Active Learning vs. Compressed Sensing

Robert Nowak
University of Wisconsin-Madison

Rui Castro  Jarvis Haupt  Rebecca Willett
Active Learning
Adaptively sense based on information gleaned from previous samples (feedback-driven sensing)

Compressed Sensing
Non-traditional samples in form of non-adaptive randomized projections (Emmanuel Candes’ talk)

both adaptive and compressive sampling are examples of Integrated Sensing and Processing since tasks of data acquisition and signal recovery are intimately intertwined

both adaptive and compressive sampling can significantly outperform traditional nyquist sampling schemes

I will first focus on adaptive sampling, and then discuss and compare with compressive sampling at the end of the talk
Adaptive Sampling Example: Laser Scanning

Goal: Image the landscape accurately as fast as possible by strategically scanning region of interest.
Environmental Monitoring with Wireless Sensor Networks

Goal: Reconstruct an accurate map of contamination by activating/querying as few sensors as possible
National Ecological Observatory Network
"What" vs. "Where" Information

"What is the value of the function at point $x$?"

"Where does the function abruptly change?"

These are two fundamentally different problems, from the perspective of sampling and learning. Classical statistical methods are primarily concerned with "what" information.

- **What** is the density of oil inside the spill? (averaging)
- **Where** is the boundary of the spill? (deciding)
Passive vs. Adaptive Sampling

**Passive Sampling**
- Sample uniformly at random (or deterministically)

**Adaptive Sampling**
- Sequentially sample using information gleaned from previous samples
Active Learning

1. **Adaptive Sampling** - Sampling locations are adaptively chosen based on past locations and responses (observations)

   - Burnashev & Zigangirov ‘74
   - Korostelev ‘99
   - Hall & Molchanov ‘03
   - Golubev & Levit ‘03
   - Castro, Willett, & RN ‘05

2. **Selective Sensing** - Sampling at random locations, but decision to obtain response/label is optional

   - Freund, Seung, Shamir, & Tishby ‘97
   - Dasgupta, Kalai, & Moteleoni ‘05
Notation

Sample locations:

\[ X_1, X_2, \ldots \]

Responses/labels:

\[ Y_1, Y_2, \ldots \]

Dependence:

Rather than observing i.i.d. pairs \((X_i, Y_i)\), sample locations and/or responses depend on past observations.
Problem Formulation

Passive Sampling:

Sample locations: $X_i \in [0, 1]^d$ are independent of $\{Y_j\}_{j \neq i}$. These do not depend in any way on $f$.

Adaptive Sampling:

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

$$X_i|(X_{i-1}, Y_{i-1}), \ldots, (X_1, Y_1)$$
Adaptive Sampling in One Dimension

Passive sampling learning rate (polynomial):

\[ E[\|f - \hat{f}\|^2] \preceq n^{-1} \]

Adaptive sampling learning rate (exponential):

\[ E[\|f - \hat{f}\|^2] \preceq e^{-c_0 n} \]
Adaptive Sampling in One Dimension

$F$ is the class of "step" functions of the form above

**Goal:** Design an estimator $\hat{\theta}_n$ to minimize

$$E|\hat{\theta}_n - \theta^*|$$
Passive Sampling in Noiseless Conditions

\[ |\hat{\theta}_n - \theta^*| \sim \frac{1}{n} \]
Adaptive Sampling in Noiseless Conditions

This is a coding problem

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

What if there is noise???
Passive Sampling in Noise

Usual parametric estimation rate:

\[ \mathbb{E}[|\hat{\theta}_n - \theta^*|] \lesssim \frac{1}{n} \]
Burnashev & Zigangirov '74 proposed a scheme and proved

\[ e^{-c_0 n} \leq \mathbb{E}[|\hat{\theta}_n - \theta^*|] \leq e^{-c_1 n} \]
A Probabilistic Bisection

BZ Method: Burnashev & Zigangirov '74

- Uniform prior

\[ q(\theta) = \text{uniform}[0, 1] \]

- Take sample at median

\[ x = 1/2 \rightarrow \text{measurement } y = 0 \text{ or } 1 \]

- Update posterior

\[ q(\theta|y) \propto \text{Prob}(y|\theta) \times q(\theta) \]

