Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz
(Labeled) Data Collection Story

First generation data analytics ignored the control over (big) data collection
  • Learning models from given (passively collected/labeled) data
  • Inference on arbitrary instances
(Labeled) Data Collection Story

First generation data analytics ignored the control over (big) data collection

- Learning models from given (passively collected/labeled) data
- Inference on arbitrary instances

In many practical/engineering settings some control over data collection

- Sensor management (run time)
- Data collection/labeling (training)
(Labeled) Data Collection Story

First generation data analytics ignored the control over (big) data collection

- Learning models from given (passively collected/labeled) data
- Inference on arbitrary instances

In many practical/engineering settings some control over data collection

- Sensor management (run time)
- Data collection/labeling (training)
Actively collecting data and controlling sensing modality
Actively collecting data
and
Controlling sensing modality
Actively collecting data and controlling sensing modality
Actively collecting data and controlling sensing modality
Actively collecting data and controlling sensing modality

informative sample must maximally reduce uncertainty
Actively collecting data and controlling sensing modality

Informative sample must maximally reduce uncertainty
Identify appropriate notion of uncertainty
Actively collecting data and controlling sensing modality

An informative sample must maximally reduce uncertainty.

Identify appropriate notion of uncertainty.

Converses: fundamental limits achievability: algorithms.
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshtabar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyuan Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz

Rate vs Reliability
Extrinsic Jensen-Shannon Divergence
Noisy Search
Active Machine Learning
Conclusion

Introduction and Motivation

Information Acquisition

A Brief History
Information Acquisition

- Stochastically varying state/parameter
- Tracked via partial/noisy yet controlled observations
- Controlled sequence of observations

- Generalizes:

<table>
<thead>
<tr>
<th>time</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>X_1</td>
<td>X_2</td>
<td>...</td>
<td>X_T</td>
</tr>
<tr>
<td>acquisition action</td>
<td>A_1</td>
<td>A_2</td>
<td>...</td>
<td>A_T</td>
</tr>
<tr>
<td>observation</td>
<td>Y_1</td>
<td>Y_2</td>
<td>...</td>
<td>Y_T</td>
</tr>
<tr>
<td>utilization action</td>
<td>U_1</td>
<td>U_2</td>
<td>...</td>
<td>U_T</td>
</tr>
</tbody>
</table>

- Distribution of Y_t determined by X_t and A_t
- Stochastic dynamics given as P(Y_{t+1}|X_t,A_t)

Objective:

\[ \text{Maximize } V \left[ \sum_{t=1}^{T} R_t(U_t(X_t), Y_t) - C_t(A_t) \right] \]
Information Acquistion

- Stochastically varying state/parameter
- Tracked via partial/noisy yet controlled observations
- Controlled sequence of observations

- Generalizes:

Comparison of Experiments

Different Conclusions!
<table>
<thead>
<tr>
<th>time</th>
<th>1</th>
<th>2</th>
<th>...</th>
<th>T</th>
</tr>
</thead>
<tbody>
<tr>
<td>state</td>
<td>$X_1$</td>
<td>$X_2$</td>
<td>...</td>
<td>$X_T$</td>
</tr>
<tr>
<td>acquisition action</td>
<td>$A_1$</td>
<td>$A_2$</td>
<td>...</td>
<td>$A_T$</td>
</tr>
<tr>
<td>observation</td>
<td>$Y_1$</td>
<td>$Y_2$</td>
<td>...</td>
<td>$Y_T$</td>
</tr>
<tr>
<td>utilization action</td>
<td>$U_1$</td>
<td>$U_2$</td>
<td>...</td>
<td>$U_T$</td>
</tr>
</tbody>
</table>

- Distribution of $Y_t$ is determined by $X_t$ and $A_t$
- Stochastic dynamic given as $\mathbb{P}(X_{t+1}|X_{1:t})$

**Objective:**

$$\text{Maximize } \mathbb{E} \left[ \sum_t R_u(U(t),X(t)) - C_a(A(t)) \right]$$
Informal

- Stochastically
Experiment Design

Introduced by Blackwell & Stein in 1952

Single-shot Problem

- Consider a single experiment $a \in A$
  - $M$ mutually exclusive hypotheses: $H_i \Leftrightarrow \{\theta = i\}$
  - Noisy observations subject to $\{q^a_i(\cdot)\}_{i,a}$
- What should $a$ be? Compare experiment $a$ with $a'$?
  - Stochastically degraded case [Blackwell ’53, Stein ’53]
Experiment Design

Introduced by Blackwell & Stein in 1952

Single-shot Problem

- Given experiment $a$:
  - True hypothesis $\theta = i$ with probability $\rho_i$
  - Output distribution $Z^a \sim \sum_{i=1}^{M} \rho_i q^a_i (\cdot)$
  - Posterior upon observation $\theta | Z^a \sim \Phi^a (\rho, Z^a)$
  - How does $\Phi^a (\cdot, \cdot)$ compare with $\Phi^{a'} (\cdot, \cdot)$
Active Hypothesis Testing

Introduced by Chernoff in 1953

- $M$ mutually exclusive hypotheses: $H_i \leftrightarrow \{ \theta = i \}$, $i = 1, 2, \ldots, M$

- Experiments $A(t) \in A$ chosen sequentially

<table>
<thead>
<tr>
<th>time</th>
<th>0</th>
<th>1</th>
<th>$\ldots$</th>
<th>$\tau - 1$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample</td>
<td>$A(0)$</td>
<td>$A(1)$</td>
<td>$\ldots$</td>
<td>$A(\tau - 1)$</td>
<td></td>
</tr>
<tr>
<td>observation</td>
<td>$Z(0)$</td>
<td>$Z(1)$</td>
<td>$\ldots$</td>
<td>$Z(\tau - 1)$</td>
<td></td>
</tr>
<tr>
<td>declaration</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\hat{\theta} = d(Z^{\tau-1}, x^{\tau-1})$</td>
</tr>
<tr>
<td>error</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$1_{{\hat{\theta} \neq \theta}}$</td>
</tr>
</tbody>
</table>

- $Z|_{\{\theta=1, A=a\}} \sim q_i^a(\cdot)$; observation density given $a \in A$ and $H_i$

- Uniform Prior $\rho(0) = [\frac{1}{M}, \frac{1}{M}, \ldots, \frac{1}{M}]$
Information Utility

Introduced by De Groot in 1963

Information Utility Heuristics:

- Measure of uncertainty $V$ [DeGroot 1962]
- Information utility associated with $V$

$$IU(a, \rho, V) = V(\rho) - \mathbb{E}[V(\Phi^a(\rho, Z))]$$

- Most informative action $\arg \max_a IU(a, \rho, V)$

Bayes operator
Which intuition works better?
Which intuition works better?

Asymptotic Optimal (Chernoff's):

Information Utility (DeGroot):
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshevar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz
A Fundamental Trade-off

Performance Measures

\[ M, \quad P_e, \quad N (= \mathbb{E}(\tau)) \]

Asymptotically reliable (given fixed hypotheses class): \( P_e \approx e^{-NE} \)

Asymptotically complex hypotheses class (given fixed reliability): \( M \approx 2^{NR} \)
Asymptotic reliability (Chernoff's)

Classical Statistics

max-min statistical discrimination

(KL-divergence)
Complex Hypothesis Class (Small Sample Regime)
Complex Hypothesis Class (Small Sample Regime)

Uncountable hypothesis classs
Complex Hypothesis Class (Small Sample Regime)

Uncountable hypothesis classes

Speedy elimination
Complex Hypothesis Class (Small Sample Regime)

Uncountable hypothesis classes

Speedy elimination
Complex Hypothesis Class (Small Sample Regime)

Uncountable hypothesis classes

Speedy elimination
Complex Hypothesis Class (Small Sample Regime)

Uncountable hypothesis classes

Speedy elimination
Complex Hypothesis Class (Small Sample Regime)

Uncountable hypothesis classes

Speedy elimination

Information Acquisition Rate
Performance Measures

\[ M, \quad P_e, \quad N(= \mathbb{E}(\tau)) \]

Asymptotically reliable (given fixed hypotheses class): \( P_e \approx e^{-NE} \)

