Information-theory based policies for learning in closed sensori-motor loops

Friedrich T. Sommer
UC Berkeley - Redwood Center for Theoretical Neuroscience

Lecture 2022
WHAT MAKES US EXPLORE AND PLAY?

THE SCIENTIST IN THE CRIB
WHAT EARLY LEARNING TELLS US ABOUT THE MIND

Alison Gopnik, Ph.D.
Andrew N. Meltzoff, Ph.D.
Patricia K. Kuhl, Ph.D.
Talk Outline and Collaborators

1) Intro: Theories on information-driven exploration

2) Exploration based on predicted information gain (PIG)

3) PIG exploration in unbounded state spaces

Daniel Little  Shariq Mobin  James Arnemann
WHY DO WE EXPLORE?
REINFORCEMENT LEARNING PERSPECTIVE

Diagram:
- State: $S_t$
- Reward: $R_t$
- Agent
- Environment
- Action: $A_t$
- Next state: $S_{t+1}$
- Next reward: $R_{t+1}$
WHY DO WE EXPLORE?
REINFORCEMENT LEARNING PERSPECTIVE

Exploitation

Exploration
WHY DO WE EXPLORE?
PERSPECTIVE OF OPTIMAL DESIGN – ACTIVE LEARNING

- Theory for learning in the brain (from little data)
- Basis for building machines that can autonomously learn
- Balance: “modeling the predictable” versus “discovering the novel”
TIME LINE OF EARLY THEORETICAL WORK

Kirstine Smith 1918: Optimal experimental design
On the standard deviations of adjusted and interpolated values of an observed polynomial function and its constants and the guidance they give towards a proper choice of the distribution of observations (Biometrika 1918)

D.V. Lindley 1956: Bayesian definition of information gain
On a measure of information provided by an experiment (Annals of Math. Stat. 1956)

E. Pfaffelhuber 1972: Definition of missing information
Learning and information theory (Intern, J. Neurosci. 1972)

MacKay 1992:
Information-based objective functions for active data selection (Neural Comp. 1992)

Oaksford & Chater 1994:
Expected information gain in selection tasks
THE BAYESIAN VIEW

Parameterized probabilistic model of observation:

\[ p(x) = \int p(x \mid \theta) p(\theta) d\theta \]

Information gain by individual observation (Lindley 56):

\[ I(x) = \int p(\theta \mid x) \log p(\theta \mid x) d\theta - \int p(\theta) \log p(\theta) d\theta \]

with prior \( p(\theta) \) and posterior \( p(\theta \mid x) \) of parameter \( \theta \).
Average information gain (Lindley 56):

\[
I = E_x I(x)
\]

\[
= \int dx p(x) \int p(\theta \mid x) \log p(\theta \mid x) \, d\theta - \int p(\theta) \log p(\theta) \, d\theta
\]

\[
= KL[p(x, \theta) \parallel p(x)p(\theta)]
\]

\[
= E_x KL[p(\theta \mid x) \parallel p(\theta)]
\]

\[
= E_\theta KL[p(x \mid \theta) \parallel p(x)]
\]

Note that:

\[
I = H(\theta \mid x) - H(\theta) = E_x KL[p(\theta \mid x) \parallel p(\theta)]
\]
Information gain for active data selection (MacKay 92):

Sampling observation history: \( h_N = \{a_i, s_i\}, i = 1, \ldots, N \)

Estimated posterior distribution: \( \hat{p}_N(\theta) = p(\theta \mid h_N) \)

(Bayesian) information gain:

\[
I = E_{s_{N+1}} KL[\hat{p}_{N+1}(\theta) \mid \hat{p}_N(\theta)] = E_{s_{N+1}} (H_N - H_{N+1})
\]

Past BIG was renamed Bayesian surprise (Baldi & Itti, 2004)
Estimate of ground truth probability:
\[ \hat{p}(x) \approx p(x) \]

