Regionalized Policy Representation for Reinforcement Learning in POMDPs

Xuejun Liao [1] xjliao@ee.duke.edu
Hui Li [1] hl1@ee.duke.edu
Ronald Parr [2] parr@cs.duke.edu
Lawrence Carin [1] lcarin@ee.duke.edu

[1] *Department of Electrical & Computer Engineering*
Duke University, 130 Hudson Hall, Box 90291
Durham, NC 27708, USA

[2] *Department of Computer Science*
Duke University, LSRC / Box 90129
Durham, NC 27708, USA

Many decision-making problems can be formulated in the framework of a partially observable Markov decision process (POMDP) [5]. The optimality of decisions relies on the accuracy of the POMDP model as well as the policy found for the model. In many applications the model is unknown and learned empirically based on experience, and building a model is just as difficult as finding the associated policy. Since the ultimate goal of decision making is the optimal policy, it is advantageous to learn an optimal policy directly from experience, without an intervening stage of model learning.

A major difficulty in direct policy learning arises from the fact that the belief state, which summarizes the history, is not available when the POMDP model is unknown. Carrying a long history of actions and observations is cumbersome and inefficient. To solve the representational issue, many methods have been proposed to compress the history and express it in a compact manner, these including reactive policies, history truncation [4, 3], finite policy graphs [8], finite state controllers [1], utile distinction HMMs [10], and recurrent neural networks [2].

We introduce the *regionalized policy representation* (RPR), a parametric framework for representing a stochastic policy in the absence of a POMDP model. The RPR expresses the policy as a distribution over actions given the history of actions and observations. The dynamics of decision states are driven jointly by actions and observations, the action-dependence implementing the control in the world-state space and the observation-dependence reflecting the agent's perception of the world-state.

We employ an off-policy method [9] to learn the parameters of an RPR, using a soft max exploration policy to ensure full exploration. The experience from the behavior policy is used to construct the empirical value function, given the RPR parameters. We then update the RPR parameters by choosing new parameters to maximize the empirical value function. We perform *maximum-value* (MV) estimation of the RPR parameters by an iterative procedure similar to expectation maximization (EM). Our algorithm is different from conventional EM in that it maximizes a value function instead of a likelihood function. This difference gives rise to some complications technically and yet offers insights into reinforcement learning. One interesting point worth mentioning about our EM-like procedure is that the E-step not only adjusts the posterior probability distribution of the decision state but also recomputes the expected future rewards. This update step reflects the change in future rewards when the updated RPR policy is followed.

The bound optimization nature of our algorithm makes it a more suitable choice than gradient based approaches [8, 1] for handling the hidden decision states, since it is less prone to local optima. Moreover, our formulation is amenable to Bayesian learning, which gives us a flexible framework for more general learning situations such as experience transfer and multitask learning.

We demonstrate the performance of RPRs on benchmark problem Hallway2 [6], in comparison to SARSA(λ) [7], RL-LSTM [2], and Utile distinction HMM (UDHMM) [10]. The results are summarized in Table 1, where UDHMM used more than 450 episodes and the RPR used 355 episodes. It is seen that the RPR outperforms the best competing algorithms, with roughly the same number of episodes.

References

[1] D. Aberdeen and J. Baxter. Scalable internal-state policy-gradient methods for POMDPs. In *ICML*, pages 3–10, 2002.

Table 1: A comparison of the RPR to other reinforcement learning algorithms on Hallway2

Method	Goal rate (%)	Median Steps to reach the goal
Random Walk	26	> 251
SARSA(λ) [7]	77	73
RL-LSTM [2]	94	61
Utile distinction HMM [10]	92	62
RPR	97	46

[2] B. Bakker. *The State of Mind: Reinforcement Learning with Recurrent Neural Networks*. PhD thesis, Unit of Cognitive Psychology, Leiden University, 2004.

[3] J. Baxter and P. L. Bartlett. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

[4] T. Jaakkola, S. P. Singh, and M. I. Jordan. Reinforcement learning algorithm for partially observable markov decision problems. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press, Cambridge, MA., 1995.

[5] L. Kaelbling, M. Littman, and A. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134, 1998.

[6] M. L. Littman, A. R. Cassandra, and L. P. Kaelbling. Learning policies for partially obsevable environments:scaling up. In *ICML*, pages 362–370, 1995.

[7] J. Loch and S. Singh. Using eligibility traces to find the best memoryless policy in partially observable markov decision processes. pages 141–150. San Francisco: Morgan Kaufmann, 1998.

[8] Nicolas Meuleau, Leonid Peshkin, Kee-Eung Kim, and Leslie Kaelbling. Learning finite-state controllers for partially observable environments. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 427–43, San Francisco, CA, 1999. Morgan Kaufmann.

[9] R. Sutton and A. Barto. *Reinforcement learning: An introduction*. MIT Press, Cambridge, MA, 1998.

[10] D. Wierstra and M. Wiering. Utile distinction hidden markov models. In *The Proceedings of the International Conference on Machine Learning*, 2004.

Topic: control
Preference: oral