Asymptotically complex hypotheses class (given fixed reliability): \( M \approx 2^{NR} \)
Performance Measures

\[ M, \quad P_e, \quad N = \frac{1}{1 - P_e} \]

Asymptotically reliable (given fixed hypotheses class): \( P_e \approx e^{-NE} \)

Asymptotically complex hypotheses class (given fixed reliability): \( M \approx 2^{NR} \)
\[ P_e \approx e^{-NE} \]
\[ M \approx 2^{NR} \]
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhamshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz
Extrinsic Jensen-Shannon

Analysis is based on geometry of belief update (filtering)

Information Utility
Divergence

Symmetrized Divergence

Extrinsic Jensen-Shannon

Dynamical System View
Recall

Information Utility Heuristics:

- Measure of uncertainty $V$ [DeGroot 1962]
- Information utility associated with $V$

$$IU(a, \rho, V) = V(\rho) - \mathbb{E}[V(\Phi^a(\rho, Z))]$$

- Most informative action $\arg \max_a IU(a, \rho, V)$
Alternative Heuristics

Another geometric approach:

**Divergence-based Selection**

- Define a “symmetrized divergence” among $q_1^a, q_2^a, \ldots, q_M^a$
- Best action must maximize the divergence
  - maximize discrimination among $H_1, H_2, \ldots, H_M$
Alternative Heuristics

Another geometric approach:

Divergence-based Selection

- Define a “symmetrized divergence” among $q_1^a, q_2^a, \ldots, q_M^a$
- Best action must maximize the divergence
  - maximize discrimination among $H_1, H_2, \ldots, H_M$
Alternative Heuristics

Another geometric approach:

**Divergence-based Selection**

- Define a “symmetrized divergence” among $q_1^a, q_2^a, \ldots, q_M^a$
- Best action must maximize the divergence
  - maximize discrimination among $H_1, H_2, \ldots, H_M$

The Kullback-Leibler (KL) divergence between $p(\cdot)$ and $q(\cdot)$:

$$D(p||q) = \sum_z p(z) \log \frac{p(z)}{q(z)}$$
Alternative Heuristics

Another geometric approach:

Divergence-based Selection

- Define a “symmetrized divergence” among $q_1^a$, $q_2^a$, \ldots, $q_M^a$
- Best action must maximize the divergence
  - maximize discrimination among $H_1$, $H_2$, \ldots, $H_M$

The Kullback-Leibler (KL) divergence between $p(\cdot)$ and $q(\cdot)$:

$$D(p||q) = \sum_z p(z) \log \frac{p(z)}{q(z)}$$
Alternative Heuristics

Another geometric approach:

Divergence-based Selection

- Define a “symmetrized divergence” among \( q_1^\alpha, q_2^\alpha, \ldots, q_M^\alpha \)
- Best action must maximize the divergence
  - maximize discrimination among \( H_1, H_2, \ldots, H_M \)

\[
\begin{align*}
J\text{-divergence [Jefferys 73]} \\
D_J(f, g) &= \frac{1}{2} D(f\|g) + \frac{1}{2} D(g\|f) \\
L\text{-divergence [Lin 91]} \\
D_L(f, g) &= \frac{1}{2} D(f\|\frac{f+g}{2}) + \frac{1}{2} D(g\|\frac{f+g}{2})
\end{align*}
\]

The Kullback-Leibler (KL) divergence between \( p(\cdot) \) and \( q(\cdot) \):

\[
D(p\|q) = \sum_z p(z) \log \frac{p(z)}{q(z)}
\]
Alternative Heuristics

Another geometric approach:

Divergence-based Selection

- Define a “symmetrized divergence”:
  - $J$-divergence [Jefferys 73]
    \[
    D_J(f, g) = \frac{1}{2} D(f||g) + \frac{1}{2} D(g||f)
    \]
  - $L$-divergence [Lin 91]
    \[
    D_L(f, g) = \frac{1}{2} D(f||\frac{f+g}{2}) + \frac{1}{2} D(g||\frac{f+g}{2})
    \]

- Best action must maximize discriminative ability
  - maximize discrimination

Jänsen-Shannon divergence [Lin 1991]

Generalizing $L$ divergence: $D_L(f, g) = \frac{1}{2} D(f||\frac{f+g}{2}) + \frac{1}{2} D(g||\frac{f+g}{2})$

The Kullback-Leibler (KL) divergence between $p(\cdot)$ and $q(\cdot)$:

\[
D(p||q) = \sum_z p(z) \log \frac{p(z)}{q(z)}
\]
Proposed Divergence

Extrinsic Jensen-Shannon Divergence [Naghshvar, J. ISIT’12]

The Extrinsic Jensen-Shannon (EJS) divergence among densities $q_1, q_2, \ldots, q_M$ with respect to $\rho = [\rho_1, \rho_2, \ldots, \rho_M]$ is defined as

$$EJS(\rho; q_1, q_2, \ldots, q_M) = \sum_{i=1}^{M} \rho_i D(q_i \| \sum_{k \neq i} \frac{\rho_k}{1-\rho_i} q_k).$$
Proposed Divergence

Extrinsic Jensen-Shannon Divergence [Naghshvar, J. ISIT’12]

The Extrinsic Jensen-Shannon (EJS) divergence among densities \( q_1, q_2, \ldots, q_M \) with respect to \( \rho = [\rho_1, \rho_2, \ldots, \rho_M] \) is defined as

\[
EJS(\rho; q_1, q_2, \ldots, q_M) = \sum_{i=1}^{M} \rho_i D(q_i \| \sum_{k \neq i} \frac{\rho_k}{1-\rho_i} q_k).
\]
Proposed Divergence

Extrinsic Jensen-Shannon Divergence [Naghshvar, J. ISIT'12]

Extrinsic Jensen-Shannon (EJS) divergence among densities $q_1, q_2, \ldots, q_M$ with respect to $\rho = [\rho_1, \rho_2, \ldots, \rho_M]$ is defined as

$$EJS(\rho; q_1, q_2, \ldots, q_M) = \sum_{i=1}^{M} \rho_i D(q_i || \sum_{k \neq i} \frac{\rho_k}{1-\rho_i} q_k).$$

Proposition

EJS is the information utility associated with the average likelihood function $U(\rho) = \sum_{i=1}^{M} \rho_i \log \frac{1-\rho_i}{\rho_i}$, i.e.

$$EJS(\rho; q_1^a, \ldots, q_M^a) = \mathcal{IU}(a, \rho, U)$$
Proposed Divergence

**Extrinsic Jensen-Shannon Divergence** [Naghshvar, J. ISIT’12]

The **Extrinsic Jensen-Shannon (EJS) divergence** among densities $q_1, q_2, \ldots, q_M$ with respect to $\rho = [\rho_1, \rho_2, \ldots, \rho_M]$ is defined as

$$EJS(\rho; q_1, q_2, \ldots, q_M) = \sum_{i=1}^{M} \rho_i D(q_i \| \sum_{k \neq i} \rho_k \frac{q_k}{1-\rho_i} \cdot q_k).$$

**Proposition**

EJS is the information utility associated with the average likelihood function $U(\rho) = \sum_{i=1}^{M} \rho_i \log \frac{1-\rho_i}{\rho_i}$, i.e.

$$EJS(\rho; q_1^\alpha, \ldots, q_M^\alpha) = IU(a, \rho, U)$$
**Dynamical System View**

Flattened geometry of Information Utility
e.g. in hypothesis testing:

- Consider (suboptimal) \( \tau = \min \{ t : \max_i \rho_i(t) \geq 1 - \epsilon \} \)
  - Stop search and declare \( \hat{\theta} = i \) if \( \rho_i(t) \geq 1 - \epsilon \) (satisfies \( \text{Pe} \leq \epsilon \))

instead work with:

- Take concave functional \( U \) (bounded \( |U(\rho(t+1)) - U(\rho(t))| \leq \Delta \))
**Dynamical System View**

Flattened geometry of Information Utility
e.g. in hypothesis testing:

- Consider (suboptimal) $\tau = \min\{t : \max_i \rho_i(t) \geq 1 - \epsilon\}$
  - Stop search and declare $\hat{\theta} = i$ if $\rho_i(t) \geq 1 - \epsilon$ (satisfies $Pe \leq \epsilon$)

instead work with:

- Take concave functional $U$ (bounded $|U(\rho(t+1)) - U(\rho(t))| \leq \Delta$)
Lyapunov-type Analysis
flatten the geometry

- guessing the right notion of uncertainty
- tracking the expected reduction in uncertainty
- converses
- achievability
Lyapunov-type Analysis

flatten the geometry

Information Utility Based Analysis: Converse

- Consider (suboptimal) \( \tau = \min \{ t : \max_i \, \rho_i(t) \geq 1 - \varepsilon \} \)
  - Stop search and declare \( \hat{\theta} = i \) if \( \rho_i(t) \geq 1 - \varepsilon \) (satisfies \( \text{Pe} \leq \varepsilon \))

- Take concave function \( U \) (bounded \( |U(\rho(t + 1)) - U(\rho(t))| \leq \Delta \)).

