

1 Introduction and objectives

The proposed research is focused at the intersection of artificial intelligence (AI) and a relatively new field known as reinforcement learning (RL) (Sutton & Barto, 1998; Bertsekas & Tsitsiklis, 1996). RL is a new body of theory and techniques for optimal control that has been developed over the last twenty-five years primarily within the machine learning and operations research communities, and which have separately become important in psychology and neuroscience. RL researchers have developed novel methods to approximate solutions to optimal control problems that are too large or too ill-defined to be solved by classical methods such as dynamic programming. For example, reinforcement learning methods have obtained the best known solutions in such diverse automation applications as helicopter flying, elevator scheduling, playing backgammon, and resource-constrained scheduling. RL methods have also been used in applications to manufacturing, telecommunications, finance, and power systems. RL methods have an extremely wide range of potential applicability and scale better than previous methods, yet are still limited in their reliability, theoretical guarantees, ease of use, data efficiency and computational requirements. A second class of challenges arises in trying to apply RL to the more ambitious goals of AI. RL methods appear to hold promise for AI because of their favorable scaling and potential for learning and maintaining knowledge from experience with less manual intervention. We feel there is a natural transition from the more advanced RL methods to mechanisms for knowledge representation, search, and human-level reasoning. The objectives of the proposed research program are to create new methods for RL that remove some of the limitations on its widespread application and to develop RL as a model of intelligence that is grounded in sensori-motor experience and could approach human abilities.

Modern RL research uses the optimal-control framework of Markov decision processes (MDPs), in which a controller (the learning agent) has to find (or, more realistically, has to approximate) an optimal feedback control rule (a policy for acting), usually starting with incomplete information about the dynamics of the controlled system (the agent's world or environment). Although MDPs have been studied intensively for many years in engineering and operations research, the methods developed by RL researchers have added some novel elements to classical solution methods. The most important novel element is to focus on sample-based methods such as temporal-difference (TD) learning and Monte Carlo tree search. These methods are applied to samples of system behavior and do not require a complete specification of the system. Sample-based RL methods have proven effective in approximating solutions to problems that are too large for classical methods to be feasible. A recent example coming in part out of the applicant's research group (Gelly & Silver, 2008), is the MoGo program for playing Computer Go, which recently defeated a world-class (9-dan) Go player for the first time in a game of 9x9 Go.

Despite application successes and many theoretical advances attributable to the MDP formalism, it has also become a limitation. Conventional MDP representations of states and dynamics tend to be flat and low-level. One would like to extend them to include more flexible, structured, and expressive representations supporting multiple levels of granularity and abstraction in state and time. Examples of more expressive representations can be found in the symbolic and relational representations of classical AI and in the factored and graphical representations of modern probabilistic AI. Other approaches are taken in intelligent control, in fuzzy logic, and in hierarchical approaches in both control and AI. However, a weakness of all of these approaches is that they obtain their greater expressiveness and abstraction at the cost of loosening or, more typically, breaking their connection to experience and data. Losing the connection to data makes learning harder and leaves the system more reliant on human intervention and maintenance. Much larger and more reliable AI systems could be constructed if their more abstract and expressive representations could remain grounded in their interaction with the world. In the last decade or so, new ideas for doing this have come out of RL research, and in particular out of research by the applicant and out of the applicant's previous NSERC-funded research.

2 Approach and past contributions

Conventionally, RL methods are divided into those that are model free and those that are model based. Model-free RL methods compute their value function and policy directly from experience, whereas model-based RL methods form a model of external world and use it as an intermediate step toward computing value functions and policies. The model corresponds to knowledge of the world, and in conventional RL it is low-level and flat—it is an estimate of the environment’s state-transition probabilities and expected rewards (the low-level parameters defining a specific MDP). It is the model that we seek to extend to use more abstract and expressive representations. The process of computing the optimal policy given a model is called *planning*. Planning with low-level models can be done by dynamic programming methods or by model-based RL methods such as Dyna (Sutton, 1990), or Monte Carlo tree search. In moving to more expressive representations we must consider *temporal abstraction*, *state abstraction*, and *planning with abstractions*. Each of these, and the RL techniques for it, are discussed in subsections below.