Missing information of current estimate (Pfaffelhuber 72):
\[ I_M(\hat{p}) = KL[p(x) \| \hat{p}(x)] \]

Information gain:
\[ I = I_M(\hat{p}_N) - I_M(\hat{p}_{N+1}) \]

Past IG “curiosity” based policies in agents (Storck et al. 1995)
Predictive IG = PIG (Little & Sommer 2011, 2013)
MODEL OF THE ENVIRONMENT

Controllable Markov Chain

\( \{A, S, p(s' | a, s)\} \)

Learning task:

Estimate: \( \hat{p}(s' | a, s) = \hat{\Theta}_{a,s,s'} \approx p(s' | a, s) = \Theta_{a,s,s'} \)
THREE TEST ENVIRONMENTS
LEARNING ACROSS A RANGE OF STRUCTURES

Dense Worlds

Mazes

1-2-3 Worlds

\[ f(\Theta_{a,s,s'}) = \text{Dir}(\alpha) = \frac{1}{Z(\alpha)} \prod_{s'} (\Theta_{a,s,s'})^{\alpha_{s'} - 1} \]
LEARNING-DRIVEN EXPLORATION
TWO SEPARATE CHALLENGES

1. Inference: Given a set of data, what is the best estimate $\hat{\Theta}$ of $\Theta$?

2. Exploration: Given $\hat{\Theta}$, how should an agent choose actions to best improve the estimate?
1. INFERENCE

Missing information for entire CMC:

\[ I_M = \sum_{a,s} \alpha_{as} KL[\Theta_{a,s,s'} \parallel \hat{\Theta}_{a,s,s'}] \]

Theorem for inference step:
Bayesian inference minimizes missing information:

\[ \hat{\Theta} = E_{\Theta|h}[\Theta] = \arg \min_{\Phi} E_{\Theta|h}[I_M(\Phi)] \]
LEARNING DURING EXPLORATION
MISSING INFORMATION IS UNEVENLY DISTRIBUTED

Control Strategies:
Random Action - an undirected baseline learner
Unembodied - an upper bound on learning
COMPARING CONTROLS
THE EMBODIMENT CONSTRAINTS ON LEARNING
2. OBJECTIVE FOR EXPLORATION

PREDICTED INFORMATION GAIN

1. Use current model to predict next sensory input: $\hat{\Theta}_{a,s,s'}$

2. Add fictive new observation $s'$ to current model: $\hat{\Theta}_{a,s \rightarrow s'}$

3. Compute predicted information gain:

$$PIG(a,s) = E_{s',\Theta|h}[KL[\Theta \parallel \hat{\Theta}] - KL[\Theta \parallel \hat{\Theta}_{a,s \rightarrow s'}]]$$

$$= \sum_{s'} \hat{\Theta}_{a,s,s'} KL[\hat{\Theta}_{a,s \rightarrow s'} \parallel \hat{\Theta}]$$
PREDICTED INFORMATION GAIN (PIG)
ACCURATE ESTIMATION OF LEARNING VALUE
GREEDY MAXIMIZATION OF PIG IMPROVES LEARNING IN 1-2-3 WORLDS
Forward search for optimal policy has exponential complexity.

Use value iteration (Bellman 1957)

\[
Q_0(a,s) = PIG(a,s)
\]

\[
Q_{\tau-1}(a,s) = PIG(a,s) + \eta \sum_{s' \in S} \hat{\Theta}_{ass'} V_{\tau}(s')
\]

\[
V_{\tau}(s') = \max_a Q_{\tau}(a,s)
\]
VALUE ITERATED MAXIMIZATION OF PIG
CLOSING THE EMBODIMENT GAP
PREVIOUS EXPLORATION STRATEGIES
PIG(VI) OUTPERFORMS ALTERNATIVES IN STRUCTURED WORLDS

