On Balancing Exploration vs. Exploitation on a Cognitive Engine for Multi-Antenna Systems

Haris I. Volos and R. Michael Buehrer

Motivation and Goals

- Multi-antenna systems have to select between several options on how to use their antenna resources.
- A cognitive engine (CE) [1] can learn and optimize the radio’s resources.
- The CE learns by analyzing prior performance data.
- When the CE has limited knowledge, it has to choose between using a configuration that is proven to work and a more promising configuration with unknown performance at the given channel conditions.
- This is the classical problem of balancing exploration vs. exploitation. Exploration refers to trying a configuration that potentially can yield better performance, exploitation refers to using a configuration with known performance.
- In this work, the performance metric (return) is the achievable capacity of the system.

The Multi-Armed Bandit Problem

- The exploration vs. exploitation problem is often studied using the Multi-Armed Bandit (MAB) framework.
- The MAB problem assumes that one has to choose between K machines that each yield an unknown reward $R$. Based on an underlying distribution.
- The goal is to find a policy that maximizes the expected return $V(s)$, where $s$ is a belief state about the machines’ return distributions:
  \[
  V(s) = E_s \sum_{n=1}^{N} \gamma^n R_n
  \]
  where
  - $E_s$ the expectation operation over $s$.
  - $N$ the maximum number of plays, $n$ the play number.
  - $\gamma$ a discount factor, $0 < \gamma < 1$.
  - $R_n$ the return at time $n$.

Exploration vs. exploitation strategies seek to maximize $V(s)$; in this work we consider two key strategies:
- The epsilon-greedy strategy
- The Gittins’ Index

The Epsilon-Greedy Strategy

- The epsilon-greedy strategy simply exploits (selects the best known configuration) $1-\epsilon$ of the time, and explores (randomly selects a configuration) $\epsilon$ of the time.
- The benefit of the epsilon-greedy strategy is its simplicity. The drawbacks is that is not guaranteed to be optimal and that it may suffer when the number of configurations is very large.
- To improve the epsilon-greedy strategy, configurations that have a capacity exceeding the capacity $C_{\text{max}}$ for a MIMO system, where excluded. $C_{\text{max}}$ (with not TX side channel info.) is given by:
  \[
  C_{\text{max}} = \min\{N_r M_t\} \log_2 \left( 1 + \frac{\text{SNR}}{\lambda_i} \right)
  \]
  where $N_r$ and $M_t$ are the number of TX and RX antennas respectively, and $\lambda_i$ the $i$th eigenvalue of $H H^H$, where $H$ is the $M_t \times N_r$ channel matrix.

The Gittins’ Index

- Gittins in [2] showed that the K-dimensional MAB problem can be solved by using a dynamic allocation index method that breaks the problem in a series of K one-dimensional problems.
- The optimal policy, for each belief state $s$ is to use the configuration $y$ with the highest index $v_y$:
  \[
  v_y(s) = \sup_{n \leq N} \frac{E_s \sum_{n=1}^{N} \gamma^n R_{y,n}}{E_s \sum_{n=1}^{N} \gamma^n}
  \]
  where $R_{y,n}$ is the return of the configuration $y$ at the $n$th trial.

Return Distributions

- The Gittins’ indices must be derived for underlying return distributions.
- Gittins’ derives and provides tables for the most common return distributions in [2].
- In this work we consider the Normal Return Process (NRP) and the Bernoulli Return Process (BRP).
- The BRP is applicable because the return is either 0 or equal to the capacity of the communication method used (if the packet was successfully delivered).

Results

### Return Distributions

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of Trials Performed</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>50</td>
</tr>
<tr>
<td>Discount Factor, $\gamma$</td>
<td>.5</td>
</tr>
<tr>
<td>Gittins’ Index, NRP</td>
<td>73</td>
</tr>
<tr>
<td>Gittins’ Index, BRP</td>
<td>60</td>
</tr>
</tbody>
</table>

### Exploration Parameter, $\epsilon$

<table>
<thead>
<tr>
<th>Method</th>
<th>.01</th>
<th>.1</th>
<th>.01</th>
<th>.1</th>
<th>.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon$-greedy strategy</td>
<td>.53</td>
<td>.65</td>
<td>.50</td>
<td>.70</td>
<td>.74</td>
</tr>
</tbody>
</table>

From the table above:

- The Gittins’ index methods outperform the epsilon-greedy strategy.
- The epsilon-greedy strategy has very good short term performance.

From the figure on the left:

- In this case, the Gittins’ index with NRP was found to have the best overall performance.
- Shows the same trends as the table above.

From the figure on the right:

- The epsilon-greedy method has better short-term results (<50 trials).
- The Gittins’ index NRP method has better long term performance.
- The Gittins’ index BRP method has bad initial performance, but it rapidly improves.

Conclusions

- The problem of balancing exploration vs. exploitation in a CE can be simply addressed by the epsilon-greedy method, albeit inefficiently, which was found to have good short term results.
- The Gittins’ indices method were found to have the best overall results, verifying the long term optimality of the method.

References