2.1 Temporal abstraction with options

Fundamental to AI and other engineering theories of control is the problem of representing knowledge about the world and about possible courses of action at a multiplicity of interrelated temporal scales. For example, a human traveler must decide which cities to go to, whether to fly, drive, or walk, and the individual muscle contractions involved in each step. In RL, the theory of *options* (Sutton, Precup & Singh, 1999) provides a generic concept of “courses of action” which includes both primitive actions such as muscle contractions and temporally extended actions such as traveling to a distant city. The theory of options is based on the theories of Markov and semi-Markov decision processes, but extends these in significant ways (see also Parr, 1998; Dietterich, 2000). Options can be used in place of actions in all of the planning and learning methods conventionally used in RL. Options and models of options can be learned for a wide variety of different subtasks, and then rapidly combined to solve new tasks. Options permit planning and learning simultaneously at a multiplicity of times scales, and toward a wide variety of subtasks, which substantially increases the efficiency and abilities of RL methods (e.g., C5¹).

Temporally extended knowledge is represented as *option models*. The model of an option gives the expected outcome of executing the option in each state where it can be applied. The outcome is in the form of a “next” state, or distribution of states, at the time the option finishes, together with the expected cumulative reward along the way. This form for the abstract knowledge is exactly as required for it to be used like low-level knowledge (state-transition probabilities) in conventional dynamic-programming planning methods. Option models are the proposed ultimate form for temporally abstract knowledge and thus the proposed primary target for learning methods. We use a terminology of *questions* and *answers*. The questions corresponding to an option model are “what will the state become if the option is executed now?” and “what will the reward be along the way?” The answers to these questions are provided by the option model. Option models are predictive and temporally abstract representations of the dynamics of the agent’s world, and in this sense they are knowledge about the cause and effect relationships in that world.

The natural way to learn option models is called *intra-option learning* because learning proceeds (by TD methods) whenever an action is taken that is consistent with the option’s policy, even if the option is not completed or even being executed at the time. Intra-option learning enables learning in parallel about many options (just as multiple Q-learning processes can learn from the same stream of random actions) and thus can be very fast. This is called *off-policy* learning and doing it reliably and efficiently in conjunction with function approximation is a major open problem in RL.

Off-policy learning Off-policy learning is learning about one way of behaving (the target policy) while

¹Citations of the C# form refer to previous research contributions in the applicant’s Form 100.

actually behaving in some other way (the behavior policy). The problem with off-policy learning is that classical TD reinforcement learning methods such as $TD(\lambda)$, Q-learning, and dynamic programming can become unstable during off-policy learning if function approximation is used. As both TD and function approximation are thought to be essential for large-scale applications, and off-policy learning is currently seen as necessary for learning temporally abstract system models, this instability is a key stumbling block to extending RL abilities.

There are methods for off-policy learning using importance sampling that are known to be stable (Precup, Sutton & Dasgupta, 2001), but these are also known to have very high variance, causing learning to be very slow. We have empirically tested and mathematically analyzed a variety of off-policy learning algorithms, including incremental least-squares methods (C14, C15), importance-sampling methods (C17), and dual-representation methods (Wang, Bowling & Schuurmans, 2007). We have developed a new concept, called a *recognizer* (C16), which achieves faster off-policy learning. A recognizer observes behavior and accepts it, or not, as something that it is learning about. It recognizes a portion of the behavior as, in effect, corresponding to the target policy. Recognizers are used to condition the predictions in temporal-difference networks (see below) on the options taken. Experiments suggest that importance sampling using recognizers has lower variance and is much better behaved than previous importance-sampling methods for off-policy learning. Most recently we have developed a new algorithm, *gradient temporal-difference learning*, or GTD, that may be a breakthrough in this area (C6). Unlike importance sampling methods, GTD is guaranteed to converge on all problems and, unlike least-squares methods, its per-step computational requirements are linear in the size of the feature vectors. GTD is the first guaranteed-convergent algorithm with this set of capabilities. We are still investigating its capabilities and practical utility.

2.2 State abstraction with predictive representations

The usual assumption in RL and in much of control theory is that the state of the system is readily observable. This assumption is reasonable for many control, gaming, and operations research problems, but not for many other problems, including the target problems in AI. In vision problems, for example, occluded objects are not perceptually available. In everyday activities we have to remember where objects are, what steps we have already done, and what we plan to do next, none of which may be immediately apparent from sensory input. The concept of state is so essential that we generally do not abandon it even if states are not readily available. Instead we use the history of what *is* available to construct an estimate of the state.