Previous Strategies:
Least Taken Action (LTA) - Si, Herrmann, and Pawelzik 2007
Counter Based (CB) - Thrun 1992
Q-Learning on Past Change (PC(Q)) - Storck, Hochreiter, Schmidhuber 1995
GENERAL UTILITY OF EFFICIENT LEARNING
INDEPENDENT GOAL-DIRECTED TASKS
Chinese Restaurant Process – CRP-PIG

\[ p_i(C_t) = \frac{c_i}{t + \theta}, i = 1,\ldots, K_t \]

\[ p_\psi(C_t) = \frac{\theta}{t + \theta} \]

Empirical Bayes Version – EB-CRP-PIG

\[ \theta(t) \approx \frac{K_t}{\ln(t) + \gamma + \frac{1}{2t} - \frac{1}{12t^2}} \]

with \( \gamma \approx 0.577 \) Euler’s Mascheroni constant
PIG IN UNBOUNDED STATE SPACE
Mobin, Arnemann, Sommer, NIPS 2014

Figure 2: Bounded Maze environment. Two transition distributions, $\pi_{sa}$, are depicted, one for $(s=13, a='left')$ and one for $(s=9, a='up')$. Dark versus light gray arrows represent high versus low probabilities. For $(s=13, a='left')$, the agent moves with highest probability left into a transporter (blue line), leading it to the absorbing state 29 (blue concentric rings). With smaller probabilities the agent moves up, down or is reflected back to its current state by the wall to the right. The second transition distribution is displayed similarly.

Figure 3: Missing Information vs. Time for EB-CRP-PIG and several other strategies in the bounded maze environment.

To directly assess how efficient learning translates to the ability to harvest reward, we consider the 5-state "Chain" problem [19], shown in Figure 4, a popular benchmark problem. In this environment, agents have two actions available, $a$ and $b$, which cause transitions between the five states. At each time step the agent "slips" and performs the opposite action with probability $p_{\text{slip}} = 0.2$. The agent receives a reward of 2 for taking action $b$ in any state and a reward of 0 for taking action $a$ in...
Figure 4: Chain Environment.

every state but the last, in which it receives a reward of 10. The optimal policy is to always choose action $a$ to reach the highest reward at the end of the chain, it is used as a positive control for this experiment. We follow the protocol in previous publications and report the cumulative reward in 1000 steps, averaged over 500 runs. Our agent EB-CRP-PIG-R executes the EB-CRP-PIG strategy for $S$ steps, then computes the best reward policy given its internal model and executes it for the remaining 1000-$S$ steps. We found $S=120$ to be roughly optimal for our agent and display the results of the experiment in Table 1, taking the results of the competitor algorithms directly from the corresponding papers. The competitor algorithms define their own balance between exploitation and exploration, leading to different results.

<table>
<thead>
<tr>
<th>Method</th>
<th>Reward</th>
<th>±</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAM-RMAX [5]</td>
<td>2810</td>
<td></td>
</tr>
<tr>
<td>BOSS [2]</td>
<td>3003</td>
<td></td>
</tr>
<tr>
<td>exploit [15]</td>
<td>3078</td>
<td></td>
</tr>
<tr>
<td>Bayesian DP [19]</td>
<td>3158±31</td>
<td></td>
</tr>
<tr>
<td>EB-CRP-PIG-R</td>
<td>3182±25</td>
<td></td>
</tr>
<tr>
<td>Optimal</td>
<td>3658±14</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Cumulative reward for 1000 steps in the chain environment.

The EB-CRP-PIG-R agent is able to perform the best and significantly outperforms many of the other strategies. This result is remarkable because the EB-CRP-PIG-R agent has no prior knowledge of the state space size, unlike all the competitor models. We also note that our algorithm is extremely efficient computationally, it must approximate the optimal policy only once and then simply execute it. In comparison, the exploit strategy [15] must compute the approximation at each time step. Further, we interpret our competitive edge over BOSS to reflect a more efficient exploration strategy. Specifically, BOSS uses LTA for exploration and Figure 3 indicates that the learning performance of LTA is far worse than the performance of the PIG-based models.

