Approximate Policy Iteration for several Environments and Reinforcement Functions

Andreas Matt and Georg Regensburger
Institute of Mathematics, Univ. of Innsbruck
EWRL-6, Nancy, France
4th-5th, September 2003
Motivation

with Simulator
Overview

• State Action Spaces
• Policy Improvement – *One Realization*
• Geometric Interpretation
• Policy Improvement - *Several Realizations*
• Approximate Policy Evaluation
• Approximate Policy Iteration
• Experiments
• Discussion
Notation

- Environment $E = (S, A, P)$
- $S$ states, $A = A(s)$ family of actions
- $P = P(- | a, s)$ family of transition probabilities on $S$

- Markov decision process (MDP) is given by $(E, R)$
- $R = R(s, a, s)$ family of Rewards in $\mathbb{R}$

- $\pi(a | s)$ probability that action $a \in A(s)$ is chosen in state $s$
- $V^\pi(s)$ (discounted) value function
- $Q^\pi(a, s)$ action-value of action $a$ in state $s$
State Action Spaces

Realizations \( (S_i, A, P_i, R_i) \)

All possible sensor values = states

\[ \pi \]

Possible actions

State Action Space \( E = (S, A) \)
Policy Improvement - One Realization

Policy Improvement Theorem

If \[ \sum_{a} Q^{\pi}(a, s) \tilde{\pi}(a | s) \geq V^{\pi}(s) \text{ for all } s \in S \]

Then \[ V^{\tilde{\pi}} \geq V^{\pi}, \ V^{\tilde{\pi}}(s) \geq V^{\pi}(s) \text{ for all } s \in S \]

Policy Iteration =
Repeat: Policy Evaluation + Policy Improvement

Usual choice \[ \tilde{\pi}(a | s) = 1 \]
\[ a = \arg\max_{a} Q^{\pi}(a, s) \]
Geometric Interpretation – One Realization

\[ \pi(a \mid s) \text{ as a point on a standard simplex} \]

\[ \sum Q^\pi(a, s) \tilde{\pi}(a \mid s) \geq V^\pi(s) \]

- \( C_{\geq}^\pi(s) \) Improving policies
- Polytope
- Vertices \( \text{vert} \left( C_{\geq}^\pi(s) \right) \)
- \( C_{>}^\pi(s) \) Strictly improving policies

\[ C_{\geq}^\pi(s) = \left\{ \tilde{\pi}(- \mid s) : \sum Q^\pi(a, s) \tilde{\pi}(a \mid s) \geq V^\pi(s) \right\} \]
Policy Improvement - Several Realizations

State Action Space
Realization \((E_1, R_1)\)
\[ \sum Q_1^\pi(a, s) \tilde{\pi}(a | s) \geq V_1^\pi(s) \]

Realization \((E_2, R_2)\)
\[ \sum Q_2^\pi(a, s) \tilde{\pi}(a | s) \geq V_2^\pi(s) \]

Find \(\tilde{\pi}\) that satisfies

- \(C_{\geq}^\pi(s)\) Improving policies
- Polytope
- Vertices \(\text{vert}(C_{\geq}^\pi(s))\)
- \(C_{>\pi}^\pi(s)\) Strictly improving policies

\[ C_{\geq}^\pi(s) = C_{1,\geq}^\pi(s) \cap C_{2,\geq}^\pi(s) \]
Policy Improvement - Several Realizations

Policy Improvement for two Realizations

Realization $(E_1, R_1)$ and Realization $(E_2, R_2)$

If $\tilde{\pi}(- | s) \in C'_\geq(s)$ for all $s \in S$

and

$\tilde{\pi}(- | s) \in C'_\geq(s)$ for at least one $s$

Then $V_1^{\tilde{\pi}} \geq V_1^\pi$, $V_2^{\tilde{\pi}} \geq V_2^\pi$

and

$V_1^{\tilde{\pi}} > V_1^\pi$ or $V_2^{\tilde{\pi}} > V_2^\pi$

That is $\tilde{\pi}$ performs better than $\pi$
Balanced Policies

π is balanced if there are no strictly improving policies:
\[ C'_\pi(s) = \emptyset \text{ for all } s \in S \]

- Stochastic policies
- Equilibrium Point
- One Environment balanced = optimal
Approximate Policy Iteration

Model-Free Case = Transition Probabilities unknown

Repeat:
1. Approximate Action-Values and Value Function
2. Improve policy with approximated values
Approximate Policy Evaluation

Approximate Action-Values and Value Function

```
repeat
    choose \( s \in S \) and \( a \in A(s) \) derived from \( \pi \)
    take action \( a \) and observe \( r \) and \( s' \)
    choose \( a' \in A(s') \) according to \( \pi \)
    \( Q(a, s) \leftarrow Q(a, s) + \alpha (r + \gamma Q(a', s') - Q(a, s)) \)
```
Approximate Policy Iteration – Several Environments

Improve policy with approximated values

repeat
  approximate $V_i^\pi$ and $Q_i^\pi$ for $i \in [n]$
  for all $s \in S$ do
    if $\text{vert}(C^\pi(s)) \neq \emptyset$ then
      choose $\pi'(\cdot | s) \in \text{vert}(C^\pi(s))$
      $\pi(\cdot | s) \leftarrow \pi'(\cdot | s)$
Experiments

- Start with random policy
- Run algorithm
- Evaluate the Value Function after each improvement step exactly using dynamic programming
Experiments with Simulator
Experiments

Obstacle avoidance

Wall following

![Graph showing normed utility over iterations for different environments and strategies.](image)
Discussion & Future

- Stochastic policies for several environments and reinforcement functions – without model
- Improve policies for new realizations
- Methods to find and choose improving vertices
- Optimistic variants – improve policies after incomplete policy evaluation steps
- Function Approximation for real robot experiments

http://mathematik.uibk.ac.at/users/rl