- Take new sample at median ("bisect" uncertainty)

\[ x_{\text{new}} = \text{median}(q(\theta|y)) \]

REPEAT

After n samples:

\[ \hat{\theta}_n = \arg \max_{\theta} q(\theta|y) \]
Adaptive Sampling in Noise

Sample at median at each step

$q_0(\theta)$
Upper bound: Motivated by coding scheme of Horstein ‘63, Burnashev & Zigangirov ‘74 showed that

\[ e^{-c_0 n} \leq \mathbb{E}[|\hat{\theta}_n - \theta|] \leq e^{-c_1 n} \]

(lower bound follows from channel coding argument)
Equivalent Communication System

Binary Symmetric Channel

Error-free transmission of $n$-bit message is equivalent to determination of $\theta \in (0, 1)$ to within $2^{-n}$

Shannon '48:

Rate < Capacity $\Rightarrow P_{err} \rightarrow 0$, exponentially
Rate > Capacity $\Rightarrow P_{err} \rightarrow 1$, exponentially

Exponential error decay is a fundamental limit
Piecewise Constant Functions, $\dim > 1$

We observe

$$Y_i = f(X_i) + W_i$$

where $f \in \mathcal{F}$, $X_i \in [0, 1]^2$

and $W_i \overset{i.i.d.}{\sim}$ zero-mean distribution
Passive Sampling and Wavelet Denoising

- Sample uniformly over domain of function

- Recursively divide the domain into hypercubes, and prune to adapt to the data

- Average samples in each cell of pruned partition to form final estimate
Passive Sampling and Denoising in Action

“true” image

noisy data

pruned partition

estimate

\[ \mathbb{E}[\|\hat{f}_n - f\|^2] \asymp n^{-\frac{1}{d}} \]

wavelets achieve best possible rate

\[ \inf_{\hat{f}_n} \sup_{f \in \mathcal{F}} \mathbb{E}[\|\hat{f}_n - f\|^2] \asymp n^{-\frac{1}{d}} \]
Can Adaptive Sampling Do Better?

**Boundary Fragments** (Korostelev and Tsybakov ‘93)

\[ x_d = g(x_1, \ldots, x_{d-1}) \]

**known constant values**
Adaptive Sampling of Boundary Fragments

Reduction to series of 1-d changepoint problems (Korostelev ‘99)

Approximate Lipschitz function with $m$ const pieces:

$$\|f_m - f\|^2 \lesssim m^{-\frac{1}{d-1}}$$

Estimation error of constant using log $n$ adaptive samples:

$$E|\hat{\theta} - \theta| \leq e^{-c_1 \log n} \leq n^{-1}$$

Use $n$ samples, $m = n/\log n$

$$\sup_{f \in \mathcal{F}} \mathbb{E}[\|\hat{f}_n - f\|^2] \asymp \left(\frac{n}{\log n}\right)^{-\frac{1}{d-1}}$$
Minimax Lower Bounds for Adaptive Sampling

- Communication analogy and Shannon capacity imply error rate cannot be improved
- Lipschitz regularity can be exploited to remove log factor

\[ \inf \sup_{\hat{f}_n, f \in F} E[\|\hat{f}_n - f\|^2] \asymp n^{-\frac{1}{d-1}} \]

compare with exp 1/d in passive case
Limitations of Boundary Fragment Model

“functional” boundary

general piecewise function

More general boundaries are not functions and the “strip” method of reduction to series of 1-d problems is not possible

General case calls for a completely different approach !!!
Multiscale Adaptive Approach

Stage 1: “Oversample” at coarse resolution

- n/2 samples uniformly distributed
- many more samples than cells
- prune partition according to standard multiscale methods
- biased, but very low variance result

“boundary zone” is reliably detected
Example: Piecewise Smooth Function

**Preview**
Partition superimposed on noisy observations

**ACTIVE**
Adaptive sampling estimate based on 17,536 samples; MSE = 0.0013

**PASSIVE**
Estimate based on 65,636 uniform samples; MSE = 0.0008

**Preview estimate**
MSE = 0.0075
Main Theorem (R. Castro, R. Willett, RN ’05)

Let $f$ be a piecewise constant function whose boundaries separating constant regions is cusp-free. Then

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

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

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

* Cusp-free boundaries cannot behave like the graph of $|x|^{1/2}$ at the origin, but milder “kinks” like $|x|$ at 0 are allowable. Boundary Fragment class is cusp-free.