- Suppose for any policy \( \epsilon \) selecting action \( a \)

  \[ I(a, \rho, U) \leq \bar{\alpha}, \text{ for some positive } \bar{\alpha}. \]

Then,

\[ \mathbb{E}[\tau^*] \gtrsim \frac{U(\rho) - U([1 - \varepsilon, \varepsilon])}{\bar{\alpha}} + \frac{\Delta}{\bar{\alpha}}. \]
Lyapunov-type Analysis
flattens the geometry

Information Utility Based Analysis: Achievability

- Consider (suboptimal) $\tau = \min \{ t : \max_i \rho_i(t) \geq 1 - \epsilon \}$
  - Stop search and declare $\hat{\theta} = i$ if $\rho_i(t) \geq 1 - \epsilon$ (satisfies $P_e \leq \epsilon$)

- Take concave function $W$ (bounded $|W(\rho(t+1)) - W(\rho(t))| \leq \Delta$).

- Suppose policy $\epsilon$ selects action $a$ such that
  \[ I(a, \rho, W) \geq \alpha, \]  for some positive $\alpha$.

Then,
\[ \mathbb{E}[\tau^*] \leq \frac{W(\rho) - W(\rho(1 - \epsilon, \epsilon))}{\alpha} + \frac{\Delta}{\alpha}. \]
Lyapunov-type Analysis
flatten the geometry

---

**Information Utility Based Analysis: Achievability**

- Consider (suboptimal) \( \tau = \min \{ t : \max_i \rho_i(t) \geq 1 - \epsilon \} \)
  - Stop search and declare \( \hat{\theta} = i \) if \( \rho_i(t) \geq 1 - \epsilon \) (satisfies \( Pe \leq \epsilon \))
- Take concave function \( W(\rho(t + 1)) - W(\rho(t)) \leq \Delta \).
- Suppose policy \( \epsilon \) selects action \( a \) such that
  \[
  I(a, \rho, W) \geq \alpha, \text{ for some positive } \alpha.
  \]

Then,
\[
E[\tau^*] \leq \frac{W(\rho) - W([1 - \epsilon, \epsilon])}{\alpha} + \frac{\Delta}{\alpha}.
\]
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshevar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz
Case Study I
Measurement-dependent Noisy Search

Motivation: UAVs for object search in rescue/survey

Challenge: Trade-off between elevation and coverage

- CV-based classifiers (convolutional neural net) fall when flying high
- Flying close inherently inefficient especially in large search scenarios
Beam Alignment for mm-wave communications
Measurement-Dependent Noisy Search

\[ a(\phi) := \alpha[1, e^{\frac{2\pi d}{\lambda} \sin \phi}, \ldots, e^{i(N-1)\frac{2\pi d}{\lambda} \sin \phi}] \]
Case Study I
Measurement-dependent Noisy Search

Motivation: UAVs for object search in rescue/survey

Challenge: Trade-off between elevation and coverage

- CV-based classifiers (convolutional neural net) fail when flying high
- Flying close inherently inefficient especially in large search scenarios

In Practice...
Measurement-Dependent Noisy Search

- Unknown parameter in an interval size \( B \) with resolution \( \delta \), \( \theta \in \{0,1\}^{B/\delta} \)
- Inspection decision \( A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta} \)
- Noisy observation \( Y^A = A \cdot \theta + \hat{Z}^A \)
  - Noise variance increases w/ \(|A(t)|\)
    - \( \text{e.g.} \ Y(t) = A(t)(\theta + Z_t), \ Z_t \sim \mathcal{N}(0, \delta \sigma I) \)

<table>
<thead>
<tr>
<th>time</th>
<th>sample</th>
<th>observation</th>
<th>declaration</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \tau - 1 )</td>
<td>( \cdots )</td>
<td>( \cdots )</td>
<td>( \tau )</td>
<td>( \hat{\theta} = d(Y_{1: \tau-1}, A_{1: \tau-1}) )</td>
</tr>
</tbody>
</table>
| 1 | \( A(1) \) | \( Y(1) \) | \( 1_{\{\hat{\theta} \neq \theta\}} \) \\
| \( \cdots \) | \( \cdots \) | \( \cdots \) | \( \cdots \) | \( \cdots \) |
Measurement-Dependent Noisy Search

- Unknown parameter in an interval size $B$ with resolution $\delta$, $\theta \in \{0,1\}^{B/\delta}$
- Inspection decision $A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta}$
- Noisy observation $Y^A = A \cdot \theta + \hat{Z}^A$
  - Noise variance increases w/ $|A(t)|$ e.g. $Y(t) = A(t)(\theta + \mathbf{Z}_t)$, $\mathbf{Z}_t \sim \mathcal{N}(0, \delta \sigma I)$

<table>
<thead>
<tr>
<th>time</th>
<th>sample</th>
<th>observation</th>
<th>declaration</th>
<th>error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$A(1)$</td>
<td>$Y(1)$</td>
<td>$\hat{\theta} = d(Y_{1:\tau-1}, A_{1:\tau-1})$</td>
<td>$1_{{\hat{\theta} \neq \theta}}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tau$</td>
<td>...</td>
<td>$Y(\tau - 1)$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\tau$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Measurement-Dependent Noisy Search

- Unknown parameter in an interval size $B$ with resolution $\delta$, $\theta \in \{0,1\}^{B/\delta}$
- Inspection decision $A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta}$
- Noisy observation $Y^A = A \cdot \theta + \hat{Z}^A$
  - Noise variance increases w/ $|A(t)|$ e.g. $Y(t) = A(t)(\theta + Z_t)$, $Z_t \sim \mathcal{N}(0, \delta \sigma I)$

| time      | 1    | ... | $\tau - 1$ | $\tau$
|-----------|------|-----|------------|-----|
| sample    | $A(1)$ | ... | $A(\tau - 1)$ | $\tau$
| observation | $Y(1)$ | ... | $Y(\tau - 1)$ | $\tau$
| declaration |       |     |             | $\hat{\theta} = d(Y_{1:\tau-1}, A_{1:\tau-1})$ |
| error     |       |     |             | $1_{\{\hat{\theta} \neq \theta\}}$ |
Measurement-Dependent Noisy Search

- Unknown parameter in an interval size $B$ with resolution $\delta$, $\theta \in \{0, 1\}^{B/\delta}$
- Inspection decision $A(t) \in \mathcal{A} \subset \{0, 1\}^{B/\delta}$
- Noisy observation $Y^A = A \cdot \theta + \hat{Z}^A$

  Noise variance increases with $|A(t)|$ e.g. $Y(t) = A(t)(\theta + Z_t)$, $Z_t \sim \mathcal{N}(0, \delta \sigma I)$

| time | 1 | ... | $\tau - 1$ | $\tau$
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>sample</td>
<td>$A(1)$</td>
<td>...</td>
<td>$A(\tau - 1)$</td>
<td></td>
</tr>
<tr>
<td>observation</td>
<td>$Y(1)$</td>
<td>...</td>
<td>$Y(\tau - 1)$</td>
<td></td>
</tr>
</tbody>
</table>
| declaration | \( \hat{\theta} = d(Y_{1:\tau - 1}, A_{1:\tau - 1}) \)
| error | $1_{\{\hat{\theta} \neq \theta\}}$ |
Measurement-Dependent Noisy Search

