Incremental Policy Iteration with Guaranteed Escape from Local Optima in POMDP Planning

Marek Grześ

University of Kent

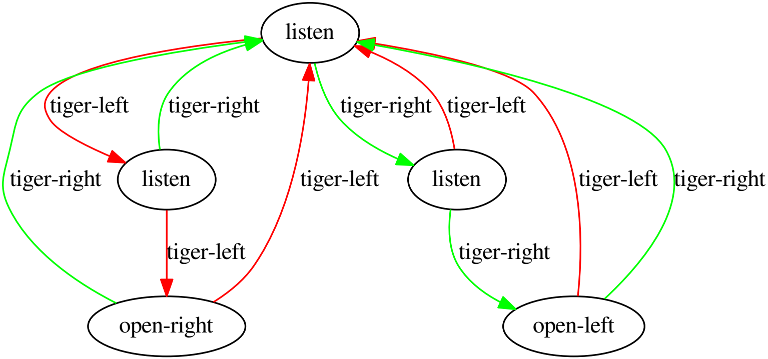
Pascal Poupart

University of Waterloo

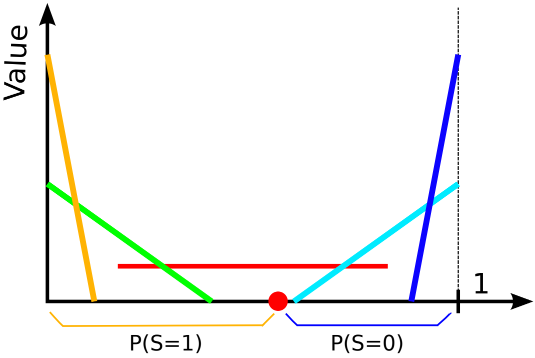
AAMAS 2015

Istanbul, May 4–8

Finite-state Controllers (FSCs) for Partially Observable Markov Decision Process (POMDPs)



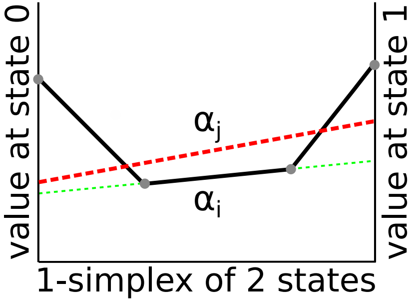
Value Function: α-vectors



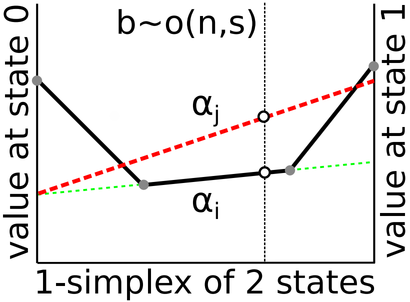
Policy Iteration:

- Compute α-vectors for a current controller
- **Use those α-vectors to improve the controller**

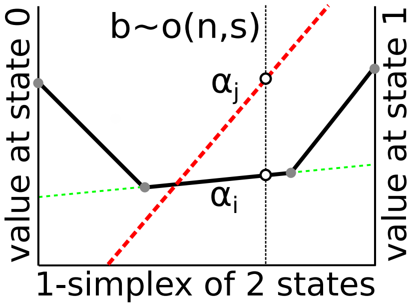
Node Improvement in Bounded Policy Iteration (BPI)



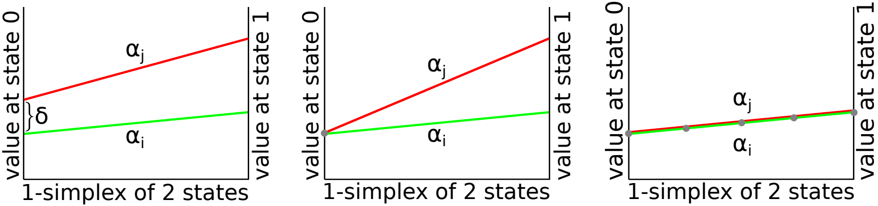
Node Improvement in Bounded Policy Iteration (BPI)