A variety of methods have been proposed for state estimation and representation when the states are not readily available, including POMDPs, history methods such as generalizations of k th-order Markov models, and *predictive state representations* (Littman, Sutton & Singh, 2002; Jaeger, 2000), a newer predictive approach that relates in many ways to the temporal abstraction ideas discussed above, and which we propose to investigate further. The standard way of addressing the hidden state problem is to study partially observable Markov decision processes (POMDPs), in which the state of the underlying MDP is not visible but must be inferred from a (typically) smaller set of observations stochastically related to the states. (POMDPs are the generalization of hidden Markov models to decision processes.) The problem with this approach is that it is strongly dependent on an accurate model of the system dynamics. In practice, history-based approaches, such as higher- or variable-order Markov models (Rissanen, 1983), are often much more effective. History-based systems immediately break symmetry, and their direct learning procedure makes them relatively simple. McCallum (1995) and others have shown in a number of examples that sophisticated history-based methods can be effective in large problems, and are often more practical than POMDP methods even in small ones.

The predictive approach to state representation is like the POMDP approach in that it updates the state

representation recursively rather than directly computing it from data. This enables it to attain generality and compactness at least equal to that of the POMDP approach. However, the predictive approach is also like the history-based approach in that its representations are grounded in data. Whereas a history-based representation looks to the past and records what did happen, a predictive representation looks to the future and represents what *will* happen. Because they are more closely tied to data, predictive state representations may be easier to learn than their POMDP counterparts, and in some cases they can be shown to generalize better (C19). A predictive state representation is based on a set of specific questions about future experience that can be compared with what actually happens. In conventional predictive state representations the questions concern the probability of specific concrete action and observation sequences. The questions are “if these actions are taken, will these observations occur?” It is natural to generalize this idea to questions about the outcomes of options as suggested in the preceding section. Predictive representations of state in which the questions are about the outcomes of options enable the representation of concepts about the long-term outcome of extended ways of behaving—a substantial generalization of prior work in RL and in some ways even going beyond conventional AI representations. For example, it is easy using option models to represent the abstract concept of “the room with the battery charger” as the expected outcome of the **dock-with-the-charger** option. Similarly, the abstract concept “chair present” could be the outcome of the **try-to-sit** option.

Temporal difference networks Temporal-difference networks (C21, C20, C18) are perhaps the final structural step in combining temporal and state abstraction while grounding knowledge in experience. This is the step of making knowledge *compositional*. This is achieved in a predictive representation by allowing the predictive questions to be about outcomes that are themselves the answers to other predictive questions. The difference is best understood with an example. Using option-based predictive representations with concrete outcomes it is easy to represent “the room with the battery charger” in terms of the outcome of the **dock-with-the-charger** option, as discussed above, because the outcome is concrete; the robot can directly sense whether or not power is trickling into its battery. But we also want to be able to represent knowledge such as “at the end of the hall is a room with a battery charger in it”. It is natural to represent this in terms of the options **follow-the-hall-to-its-end** and **dock-with-the-charger**, one composed with the other. That is, the knowledge is that at the end of **follow-the-hall-to-its-end** the robot will reach a state in which the prediction for **dock-with-the-charger** is high. The outcome is the answer to another question rather than a concrete observation.

The idea of temporal-difference networks with option-based predictive questions is a combination of several steps that are each conceptually simple, but together constitute significant complexity. The ramifications are far from clear and the extent and generality of the mechanisms designed so far has yet to be determined. One focus is on designing an off-policy learning algorithm for the general case. We have completed successful experiments with one such algorithm based on importance sampling (C17; Rafols, 2006), but the algorithm is slow to learn, largely because of the special challenges of off-policy learning discussed previously. Much more foundational work is needed to increase the efficiency of the off-policy learning or to find ways of using it so that variance does not become an issue. Creating compelling examples and expositions of our experience-based approach to artificial intelligence is also a priority for the proposed research (e.g., Koop, 2007).

2.3 Planning with approximations and abstractions

Planning refers to the use of state representations and corresponding models of the world to anticipate the consequences of alternative courses of action and to pick among them. Planning is a core topic for the research program, corresponding roughly to reasoning in people and to optimal decision making in control theory. Much of the proposed research program has been directed toward constructing flexible,

expressive, multi-scale approximate models of the world, but it is not enough to know how the world will behave; that knowledge must be used flexibly to support decision making. Planning with the kind of function approximation needed for large state spaces is largely an open problem. There are a number of possible approaches arising out of our previous work with iLSTD (C14, C15) and with Dyna-style planning systems (Sutton, 1990; Paduraru, 2006). Recently we extended Dyna-style planning to use linear function approximation in the world model, including a linear extension of prioritized sweeping (C9). Next steps will be to extend this to incorporate models based on options and TD networks. In the domain of Computer Go, we have extensively explored planning with linear function approximation in the value function (but using exact self-play for the model). Our value function is based on 1.5 million features, each corresponding to a local shape in a specific region of the board. With this technique we were able to learn the best known static evaluation function for 9x9 Go (C10).