Compare with passive rate $\exp 1/d$ and with minimax rate $\exp 1/(d-1)$
**Adaptive vs. Passive Sampling**

**Smooth Functions**

Learning Rates:
Passive = Adaptive
\[ n^{-2\alpha/(2\alpha+d)} \]

**Piecewise Smooth**

Learning Rates:
Passive
\[ n^{-1/d} \]
Adaptive
\[ n^{-1/(d-1)} \]
Compressive Sampling

Non-traditional samples in form of non-adaptive randomized projections (Emmanuel Candes' talk)

\[ Y_i = \left\langle \begin{bmatrix} f_1^* \\ f_2^* \\ \vdots \\ f_n^* \end{bmatrix}, \begin{bmatrix} \pm \frac{1}{\sqrt{n}} \\ \vdots \\ \pm \frac{1}{\sqrt{n}} \end{bmatrix} \right\rangle + \text{noise}(\sigma^2) \]

n-point signal random vector
A Simple Example

Suppose that \( f^* \in \mathbb{R}^n \) has a single non-zero entry that is strictly greater than zero. How many random projections are required to perfectly reconstruct \( f^* \)?

If \( \phi' f^* > 0 \), then non-zero element is located at one of the \( +1 \) locations in \( \phi \), otherwise it must be at one of the \( -1 \) locations. Repeat.
A Complicated Example: MRI

Compressive sampling theory suggests that accurate reconstructions are possible from vastly undersampled Fourier data.
Performance of Compressive Sampling

Suppose that we take $k$ random projection samples

$$Y_i = \langle f^*, \phi_i \rangle + \text{noise}(\sigma^2), \quad i = 1, \ldots, k$$

And that $f^*$ is compressible, in the sense that for each $m \geq 1$ there exists an $m$-term approximation $f_m$ (in some basis) satisfying

$$\|f^* - f_m\|^2 \leq m^{-\alpha}, \quad \alpha > 1$$

Then we can compute a reconstruction $\hat{f}$ from these data satisfying:

$$\sigma^2 = 0 : \quad E \left[ \|f^* - \hat{f}\|^2 \right] \leq k^{-\alpha}$$

$$\sigma^2 > 0 : \quad E \left[ \|f^* - \hat{f}\|^2 \right] \leq k^{-\alpha/(\alpha+1)}$$

Candes, Romberg, Tao '04
Donoho '04

Haupt & RN '05,
E. Candes & T. Tao '05
Adaptive vs. Compressive Sampling

Noiseless CS:

$$\sigma^2 = 0 : \ E \left[ \| f^* - \hat{f} \|^2 \right] \leq k^{-\alpha}$$

Noisy CS:

$$\sigma^2 > 0 : \ E \left[ \| f^* - \hat{f} \|^2 \right] \leq k^{-\alpha/(\alpha+1)}$$

CS is truly optimal in noiseless situations, but what about noisy cases?

How does performance degrade with noise?

How does CS compare to Adaptive Sampling in noise?
Adaptive vs. Compressive Sampling

Compare adaptive sampling (AS) and compressive sampling (CS) for recovery of step functions, where $\theta^* \in \{1, \ldots, n\}$

$$f_{\theta^*} = \begin{bmatrix} -1 \\ \vdots \\ -1 \\ +1 \\ \vdots \\ +1 \end{bmatrix}$$

change @

signal

vector representation
Adaptive vs. Compressive Sampling

**AS:**

\[ Y_i = \left\langle \begin{bmatrix} -1 \\ \\
\vdots \\
-1 \\
+1 \\
+1 \end{bmatrix} , \begin{bmatrix} 0 \\
\vdots \\
0 \\
1 \\
0 \end{bmatrix} \right\rangle + \text{noise}(\sigma^2) \]

Sample at adaptively chosen location

**CS:**

\[ Y_i = \left\langle \begin{bmatrix} -1 \\ \\
\vdots \\
-1 \\
+1 \\
+1 \end{bmatrix} , \begin{bmatrix} \pm \frac{1}{\sqrt{n}} \\
\vdots \\
\pm \frac{1}{\sqrt{n}} \\
\pm \frac{1}{\sqrt{n}} \end{bmatrix} \right\rangle + \text{noise}(\sigma^2) \]