Node Improvement in Bounded Policy Iteration (BPI)



The Need to Escape Local Optima



Find a New Node with Maximal Improvement Over Entire Belief Simplex

$$\Gamma^{a,o} \leftarrow V_n^{a,o}(s) = \frac{R^a(s)}{|O|} + \gamma \sum_{s' \in S} P(o|s', a) P(s'|a, s) V_n^{\pi}(s'), \forall_n$$

$$\begin{array}{ll}
\text{max:} & \sum_{a,n',o,s} w(s) P(n', a|o) V_{n'}^{a,o}(s) - \beta \\
\text{s.t.} & \sum_s w(s) = 1; \sum_{n',a} P(n', a|o) = 1; \\
 & \forall_{a,o_1,o_2} \sum_{n_1} P(n_1, a|o_1) = \sum_{n_2} P(n_2, a|o_2) \\
 & \forall_n \beta \geq \sum_s w(s) V_n^{\pi}(s); \\
 & \forall_s w(s) \in R; \forall_{n',a,o} P(n', a|o) \in [0, 1]
\end{array}$$

- Quadratic objective
- All constraints are linear

Optimal Solution to the Escape Problem

Theorem 1: There exists an optimal solution that corresponds to a deterministic node.

How to Solve this Quadratic Programme?

- Quadratic terms are products of two probabilities—the belief state, $w(s)$, and the edge probability, $P(n', a|o)$; thus, McCormick relaxation can be applied
- Thanks to Theorem 1, McCormick relaxation finds an optimal, deterministic node
- But, McCormick relaxation leads to a mixed-integer linear programme (MILP) which is intractable
- Fortunately, we don't need an optimal solution to our MILP; solutions that yield a non-trivial improvement at $w(s)$ will eventually help the policy iteration algorithm
- Even a linear relaxation of our MILP can be sufficient (see the paper for interesting properties)

Heuristic Tricks to Avoid Heavy Guns (i.e. CPLEX)

- One-step lookahead (on-policy and off-policy)
- Node splitting
- Checking corners

Some Results from our Paper

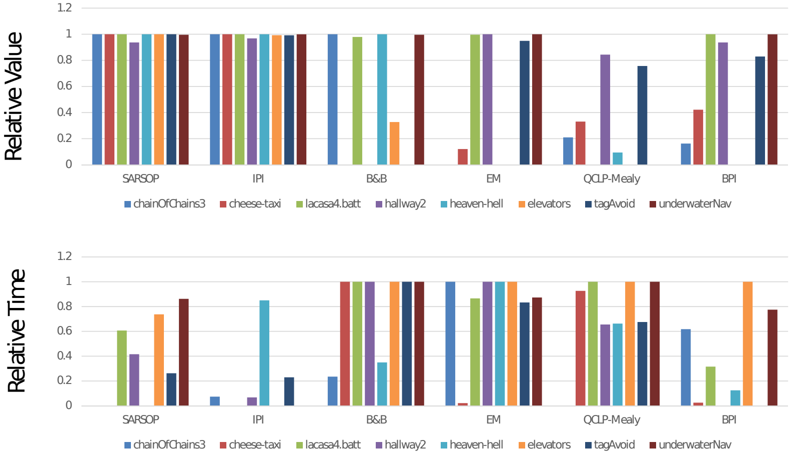


Figure : **Relative values:** normalised values so that the SARSOP upper bound is 1 and the worst value achieved by any algorithm is 0. **Relative time:** normalised time where the longest time is 1 and the shortest time is 0.

Conclusion

1. A new view on principled methods for policy iteration in POMDPs
2. A new efficient method for improving individual nodes
3. An intuitive explanation of local optima and challenges in escaping it
4. A guaranteed method for escape that facilitates fast, anytime execution
5. Deterministic nodes appear to be sufficient for node improvement, and the best node for escape is deterministic too
6. Heuristic methods analysed (with new connections identified—node splitting vs. node improvement) and used in a practical and well-justified manner