2.4 Sensor-rich robotics

We propose to add a new robotics effort to challenge, direct, and inspire the research on grounded AI. This robot will be small and mobile and outfitted with an unusually rich set of sensors, including sensors for touch, acceleration, motion, sound, vision, and several kinds of proximity. Here we will explicitly face the challenge of relating low-level sensors and actuators, such as pixels and motor torques updated 100 times per second, to higher-level knowledge of space, objects, and people. The initial objective will be for the robot to form an extended multi-level model of the relationships among its sensors and between its sensors and its actuators. We have proposed that higher-level knowledge can be grounded in raw data of sensations and actions; this robotic platform will challenge and inspire us to see if it can really be done.

2.5 Computational models of natural learning systems

RL is studied in psychology and neuroscience as well as in engineering and artificial intelligence. In psychology, RL methods are important models of elemental learning processes in animals, such as classical conditioning (Sutton & Barto, 1981, 1990). In neuroscience, RL methods such as TD(λ) (Sutton, 1988) have independently become the dominant models of reward systems in the brain, in particular of the dopamine system (Schultz, Dayan & Montague, 1997). We propose a small component to the research program for modeling natural learning phenomena at the behavioral and neurophysiological level. Past contributions here include a new TD model of the dopamine brain-reward system, incorporating a multi-dimensional and temporally extended stimulus representation (C3,C7). This work has also led to contributions in the purely psychological literature (C1, C4).

3 Contributions to training of HQP

All of this research has been and will be done in collaboration with and furthering the training of students and postdoctoral fellows at the University of Alberta. In the five years that I have been here I have supervised and graduated six MSc students. I currently supervise or co-supervise seven PhD students, two MSc students, two postdocs, and one robotics technician. The proposed research will positively impact the training of all of these HQP.

4 Anticipated significance of the research

The proposed research will contribute to the growth of knowledge in AI, optimal control, and the understanding of natural learning systems. RL methods address ubiquitous problems of prediction and control, and as such improved methods have potential applications in a vast number of problems of scientific and economic importance. Connecting the knowledge of AI systems to experience, as to be explored in the proposed research, could enable them to be maintained, verified and even learned without human intervention. This ability may be essential to the practical development of very large AI systems.

References

- Bertsekas, D. P., Tsitsiklis, J. N. (1996). *Neural Dynamic Programming*. Athena Scientific, Belmont, MA.
- Dietterich, T. G. (2000). Hierarchical reinforcement learning with the MAXQ value function decomposition. *Journal of Artificial Intelligence Research* 13:227-303.
- Jaeger, H. (2000). Observable operator models for discrete stochastic time series. *Neural Computation*, 12, 1371–1398.
- Koop, A. (2007). Understanding experience: Temporal coherence and empirical knowledge representation. Master's thesis, Dept. of Computer Science, University of Alberta, 2007.
- Littman, M. L., Sutton, R. S., Singh, S. (2002). Predictive representations of state. *Advances in Neural Information Processing Systems 14*. MIT Press.
- McCallum, A. K. (1995). *Reinforcement learning with selective perception and hidden state*. Doctoral dissertation, Department of Computer Science, University of Rochester.
- Paduraru, C. (2006). Planning with approximate and learned MDP models. Master's thesis, Department of Computer Science, University of Alberta.
- Parr, R. (1998). Hierarchical Control and Learning for Markov Decision Processes. Ph.D. Dissertation, Department of Computer Science, University of California at Berkeley.
- Precup, D., Sutton, R. S., Dasgupta, S. (2001). Off-policy temporal-difference learning with function approximation. *Proceedings of the 18th International Conference on Machine Learning*.
- Rafols, E. (2006). *Temporal Abstraction in Temporal-difference Networks*. PhD thesis, Dept. of Computer Science, University of Alberta.
- Rissanen, J. (1983). A universal data compression system. *IEEE Transactions on Information Theory* 29 (5): 656–664.
- Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning*, 3(1):9–44.
- Sutton, R. S. (1990). Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Proceedings of the Seventh International Conference on Machine Learning*, pages 216–224. San Mateo, CA.
- Sutton, R. S., and Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88:135–170.
- Sutton, R. S., Barto, A. G. (1990). Time-derivative models of pavlovian reinforcement. In *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, M. Gabriel and J. Moore, Eds., pp. 497–537. MIT Press.
- Sutton, R. S., Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
- Sutton, R. S., Precup, D., Singh, S. (1999). Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112:181–211.
- Wang, T., Bowling, M., Schuurmans, D. (2007). Dual representations for dynamic programming and reinforcement learning. In *Proceedings of the 2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, 44-51.

