

2: while not converged do

3: choose $X_n$ such that $\int_0^{X_n} \pi_n(x) dx = 1/m$

4: $Y_n \leftarrow f(X_n)$

5: Perform Bayesian update to obtain $\pi_{n+1}(x)$:

6: if $Y_n = 0$ then

7: $\pi_{n+1}(x) = \begin{cases} 
(1 - p) \left( \frac{m}{1+(m-2)p} \right) \pi_n(x) & x \leq 1/m \\
 p \left( \frac{m}{1+(m-2)p} \right) \pi_n(x) & x > 1/m 
\end{cases}$

8: else

9: $\pi_{n+1}(x) = \begin{cases} 
 p \left( \frac{m}{1+(m-2)p} \right) \pi_n(x) & x \leq 1/m \\
(1 - p) \left( \frac{m}{1+(m-2)p} \right) \pi_n(x) & x > 1/m 
\end{cases}$

10: end if

11: end while

12: return $\hat{\theta}_n$ such that $\int_0^{\hat{\theta}_n} \pi_{n+1}(x) dx = 1/2$
Rather than sampling at $1/m$ into the feasible interval, we sample $1/m$ into the posterior distribution on $\theta$.

For $m=3$, $p=0.2$: 
The discretized PQS algorithm for $m \geq 2$ satisfies

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

This bound matches Castro and Nowak for $m=2$. In this case, comparing to $O(1/n)$ error bound given by grid sampling, we have a $(1/2)^{n/2}$ bound.

It’s loose in practice, but finding the expected number of samples and any distance bounds or expectations is significantly more technical, because of the less obvious interaction of sample locations with the posterior in PQS.
Future: Hölder smooth

To generalize we could consider a more general class of functions, Hölder smooth functions, and estimate the level set at \( \gamma \). A function is \( \alpha \)-Hölder smooth for \( \delta > 0 \) so that \( \forall x : |f(x) - \gamma| \leq \delta \) we have:

\[
|f(x) - \gamma| \geq c|x - \theta^*|^{\alpha}.
\]

This is wrong: the idea is that we have something that is unsmooth at the transition.

\[\alpha = 0\]

\[\alpha = 1\]

Figures courtesy Rui Castro

Results on sample complexity only are known in the case of binary bisection and for d-dimensional signals that have the additional assumption of smooth level sets.
Thank you!

Questions?