

Two reinforcement learning problems

Pole Balancing • Maze learning

David Keil
University of Connecticut, October, 2002

Source: Donald Michie and R. A. Chambers, "BOXES: An experiment in adaptive control," In E. Dale and D. Michie, *Machine intelligence 2*, Oliver & Boyd, 1968

The BOXES program

- Focused on "information-versus payoff dilemma"
- Solution to complexity: reduce complex game to simpler model, then solve subgames
- Pole balancing is "game against nature"
- Infinite continuous state set is reduced by quantization to discrete state set
- No optimal policy was known in control theory

Problem definition

- Inputs: either a state tuple or a failure signal
- State:
 - cart position (5 values)
 - angle of pole (5 values)
 - cart velocity (3 values)
 - angle rate of change (3 values)
- Actions: {left, right}, i.e., "bang-bang" control
- Interval from action to percept: 0.05 sec

Data used in learning

- For each of the 225 states, a "demon" keeps score of the following, based on past inputs:
- *LL*, left life: sum of weighted durations of past trials after state was exited in a left action
 - *RL*, right life
 - *LU*, left usage: weighted sum of number of left actions on entry to state in past runs
 - *RU*, right usage
 - *Target*, desired level of attainment set by supervising demon
 - $T_1 \dots T_n$, times at which state has been entered during current run

Policy details

- Decision rule maps $LL \times RL \times LU \times RU \times target \rightarrow \{left, right\}$
- Let $N = \#$ times a state has been entered this run
 $DK =$ weighting factor, favors recent inputs
 $K =$ weighting factor, favors global over local inputs
- Local totals are updated, global demon updates *target*
- $Value_L = (LL \times K \times target) / (LU + K)$; similarly for right value $Value_R$
- If $Value_L > Value_R$ then $action \leftarrow left$, else $action \leftarrow right$

Policy learning overview

- "State signal" represents percept, which contributes to state knowledge
- Policy maps from states to actions
- State value is computed using accumulated experience
- Learner adapts policy online, using this experience (*LL, LU, RL, RU, target*)

Results of [MC68]

- DK , K were adjusted experimentally to 0.99, 20.0, resp.
- “Merit,” measuring success value of a trial, is computed as weighted global-live / global-usage, i.e., roughly the number of steps per trial before failure, weighted to favor later trials
- System achieved a run of over 72,000 decisions (i.e., about 1 hour of real-time control)

Discussion

- (Michie and Chambers, 1968) was early RL research aimed at solving a difficult adaptive control problem
- Insights:
 - Use of trial and error
 - Decomposition of problem
 - Separation of exploitation and exploration, rejection of greedy action
 - Use of optimistic target performance level to encourage exploration (stretch system to explore to meet high goals)