Non-adaptive, unit norm, random projection
Bayesian Recovery Strategy

Initialize:

\[ i = 0; \quad q_0(\theta) = \text{uniform}[0, 1] \]

Sample:

\[ \phi_i = \text{adaptive or non-adaptive point sample or random projection} \]

\[ Y_i = \langle f^*, \phi_i \rangle + \text{noise} \]

Update Posterior (Bayes Rule):

\[ q_{i+1}(\theta) \propto \Pr(Y_i|\theta) q_i(\theta) \]

\[ \tilde{\theta}_k = \arg\max_\theta q_k(\theta) \]
Analysis of Bayesian Recovery Strategy

\[ \hat{\theta}_k = \arg \max_{\theta} q_k(\theta) \]

Using similar analysis approach to that of Burnashev & Zigangirov ‘74 one can bound the probability

\[ \Pr\left(\hat{\theta}_k \neq \theta^*\right) \]

for each of the following sampling schemes:

- PS : passive (non-adaptive) point sampling
- CS : compressive sampling with white noise projections
- AS : adaptive point sampling
Error Bounds (R. Castro & RN ’05)

\[ Pr \left( \hat{\theta}_k \neq \theta \right) \leq \]

\[ PS : n \left( 1 - \frac{1}{n} \left( 1 - e^{-\frac{1}{2\sigma^2}} \right) \right)^k \]

\[ CS : n \max \left\{ \left( \frac{1}{2} + \frac{1}{2} e^{-\frac{1}{2n\sigma^2}} \right)^k , e^{-\frac{k}{2n\sigma^2}} \right\} \]

\[ AS : n \left( \frac{1}{2} + \frac{1}{2} e^{-\frac{1}{2\sigma^2}} \right)^k \]

\[ Pr \left( \hat{\theta}_k \neq \theta \right) \leq n \left[ \alpha(n, \sigma^2) \right]^k \]

\[ AS : \alpha(n, \sigma^2) = \alpha(\sigma^2) \]

\[ PS : \alpha(n, \sigma^2) \text{ depends on } n \]

\[ CS : \alpha(n, \sigma^2) \text{ dependency on } n \text{ negligible at high SNR} \]
High SNR Regime

As $\sigma^2 \to 0$

\[ PS : \quad n \left(1 - \frac{1}{n}\right)^k \]

\[ CS : \quad n \ 2^{-k} \]

\[ AS : \quad n \ 2^{-k} \]

AS and CS bounds are equivalent in low-noise limit (noise-free case), and significantly better than PS.
Low SNR Regime

As $\sigma^2 \to \infty \Rightarrow e^{-\frac{1}{2\sigma^2}} \approx -\frac{1}{2\sigma^2}$

$PS : n \left(1 - \frac{1}{2n\sigma^2}\right)^k$

$CS : n \left(1 - \frac{1}{2n\sigma^2}\right)^k$

$AS : n \left(1 - \frac{1}{4\sigma^2}\right)^k$

PS and CS bounds are equivalent in low SNR limit, and significantly worse than AS
Adaptive vs. Passive vs. Compressive Sampling

\[ \Pr(\hat{\theta}_k \neq \theta) \]

\( n = 1000, k = 30 \) samples

PS is bias limited to accuracy on the order of \( k/n \)

AS and CS are not!
Ex. Boundary Fragment Reconstruction

original
noisy

(a) Conventional pixel sampling
   $k=4096$, $MSE=97.22$

(b) Adaptive pixel sampling
   $k=4096$, $MSE = 29.55$

(c) Compressive sampling
   $k=4096$, $MSE = 30.48$

(d) Compressive sampling
   $k=1024$, $MSE = 103.01$
Conclusions

For piecewise constant/smooth functions

Adaptive sampling (special-purpose) > Compressive sampling (universal) > Passive (point) sampling (universal)
Spatially adaptive estimators based on “sparse” model selection (e.g., wavelet thresholding) may provide automatic mechanisms for guiding active learning processes.

Spatially adaptive (nonlinear) estimators and spatially adaptive (nonlinear) sampling seem to naturally go hand in hand.

Can active learning work in even more realistic situations and under little or no prior assumptions?