- Unknown parameter in an interval size $B$ with resolution $\delta$, $\theta \in \{0,1\}^{B/\delta}$
- Inspection decision $A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta}$
- Noisy observation $Y^A = A \cdot \theta + \hat{Z}^A$
  - Noise variance increases w/ $|A(t)|$ e.g. $Y(t) = A(t)(\theta + Z_t)$, $Z_t \sim \mathcal{N}(0, \delta \sigma I)$

<table>
<thead>
<tr>
<th>time</th>
<th>1</th>
<th>...</th>
<th>$\tau - 1$</th>
<th>$\tau$</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample</td>
<td>$A(1)$</td>
<td>...</td>
<td>$A(\tau - 1)$</td>
<td></td>
</tr>
<tr>
<td>observation</td>
<td>$Y(1)$</td>
<td>...</td>
<td>$Y(\tau - 1)$</td>
<td></td>
</tr>
<tr>
<td>declaration</td>
<td></td>
<td></td>
<td></td>
<td>$\hat{\theta} = d(Y_{1:\tau-1}, A_{1:\tau-1})$</td>
</tr>
<tr>
<td>error</td>
<td></td>
<td></td>
<td></td>
<td>$1_{{\hat{\theta} \neq \theta}}$</td>
</tr>
</tbody>
</table>
Measurement-Dependent Noisy Search

- Unknown parameter in an interval size $B$ with resolution $\delta$, \( \theta \in \{0,1\}^{B/\delta} \)
- Inspection decision \( A(t) \in \mathcal{A} \subset \{0,1\}^{B/\delta} \)
- Noisy observation \( Y^A = A \cdot \theta + \hat{Z}^A \)
  - Noise variance increases with \( |A(t)| \) e.g. \( Y(t) = A(t)(\theta + Z_t), \quad Z_t \sim \mathcal{N}(0, \delta \sigma I) \)

<table>
<thead>
<tr>
<th>time</th>
<th>1</th>
<th>...</th>
<th>( \tau - 1 )</th>
<th>( \tau )</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample</td>
<td>( A(1) )</td>
<td>...</td>
<td>( A(\tau - 1) )</td>
<td>( A(\tau) )</td>
</tr>
<tr>
<td>observation</td>
<td>( Y(1) )</td>
<td>...</td>
<td>( Y(\tau - 1) )</td>
<td>( Y(\tau) )</td>
</tr>
<tr>
<td>declaration</td>
<td>( \hat{\theta} = d(Y_{1:\tau-1}, A_{1:\tau-1}) )</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>error</td>
<td>( 1_{{\hat{\theta} \neq \theta}} )</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Two Important Findings:

Advantage of allowing for wide-area Searches

Advantage of adaptive search strategies over open-loop

Two Important Findings:

Advantage of allowing for wide-area Searches

Advantage of adaptive search strategies over open-loop

Two Important Findings:

Advantage of allowing for wide-area Searches

Advantage of adaptive search strategies over open-loop

Two Important Findings:

Advantage of allowing for wide-area Searches

Advantage of adaptive search strategies over open-loop

Main Analytical (Information Theoretic) Insight
Main Analytical (Information Theoretic) Insight

Observation: \( Y^A = \underbrace{1_{\sup(A) \cap \sup(\theta) \neq 0}}_{X^A} + \hat{Z}^A \)
Main Analytical (Information Theoretic) Insight

Observation: $Y^A = \underbrace{1_{\sup(A) \cap \sup(\theta) \neq 0}}_{X^A} + Z^A$

Binary input additive (Gaussian) channel
Main Analytical (Information Theoretic) Insight

**Observation:** \[ Y^A = \frac{1_{\text{sup}(A) \cap \text{sup}(\theta) \neq 0}}{X^A} + \hat{Z}^A \]

Binary input additive (Gaussian) channel
Main Analytical (Information Theoretic) Insight

Observation: \( Y^A = \mathbb{1}_{\text{sup}(A) \cap \text{sup}(\theta) \neq 0} + \hat{Z}^A \)

Binary input additive (Gaussian) channel \( \Rightarrow \mathbb{E}[\tau] \approx \frac{\log B/\delta \epsilon}{I(X, Y^A)} \) is sufficient.
Main Analytical (Information Theoretic) Insight

Observation: \( Y^A = \frac{1}{\sup(A) \cap \sup(\theta) \neq 0} + \hat{Z}^A \)

\[ Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \quad Z^q \sim \mathcal{N}(0, qB\sigma^2) \]

\[ \mathbb{E}[\tau^N_{\epsilon}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q, \sigma \sqrt{qB})} \]

![Table and Diagram](image-url)
Main Analytical (Information Theoretic) Insight

Observation: \( Y^A = \mathbb{1}_{\sup(A) \cap \sup(\theta) \neq 0} + \hat{Z}^A \)

\( Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \quad Z^q \sim \mathcal{N}(0, qB\sigma^2) \)

\[ \mathbb{E}[\tau^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma \sqrt{q^*B})} \]

\[
\begin{array}{|c|c|c|c|c|}
\hline
   & 1 & 2 & 3 & \ldots & N \\
\hline
1 & 0 & 1 & 0 & \ldots & 1 \\
2 & 1 & 1 & 0 & \ldots & 1 \\
3 & 0 & 0 & 1 & \ldots & 1 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
M & 1 & 0 & 1 & \ldots & 0 \\
\hline
\end{array}
\]

Decoder
Main Analytical (Information Theoretic) Insight

Observation: \[ Y^A = 1_{\text{sup}(A) \cap \text{sup}(\theta) \neq 0} + \hat{Z}^A \]

\[ Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \quad Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau^N_{\epsilon}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{\beta_1(q^*, \sigma^2q^*B)} \]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>M</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>
Main Analytical (Information Theoretic) Insight

Observation: \( Y^A = \mathbf{1}_{\sup(A) \cap \sup(\theta) \neq 0} + \hat{Z}^A \)

\( Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \quad Z^q \sim \mathcal{N}(0, qB\sigma^2) \)

\[ \mathbb{E}[\tau^A_{\epsilon}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma \sqrt{q^*B})} \]

\[
\begin{array}{ccccccc}
1 & 0 & 1 & 0 & \ldots & \ldots & 1 \\
2 & 1 & 1 & 0 & \ldots & \ldots & 1 \\
3 & 0 & 0 & 1 & \ldots & \ldots & 1 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \vdots \\
M & 1 & 0 & 1 & \ldots & 0 \\
\end{array}
\]
Main Analytical (Information Theoretic) Insight

**Observation:** \[ Y^A = 1_{\sup(A) \cap \sup(\theta) \neq 0} + \hat{Z}^A \]

\[ Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \ Z^q \sim \mathcal{N}(0, qB\sigma^2) \]

\[ \mathbb{E}[\tau^{NA}_\epsilon] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{BPSK}(q^*, \sigma \sqrt{q^*B})} \]
Main Analytical (Information Theoretic) Insight

Observation: \( Y^A = 1_{\sup(A) \cap \sup(\theta) \neq 0} + \hat{Z}^A \)

\( Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \; Z^q \sim \mathcal{N}(0, qB\sigma^2) \)

\( \mathbb{E}[\tau_{\epsilon}^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma \sqrt{q^* B})} \)
Main Analytical (Information Theoretic) Insight

Observation:  \( Y^A = \underbrace{1_{\sup(A) \cap \sup(\theta) \neq 0}}_{X^A} + \hat{Z}^A \)

\[ Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \ Z^q \sim \mathcal{N}(0, qB\sigma^2) \]
\[ \mathbb{E}[\tau^\text{NA}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})} \]

- Adaptive strategy builds on posterior matching
  - Ensures high \( I(\theta, Y(t)) \)
Main Analytical (Information Theoretic) Insight

\[ Y^A = \frac{1_{\sup(A) \cap \sup(\theta) \neq 0}}{\sup(A)} + \hat{Z}^A \]

\[ Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \quad Z^q \sim \mathcal{N}(0, qB\sigma^2) \]

\[ \mathbb{E}[\tau^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})} \]