Figure 5: Missing Information vs. Time for EB-CRP-PIG and CRP-PIG in the unbounded maze environment.

Finally, we consider an unbounded maze environment with $|S|$ being infinite and with multiple absorbing states. Figure 5 shows the decrease of missing information (11) for the two CRP based strategies. Interestingly, like in the bounded maze the Empirical Bayes version reduces the missing information more rapidly than a CRP which has a fixed, but experimentally optimized, parameter value. What is important about this result is that EB-CRP-PIG is not only better but it requires no prior parameter tuning since $\theta$ is adjusted intrinsically. Figure 6 shows how an EB-CRP-PIG and an LTA agent explore the environment over 6000 steps. The missing information for each state is

Reduction in Missing Information in unbounded maze
PIG IN UNBOUNDED STATE SPACE

Exploration in unbounded environment

Figure 6: Unbounded Maze environment. Exploration is depicted for two different agents (a) EB-CRP-PIG and (b) LTA, after 2000, 4000, and 6000 exploration steps respectively. Initially all states are white (not depicted), which represent unexplored states. Transporters (blue lines) move the agent to the closest gravity well (small blue concentric rings). The current position of the agent is indicated by the purple arrow.

The two agents are also tested in a reward task in the unbounded environment for assessing whether the exploration of EB-CRP-PIG leads to efficient reward acquisition. Specifically, we assign a reward to each state equal to the Euclidean distances from the starting state. Like for the Chain problem before, we create two agents EB-CRP-PIG-R and LTA-R which each run for 1000 total steps, exploring for S=750 steps (defined previously) and then calculating their best reward policy and executing it for the remaining 250 steps. The agents are repositioned to the start state after S steps and the best reward policy is calculated. The simulation results are shown in Table 2. Clearly, the increased coverage of the EB-CRP-PIG agent also results in higher reward acquisition.

<table>
<thead>
<tr>
<th>Method</th>
<th>Reward</th>
</tr>
</thead>
<tbody>
<tr>
<td>EB-CRP-PIG-R</td>
<td>1053</td>
</tr>
<tr>
<td>LTA-R</td>
<td>812</td>
</tr>
</tbody>
</table>

Table 2: Cumulative reward after 1000 steps in the unbounded maze environment.
SUMMARY

1. Learning in action-perception loops is the old optimal design problem

2. State-space information gain (PIG) versus Bayesian information gain

3. Maximizing PIG minimizes missing information

4. Nongreedy optimization is critical in interesting environments

5. Extension of PIG to unbounded environments:
   - CRP+ Empirical Bayes works best
   - Surprise-based information seeking does not eliminate surprise
   - Balance between eliminating uncertainty and discovering more states is model depending
Other objectives for exploration

1. Homekinesis (Der, 2000)
Other objectives for exploration

1. Free Energy Principle  
   (Friston, 2010)

-> Dark corner problem
Other objectives for exploration

Learning Progress (Kaplan & Oudeyer, 2007)

= derivative of prediction error

Does not require information estimation and still avoids problems of policies of directly minimizing prediction error
MEASURES OF UTILITY TOWARDS LEARNING
IN THEORETICAL PSYCHOLOGY

Predicted Information Gain (PIG) (Oaksford & Chater 1994)

\[ \sum_{s'} \Theta_{a,s,s'} D_{KL}(\Theta_{a,s \rightarrow s'} || \Theta) \]

Predicted Mode Change (PMC) (Baron 2005, Nelson 2005)

\[ \sum_{s'} \Theta_{a,s,s'} \left[ \max_{s^*} \Theta_{a,s \rightarrow s'} - \max_{s^*} \Theta_{a,s,s^*} \right] \]

Predicted L1 Change (PLC) (Klayman & Ha 1987)

\[ \sum_{s'} \Theta_{a,s,s'} \left[ \sum_{s^*} \left| \Theta_{a,s \rightarrow s'} - \Theta_{a,s,s^*} \right| \right] \]