- Adaptive strategy builds on posterior matching
- Ensures high \( I(\theta, Y(t)) \)

![Network Diagram]

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Search Width \( \delta \)
Main Analytical (Information Theoretic) Insight

Observation: \( Y^A = \underbrace{1_{\text{sup}(A) \cap \text{sup}(\theta) \neq 0}}_{X^A} + \hat{Z}^A \)

\[ Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \ Z^q \sim \mathcal{N}(0, qB\sigma^2) \]

\[ \mathbb{E}[\tau_{\epsilon}^{\text{NA}}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})} \]

- Adaptive strategy builds on posterior matching
  - Ensures high \( I(\theta, Y(t)) \)
Main Analytical (Information Theoretic) Insight

Observation: \[ Y^A = 1_{\sup(A) \cap \sup(\theta) \neq 0} + \hat{Z}^A \]

\[ Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \quad Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau^\text{NA}] \geq \frac{(1 - \epsilon) \log \frac{B}{\delta} - h(\epsilon)}{C_{\text{BPSK}}(q^*, \sigma\sqrt{q^*B})} \]

- Adaptive strategy builds on posterior matching
- Ensures high \( I(\theta, Y(t)) \)
Main Analytical (Information Theoretic) Insight

Observation: \[ Y^A = \frac{1}{\sup(A) \cap \sup(\theta) \neq 0} + \hat{Z}^A \]

\[ Y = X^q + Z^q, \quad X^q \sim \text{Ber}(q), \quad Z^q \sim \mathcal{N}(0, qB\sigma^2) \quad \mathbb{E}[\tau^N] \geq \frac{(1 - \varepsilon) \log \frac{B}{\delta} - h(\varepsilon)}{C_{\text{BPSK}}(q^*, \sigma \sqrt{q^*B})} \]

- Adaptive strategy builds on posterior matching

- Ensures high \( I(\theta, Y(t)) \)

![Diagram showing a bar graph and a schematic of a process with inputs and outputs.]
Summary: Two Important Questions Answered

- Role of allowable actions set $A$
  - Designing $A$ can significantly reduce the overhead
    ‣ Even when noise variance increases w/ $|A(t)|$ (linearly)!

- Adapt $A(t)$ to past observations (feedback) or not?
  - Adaptive policies are computationally expensive but significant adaptivity gain in low SNR regimes
In theory, there is no difference between theory and practice...

In practice, there is!
Measurement-Dependent Noisy Search

\[ y_t = \sqrt{P} w_t^H a(\phi) + w_t^H n_t \]

\[ a(\phi) := \alpha[1, e^{j\frac{2\pi d}{\lambda} \sin \phi}, ..., e^{j(N-1)\frac{2\pi d}{\lambda} \sin \phi}] \]
Measurement-Dependent Noisy Search

\[\begin{align*}
\text{time} & : 1 \ldots \tau - 1 \\
\text{beam} & : W_1 \ldots W_{\tau - 1} \\
\text{observe} & : y_1 \ldots y_{\tau - 1} \\
\text{detect} & \\
\text{error} & \\
\end{align*}\]

Adaptive Beamforming Design

\[y_{1:t} \rightarrow w_{t+1} \in \mathcal{W}\]

Single RF chain

\[y_t = \sqrt{P} w_t^H a(\phi) + w_t^H n_t\]

\[a(\phi) := [\alpha, e^{j \frac{2\pi d}{\lambda} \sin \phi}, \ldots, e^{j (N-1) \frac{2\pi d}{\lambda} \sin \phi}]\]

Region of Interest (for sector 1)

3 sectors, 120 degrees each sector

not known
Measurement-Dependent Noisy Search

\[ \text{time} \quad 1 \quad \ldots \quad \tau - 1 \]

<table>
<thead>
<tr>
<th>beam</th>
<th>( W_1 )</th>
<th>( W_{\tau - 1} )</th>
</tr>
</thead>
</table>

| observe | \( y_1 \) | \( y_{\tau - 1} \) |

detect error

Adaptive Beamforming Design

\[ y_{1:t} \to \mathbf{w}_{t+1} \in \mathbb{W} \]

\[ \mathbf{Y}_t \leftarrow \mathbf{y}_t \]

\[ \mathbf{y}_t = \sqrt{p} \mathbf{a}(\phi) + \mathbf{w}_t \]

\[ \phi = d(y_{1:\tau - 1}, W_{1:\tau - 1}) \]

\[ \mathbf{1}_{\{\mathbf{w} \neq \mathbf{w}^*\}} \]

not arbitrary

\[ a(\phi) := [1, e^{j \frac{2\pi d}{\lambda} \sin \phi}, \ldots, e^{j(N-1) \frac{2\pi d}{\lambda} \sin \phi}] \]

not known
Measurement-Dependent Noisy Search

Hierarchical Beam Patterns

Binary search
Repeat to increase SNR (linear in beam width)


not arbitrary

\[ a(\phi) := \alpha \mathbf{1}, e^{j \frac{2\pi d}{\lambda} \sin \phi}, \ldots, e^{j(N-1) \frac{2\pi d}{\lambda} \sin \phi} \]

not known
Measurement-Dependent Noisy Search

Hierarchical Beam Patterns

Adaptive Beamforming Design

\[ y_{1:t} \rightarrow \mathbf{w}_{t+1} \in \mathcal{W} \]

Code over the beam patterns water filling in angular space

\[ y_t = \sqrt{P_t} \mathbf{a}(\phi) + \mathbf{w}_t \]

\[ \mathbf{w}^{H} \mathbf{a}(\phi) \]

\[ \alpha, e^{j\frac{2\pi d}{\lambda} \sin \phi}, \ldots, e^{j(N-1)\frac{2\pi d}{\lambda} \sin \phi} \]

not arbitrary

not known

Demo: Parrot Platform
Demo: Parrot Platform

- Cell phone acts as a server for downloading images from, classify/compute belief, and control commands to drone
- Implementing/tuning MobileNet CNN
- Java ServerSocket for receiving/sending messages on drone
- Mobile app built with Android Studio + Parrot Software Development Kit

Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghshevar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz

Active Machine Learning
Information Acquisition
Introduction and Motivation
A Brief History
Rate vs Reliability
Extrinsic Jensen-Shannon Divergence
Noisy Search

UC San Diego
Jacobs School of Engineering
Center for Wireless Communications

Machine Integrated Computing and Security (MICAS)

NSF
Case Study II

Active Learning from Imperfect Labeler
Active Learning from Logged Data
Adaptive Hyper-parameter tuning
Problem Statement

Case Study II
Active Learning from Imperfect Labeler
Active Learning from Logged Data
Adaptive Hyper-parameter tuning

Prior Work

Our
Problem Statement

Case Study II

Active Learning from Imperfect Labeler

Active Learning from Logged Data

Adaptive Hyper-parameter tuning

- Classification
- Classical Approach: Passive Learning
- Problem: Largely Redundant Labels
- Active Learning

Prior Work

Our
Problem Statement

Case Study II

Active Learning from Imperfect Labeler

Active Learning from Logged Data

Adaptive Hyper-parameter tuning

Prior Work

- Classification
- Classical Approach: Passive Learning
- Problem: Largely Redundant Labels
- Active Learning

Our
Problem Statement

Case Study II
Active Learning from Imperfect Labeler
Active Learning from Logged Data
Adaptive Hyper-parameter tuning

Prior Work

Our
Case Study II

Active Learning from Imperfect Labeler

Active Learning from Logged Data

Adaptive Hyper-parameter tuning
Case Study II
Active Learning from Imperfect Labeler
Active Learning from Logged Data
Adaptive Hyper-parameter tuning

Prior Work

Our
Black-box optimization

Consider optimizing a function $f : \mathcal{X} \rightarrow \mathbb{R}$

Caveat:
- $f$ is not known explicitly: accessed only through noisy and expensive evaluation queries
Black-box optimization

Consider optimizing a function $f : \mathcal{X} \to \mathbb{R}$

**Caveat:**
- $f$ is not known explicitly: accessed only through noisy and expensive evaluation queries

**Goal:** Design a sequential strategy of selecting $n$ query points $x_1, \ldots, x_n$ to efficiently optimize $f$ over the horizon $n$
Black-box optimization

Consider optimizing a function \( f : \mathcal{X} \rightarrow \mathbb{R} \)

**Caveat:**

- \( f \) is not known explicitly: accessed only through noisy and expensive evaluation queries

**Goal:** Design a sequential strategy of selecting \( n \) query points \( x_1, \ldots, x_n \) to efficiently optimize \( f \) over the horizon \( n \)

- Performance measures:
  - Simple regret: \( S_n = f(x^*) - f(x_n) \)
Black-box optimization

Consider optimizing a function \( f: \mathcal{X} \to \mathbb{R} \)

**Caveat:**

\( f \) is not known explicitly: accessed only through noisy and expensive evaluation queries

**Goal:** Design a sequential strategy of selecting \( n \) query points \( x_1, \ldots, x_n \) to efficiently optimize \( f \) over the horizon \( n \)

**Performance measures:**

- Simple regret: \( S_n = f(x^*) - f(x_n) \)

- Cumulative regret: \( R_n = \sum_{t=1}^{n} f(x^*) - f(x_t) \)
Black-box optimization

Consider optimizing a function $f : \mathcal{X} \to \mathbb{R}$ (consider $\mathcal{X} \subset \mathbb{R}^D$)

**Caveat:**
- $f$ is not known explicitly: accessed only through noisy and expensive evaluation queries

**Goal:** Design a sequential strategy of selecting $n$ query points $x_1, \ldots, x_n$ to efficiently optimize $f$ over the horizon $n$

- Performance measures:
  - Simple regret: $S_n = f(x^*) - f(x_n)$
  - Cumulative regret: $\mathcal{R}_n = \sum_{t=1}^{n} f(x^*) - f(x_t)$
Black-box optimization

Consider optimizing a function \( f : \mathcal{X} \rightarrow \mathbb{R} \)

**Caveat:**
- \( f \) is not known explicitly: accessed only through noisy and expensive evaluation queries

**Goal:** Design a sequential strategy of selecting \( n \) query points \( x_1, \ldots, x_n \) to efficiently optimize \( f \) over the horizon \( n \)

- Performance measures:
  - Simple regret: \( S_n = f(x^*) - f(x_n) \)
  - Cumulative regret: \( R_n = \sum_{t=1}^{n} f(x^*) - f(x_t) \)

- Ill-posed unless learning \( f(x) \) gives information about \( f(x') \)
Bayesian Setting: Gaussian Prior and Additive Noise

- $f$ is a sample drawn from a zero mean Gaussian process with covariance function $K(x, x') = \mathbb{E}[f(x)f(x')]$

- Observation model: $y = f(x) + \eta$ with $\eta \sim N(0, \sigma^2)$

- Gaussian posterior
  - Posterior mean and variance at $x$:
    
    $$\mu_t(x) = k_t(x)^T(K_t + \sigma^2 I)^{-1}y_{1:t-1}$$
    $$\sigma_t^2(x) = k(x, x) - k_t(x)^T(K_t + \sigma^2 I)^{-1}k_t(x)$$

    where
    $k_t(x) = \text{cov}(f(x), f(x_{1:t-1}))$
    and
    $K_t = \text{cov}(f(x_{1:t-1}), f(x_{1:t-1}))$
Application 1: Hyper-parameter tuning in ML models

- Training with hyperparameters $\theta$ outputs classifier $A(\theta)$
  - $\mathcal{X}$ = space of hyperparameters.

- $f(\theta)$ = performance of $A(\theta)$ on some test set.

- Finite $n$ trials $\Rightarrow$ good $\theta_n$ (pure exploration)

**Goal:** After $n$ rounds, output $\theta_n^*$ to minimize simple regret:

$$S_n = f(\theta^*) - f(\theta_n^*)$$
Application 1: Hyper-parameter tuning in ML models

- Training with hyperparameters $\theta$ outputs classifier $A(\theta)$
  - $\mathcal{X}$ = space of hyperparameters.
  - $f(\theta)$ = performance of $A(\theta)$ on some test set.
  - Finite $n$ trials $\Rightarrow$ good $\theta_n$ (pure exploration)

**Goal:** After $n$ rounds, output $\theta_n^*$ to minimize simple regret:

$$S_n = f(\theta^*) - f(\theta_n^*)$$

- Practically formulated as GP w Matérn family:

$$K_{\nu}^{\text{Matérn}}(x, x') = K(0) \left(1 + \sum_{i=1}^{m} a_i ||x - x'||^i \right) e^{-c_1 \sqrt{\nu} ||x - x'||},$$

(for half integer values of $\nu = m + 1/2$ and some $a_i > 0$)
Application 1: Hyper-parameter tuning in ML models

- Training with hyperparameters $\theta$ outputs classifier $A(\theta)$
  - $\mathcal{X}$ = space of hyperparameters.
  - $f(\theta)$ = performance of $A(\theta)$ on some test set.
  - Finite $n$ trials $\Rightarrow$ good $\theta_n$ (pure exploration)

Goal: After $n$ rounds, output $\theta_n^*$ to minimize simple regret:

$$S_n = f(\theta^*) - f(\theta_n^*)$$

- Practically formulated as GP w Matérn family:

$$K_{\nu}^{\text{Matérn}}(x, x') = K(0)(1 + \sum_{i=1}^{m} a_i ||x - x'||^i) e^{-c_1 \sqrt{\nu} ||x - x'||},$$

(for half integer values of $\nu = m + 1/2$ and some $a_i > 0$)
Gaussian Process Optimization (GP): Prior Work

- Zero mean Gaussian prior with known covariance function
- Update the posterior $\mathbb{P}_t$ based on $x_{1:t-1}, y_{1:t-1}$
- Query point according to acquisition rule:

$$x_t = \arg \max_{x \in \mathcal{X}} \alpha(x)$$
Gaussian Process Optimization (GP): Prior Work

- Zero mean Gaussian prior with known covariance function
- Update the posterior $P_t$ based on $x_{1:t-1}, y_{1:t-1}$
- Query point according to acquisition rule:
  $$x_t = \arg \max_{x \in \mathcal{X}} \alpha(x)$$
- $\alpha(x)$ is the utility of querying $x$ balances exploration and exploitation
- Commonly used $\alpha(\cdot)$:
  - Probability of Improvement: $\alpha_{PI}(x) = P_t(f(x) > \tau)$
  - Expected Improvement: $\alpha_{EI}(x) = \mathbb{E}_t[(f(x) - \tau)1_{\{f(x) > \tau\}}]$ 
  - Upper Confidence Bound: $\alpha_{UCB}(x) = \mu_t(x) + \beta_n \sigma_t(x)$
Gaussian Process Optimization (GP): Prior Work

- Zero mean Gaussian prior with known covariance function
- Update the posterior $P_t$ based on $x_{1:t-1}, y_{1:t-1}$
- Query point according to acquisition rule:
  \[ x_t = \arg \max_{x \in \mathcal{X}} \alpha(x) \]
- $\alpha(x)$ is the utility of querying $x$ balances exploration and exploitation
- Commonly used $\alpha(\cdot)$:
  - Probability of Improvement: $\alpha_{PI}(x) = P_t(f(x) > \tau)$
  - Expected Improvement: $\alpha_{EI}(x) = \mathbb{E}_t[(f(x) - \tau)1_{\{f(x) > \tau\}}]$  
  - Upper Confidence Bound: $\alpha_{UCB}(x) = \mu_t(x) + \beta_n \sigma_t(x)$
- Non-convex optimization w many local maxima
Gaussian Process Optimization (GP): Prior Work

- Zero mean Gaussian prior with known covariance function
- Update the posterior $\mathbb{P}_t$ based on $x_{1:t-1}, y_{1:t-1}$
- Query point according to acquisition rule:
  \[ x_t = \arg \max_{x \in \mathcal{X}} \alpha(x) \]
- $\alpha(x)$ is the utility of querying $x$ balances exploration and exploitation
- Practically we rely on a discretization
  \[ x_t = \arg \max_{x \in \mathcal{X}} \alpha(x) \]
Gaussian Process Optimization (GP): Prior Work

- Zero mean Gaussian prior with known covariance function

- Update the posterior $P_t$ based on $x_{1:t-1}, y_{1:t-1}$

- Query point according to acquisition rule:
  \[ x_t = \arg \max_{x \in \mathcal{X}} \alpha(x) \]

- $\alpha(x)$ is the utility of querying $x$ balances exploration and exploitation

- Practically, we rely on a discretization $\{\mathcal{X}_t\}_t$
  \[ x_t = \arg \max_{x \in \mathcal{X}_t} \alpha(x) \]

- Prior work: off-line discretization $\{\mathcal{X}_t\}_t$ with $|\mathcal{X}_t| = \mathcal{O}(t^D)$
Prior Work: Information-type Regret Bounds

- Existing bounds on $\mathcal{R}_n$ have the general form:

  $$\mathcal{R}_n \leq \mathcal{O}(\sqrt{n\gamma_n \log n})$$  \hspace{1cm} (1)

- Here $\gamma_n$ is the maximum information gain from $n$ observations

  $$\gamma_n = \sup_{S \subseteq \mathcal{X} : |S| = n} I(y_S; f)$$  \hspace{1cm} (2)

- For specific kernels, bounds on $\gamma_n$ can be obtained

  $$\gamma_n^{\text{Matérn}}(\nu) = \begin{cases} 
  \mathcal{O}(n^{\frac{D(D+1)}{2(D+1)+2\nu}} \log n) & \nu > 1 \\
  \mathcal{O}(\log n)^{D+1} & \nu = \infty 
  \end{cases}$$
Prior Work: Information-type Regret Bounds

- Existing bounds on $\mathcal{R}_n$ have the general form:

$$\mathcal{R}_n \leq \mathcal{O}(\sqrt{n\gamma_n \log n})$$  \hspace{1cm} (1)

- Here $\gamma_n$ is the maximum information gain from $n$ observations

$$\gamma_n = \sup_{S \subseteq \mathcal{X} : |S| = n} I(y_S; f)$$  \hspace{1cm} (2)

- $\gamma_n$: maximum information about $f$, and not necessarily $x^*$.

- For specific kernels, bounds on $\gamma_n$ can be obtained

$$\gamma_n^{\text{Matérn}}(\nu) = \begin{cases} 
\mathcal{O}(n^{\frac{D(D+1)}{2D+2} + 2\nu} \log n) & \nu > 1 \\
\mathcal{O}(\log n)^{D+1} & \nu = \infty
\end{cases}$$
A toy example

Suppose $\mathcal{X} = [0, 1]$ and let $f : [0, 1] \to \mathbb{R}$ be a sample from:

$$f(x) = \sum_{i=1}^{\infty} a_i X_i \left( \psi(3^i x - 1) - \psi(3^i x - 2) \right)$$

$$\psi(x) = 1 - 4(x - 0.5)^2$$

where $(a_i)_{i \geq 1}$ non-increasing positive constants and $X_i \sim \mathcal{N}(0, 1)$. 

![Graph of f(x)](image)
1. Algorithmic Improvement: Adaptive discretization for GP
   - Opportunistically adapts to the (simple) structure of $f$
   - Strictly lower complexity ($O(D)$) for $\mathcal{X} \subset \mathbb{R}^D$

2. Analytic Improvement: Dimensional-type regret bound
   - As good or better regret bound than prior work
   - First sublinear bound for exponential kernels (Matérn-$\nu = 1/2$)
   - Strictly tighter bounds for Matérn kernels if $D > \nu - 1$

GP Optimization with Adaptive Discretization

Idea: Piecewise constant upper-bound confidence function
GP Optimization with Adaptive Discretization

**Idea:** Piecewise constant upper-bound confidence function

- Rely on two bounds:
  - Compute UCB on the function value at a point.
  - Upper bound the variation of the function in a region.

- Maximally select the best region (piecewise constant UCB)
GP Optimization with Adaptive Discretization

**Idea:** Piecewise constant upper-bound confidence function

- Rely on two bounds:
  - Compute UCB on the function value at a point.
  - Upper bound the variation of the function in a region.
GP Optimization with Adaptive Discretization

Idea: Piecewise constant upper-bound confidence function

- Rely on two bounds:
  - Compute UCB on the function value at a point.
  - Upper bound the variation of the function in a region.

- Maximally select the best region (piecewise constant UCB)
Idea: Piecewise constant upper-bound confidence function

- Rely on two bounds:
  - Compute UCB on the function value at a point.
  - Upper bound the variation of the function in a region.

- Maximally select the best region (piecewise constant UCB)

- Refine the discretization after sufficient number of observations in a region.
**Idea:** Piecewise constant upper-bound confidence function

- Rely on two bounds:
  - Compute UCB on the function value at a point.
  - Upper bound the variation of the function in a region.

- Maximally select the best region (piecewise constant UCB)

- Refine the discretization after sufficient number of observations in a region.
Algorithm 1: Tree-based Adaptive Discretization

- Work with a fixed tree of partitions (with fan-out $N$):
  - Increasingly refined subsets $\mathcal{X}_h = \{x_{h,i} : 1 \leq i \leq N^h\}$
  - for each $x_{h,i}$, we have a cell
    $$\mathcal{X}_{h,i} = \{x \in \mathcal{X} : l(x, x_{h,i}) \leq l(x, x_{h,j}) \quad \forall j \neq i\}$$

- Assume cells $\mathcal{X}_{h,i}$ satisfy:
  - The radius of $\mathcal{X}_{h,i}$ geometrically decaying with $h$
    $$B(x_{h,i}, \gamma \rho^h) \subset \mathcal{X}_{h,i} \subset B(x_{h,i}, \gamma^{-1} \rho^h)$$
  - For a fixed $h$, $\cup_i \mathcal{X}_{h,i} = \mathcal{X}$
Algorithm-1: Tree-based Adaptive Discretization

- Work with a fixed tree of partitions (with fan-out $N$):
  - Increasingly refined subsets $\mathcal{X}_h = \{x_{h,i} : 1 \leq i \leq N^h\}$
  - for each $x_{h,i}$, we have a cell

$$\mathcal{X}_{h,i} = \{x \in \mathcal{X} : l(x, x_{h,i}) \leq l(x, x_{h,j}) \quad \forall j \neq i\}$$

- Assume cells $\mathcal{X}_{h,i}$ satisfy:
  - The radius of $\mathcal{X}_{h,i}$ geometrically decaying with $h$

$$B(x_{h,i}, \gamma \rho^h) \subset \mathcal{X}_{h,i} \subset B(x_{h,i}, \gamma^{-1} \rho^h)$$

  - For a fixed $h$, $\bigcup_i \mathcal{X}_{h,i} = \mathcal{X}$

- Can be constructed if $\mathcal{X} = [a, b]^D$
Algorithm-1: Tree-based Adaptive Discretization

- In round $t$, maintain a set of leaf nodes $\mathcal{L}_t$ partitioning $\mathcal{X}$
Algorithm-1: Tree-based Adaptive Discretization

- In round $t$, maintain a set of leaf nodes $L_t$ partitioning $\mathcal{X}$

- Point $x_{h,i}$ represents the center of cell $\mathcal{X}_{h,i}$ and $V_h$ is a h.p.u.b on the maximum function variations
Algorithm-1: Tree-based Adaptive Discretization

- In round $t$, maintain a set of leaf nodes $\mathcal{L}_t$ partitioning $\mathcal{X}$

- Point $x_{h,i}$ represents the center of cell $\mathcal{X}_{h,i}$ and $V_h$ is a h.p.u.b on the maximum function variations

- Select a node/point from $\mathcal{L}_t$ by maximizing index

$$I_t(x_{h,i}) = \bar{U}_t(x_{h,i}) + V_h,$$
Algorithm-1: Tree-based Adaptive Discretization

- In round $t$, maintain a set of leaf nodes $L_t$ partitioning $\mathcal{X}$

- Point $x_{h,i}$ represents the center of cell $\mathcal{X}_{h,i}$ and $V_h$ is a h.p.u.b on the maximum function variations

- Select a node/point from $L_t$ by maximizing index $I_t(x_{h,i}) = \bar{U}_t(x_{h,i}) + V_h$,

  - Refine: If $\beta_n \sigma_{t-1}(x_{h,i}) \leq V_h$, then the node $\mathcal{X}_{h_t,i_t}$ is expanded into and replaced by its $N$ children nodes
Algorithm-1: Tree-based Adaptive Discretization

- In round $t$, maintain a set of leaf nodes $\mathcal{L}_t$ partitioning $\mathcal{X}$

- Point $x_{h,i}$ represents the center of cell $\mathcal{X}_{h,i}$ and $V_h$ is a h.p.u.b on the maximum function variations

- Select a node/point from $\mathcal{L}_t$ by maximizing index

  $$I_t(x_{h,i}) = \bar{U}_t(x_{h,i}) + V_h,$$

  - **Refine**: If $\beta_n \sigma_{t-1}(x_{h_{t-1},i_{t}}) \leq V_h$, then the node $\mathcal{X}_{h_{t},i_{t}}$ is expanded into and replaced by its $N$ children nodes

  - **Evaluate**: Otherwise, observe the noisy function value $y_t = f(x_{h_{t},i_{t}}) + \eta_t$ and update the posterior distribution of $f$
Algorithm-1: Tree-based Adaptive Discretization

- In round $t$, maintain a set of leaf nodes $\mathcal{L}_t$ partitioning $\mathcal{X}$

- Point $x_{h,i}$ represents the center of cell $\mathcal{X}_{h,i}$ and $V_h$ is a h.p.u.b on the maximum function variations

- Select a node/point from $\mathcal{L}_t$ by maximizing index $I_t(x_{h,i}) = \bar{U}_t(x_{h,i}) + V_h$

  - Refine: If $\beta_n \sigma_{t-1}(x_{h,t,i_t}) \leq V_h$, then the node $\mathcal{X}_{h,t,i_t}$ is expanded into and replaced by its $N$ children nodes

  - Evaluate: Otherwise, observe the noisy function value $y_t = f(x_{h,t,i_t}) + \eta_t$ and update the posterior distribution of $f$

- Complexity: $\mathcal{O}(N n^4 \log n + N D n \log n)$
Algorithm-1: Regret Bounds

Theorem-1

Under mild technical conditions on $K(\cdot, \cdot)$, with high probability we have that

$$
\mathcal{R}_n = \sum_{i \leq n} f(x^*) - f(x_t) \leq \min \{ \mathcal{O}(\sqrt{n} \gamma_n \log n), \mathcal{O}(n^{1-\frac{\alpha}{\tilde{D} + 2\alpha}}) \}
$$

$$
\mathcal{S}_n = f(x^*) - f(x(n)) \leq \tilde{O}(n^{-\alpha/(\tilde{D} + 2\alpha)})
$$

- $\tilde{D}$ is a notion of dimension of the near optimal regions of $f$
  - Regret bound is a random variable, a function of dimensionality of $f$ around its maxima
  - For almost all realization of $f$, $\tilde{D} \leq D$
Improved bounds for Matérn kernels

- Matérn kernels parameterized by $\nu = m + 1/2$:

$$K_\nu^{\text{Matérn}}(x, x') = K(0)(1 + \sum_{i=1}^{m} a_i ||x-x'||^i) e^{-c_1 \sqrt{\nu} ||x-x'||}$$

- Improved bound for all when $D \geq \nu - 1$
Improved bounds for Matérn kernels

- Matérn kernels parameterized by $\nu = m + 1/2$:

$$K^\text{Matérn}_{\nu}(x, x') = K(0)(1+\sum_{i=1}^{m} a_i \|x-x'\|^\nu) e^{-c_1 \sqrt{\nu} \|x-x'\|}, \quad \left\{ \begin{array}{l} \nu = 1/2 \\ \nu = 3/2 \\ \nu = 5/2 \end{array} \right.$$ 

- Improved bound for all when $D \geq \nu - 1$
Improved bounds for Matérn kernels

Matérn kernels parameterized by $\nu = m + 1/2$:

$$K^{\text{Matérn}}_{\nu}(x, x') = K(0) \left(1 + \sum_{i=1}^{m} a_i ||x - x'||^i \right) e^{-c_1 \sqrt{\nu} ||x - x'||},$$

When $\mathcal{X} \subset [0, 1]^D$ and $\nu > 2$, $\tilde{D} \leq 3D/4$ with high prob.

Improved bound for all when $D \geq \nu - 1$
Improved bounds for Matérn kernels

- Matérn kernels parameterized by $\nu = m + 1/2$:

$$K_{\nu}^{\text{Matérn}}(x, x') = K(0)(1 + \sum_{i=1}^{m} a_i \|x - x'\|^i) e^{-c_1 \sqrt{\nu} \|x - x'\|},$$

- Our bounds improve on the existing bounds in two ways:
  - For $\nu = 1/2$, we provide the first explicit sublinear bounds on cumulative regret.
  - For $\nu = 3/2$ and $\nu = 5/2$ our bounds are tighter for $D \geq 2$.

- Improved bound for all when $D \geq \nu - 1$
Algorithm-1: Regret Bounds for Noiseless Observations

Theorem-2

If in addition to the assumptions of Theorem-1, we further assume $\sigma = 0$. With high probability,

$$\mathcal{R}_n \leq \tilde{O}(n^{1-\frac{\sigma}{\tilde{D}}})$$

(3)

$$\mathcal{S}_n \leq \tilde{O}(n^{-\alpha/\tilde{D}})$$

(4)

if $\tilde{D} > 0$, and

$$\mathcal{R}_n \leq \tilde{O}(1)$$

(5)

$$\mathcal{S}_n \leq \tilde{O}(e^{-c_1 \log(1/\rho)n})$$

(6)

if $\tilde{D} = 0$ and $h_{\max} = \Omega(n)$, for some constant $c_1 > 0$. 

150.
Comparison with BaMSOO

- Our tree algorithm motivated by Bayesian Multi-Scale Optimistic Optimization (BaMSOO)
  - BaMSOO relies on an adaptive construction of a partition tree
  - BaMSOO only works with noiseless observations

Our method has some advantages:

- BaMSOO’s regret analysis only under very restrictive conditions on $K$, e.g. excludes Matérn $\nu = 1/2$

- BaMSOO has strictly worse simple regret, $S_n$, for some fairly practical cases
Summary of Results

1. Algorithmic Contribution: Adaptive discretization for GP
   ▶ Opportunistically adapts to the (simple) structure of $f$
   ▶ Strictly lower complexity ($O(D)$) for $\mathcal{X} \subset \mathbb{R}^D$

2. Analytic Contribution: Dimensional-type regret bound
   ▶ As good or better regret bound than prior work
   ▶ First sublinear bound for exponential kernels (Matérn-$\nu = 1/2$)
   ▶ Strictly tighter bounds for Matérn kernels if $D > \nu - 1$
Summary & Extensions

Information Acquisition and Evolution of Belief Vector

Generalized notions of rate and reliability to acquisition rate--reliability

Establishing acquisition rate--reliability trade-off reminiscent of that of codes

Uncertainty measure beyond entropy

More dynamic notion of uncertainty

Converses that account for the unpredictable component of the state
DetecDrone

Drones that Actively Seek Information and Learn

Intelli-Ranch
- Camera Enabled Drones
- Monitoring Livestock
- Alarm and Rescue
- Cross-referencing

Wide Area Object Tracking
- Camera Enabled Drones
- Flight Path Optimization for Maximum Battery Life
- Multi-resolution Mapping

Assisted Living
- Camera Enabled Drones
- Voice Activation
- Augmented Sensing with Easy Voice Control

155.
Information Acquisition and Active Learning

Tara Javidi
University of California
San Diego

Mohammad Naghsvar

Sung-En Chiu
Anusha Lalitha
Yongxi Lu
Nancy Ronquillo
Shubhanshu Shekhar
Ziyao Tang
Songbai Yan

Kamalika Chaudhuri
Yonatan Kaspi
Ofer Shayevitz