2 Research Contributions (2003-2008)

Student authors are **bolded**. Non-student supervised-HQP authors are underlined.

Articles in Refereed Journals

1. Kehoe, E. J., **Olsen, K. N.**, Ludvig, E. A., Sutton, R. S., “Scalar timing varies with response magnitude in classical conditioning of the rabbit nictitating membrane response (*Oryctolagus cuniculus*),” *Behavioral Neuroscience* (in press, 21 manuscript pages).
2. Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., **Lee, M.**, “Natural actor–critic algorithms,” to appear in *Automatica* (accepted October 2007), 35 manuscript pages.
3. Ludvig, E. A., Sutton, R. S., Kehoe, E. J., “Stimulus representation and the timing of reward-prediction errors in models of the dopamine system,” *Neural Computation* 20:3034–3054, 2008.
4. Kehoe, E. J., Ludvig, E. A., **Dudeny, J. E.**, **Neufeld, J.**, Sutton, R. S., “Magnitude and timing of nictitating membrane movements during classical conditioning of the rabbit (*Oryctolagus cuniculus*),” *Behavioral Neuroscience* 122(2):471–476, 2008.
5. Stone, P., Sutton, R. S., **Kuhlmann, G.**, “Reinforcement Learning for RoboCup-Soccer Keepaway.” *Adaptive Behavior* 13(3):165–188, 2005.

Refereed Articles in Conference Proceedings

6. Sutton, R. S., Szepesvari, Cs., **Maei, H. R.**, “A convergent $O(n)$ algorithm for off-policy temporal-difference learning with linear function approximation,” *Advances in Neural Information Processing Systems 21* (to appear, 8 pages). MIT Press.
7. Ludvig, E., Sutton, R. S., **Verbeek, E.**, Kehoe, E. J., “A computational model of hippocampal function in trace conditioning,” *Advances in Neural Information Processing Systems 21* (to appear, 8 pages). MIT Press.
8. **Cutumisu, M.**, Szafron, D., Bowling, M., Sutton R. S., “Agent learning using action-dependent learning rates in computer role-playing games,” *Proceedings of the 4th Conference on Artificial Intelligence and Interactive Digital Entertainment* (to appear, 8 pages).
9. Sutton, R. S., Szepesvari, Cs., **Geramifard, A.**, Bowling, M., “Dyna-style planning with linear function approximation and prioritized sweeping,” *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 2008.
10. **Silver, D.**, Sutton, R. S., Müller, M., “Sample-based learning and search with permanent and transient memories,” *Proceedings of the 25th International Conference on Machine Learning*, 2008. (27% Acceptance)

11. Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., **Lee, M.**, “Incremental natural actor-critic algorithms,” *Advances in Neural Information Processing Systems 20*, 2008.
12. Sutton, R. S., **Koop, A.**, **Silver, D.**, “On the role of tracking in stationary environments,” *Proceedings of the 24th International Conference on Machine Learning*, 2007.
13. **Silver, D.**, Sutton, R. S., Müller, M., “Reinforcement learning of local shape in the game of Go,” *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, 2007.
14. **Geramifard, A.**, Bowling, M., Zinkevich, M., Sutton, R. S., “iLSTD: Eligibility traces and convergence analysis,” *Advances in Neural Information Processing Systems 19*, 2007.
15. **Geramifard, A.**, Bowling, M., Sutton, R. S., “Incremental least-squares temporal difference learning,” *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 356-361, 2006.
16. Precup, D., Sutton, R. S., **Paduraru, C.**, **Koop, A.**, Singh, S., “Off-policy learning with options and recognizers,” *Advances in Neural Information Processing Systems 18*, 2006.
17. Sutton, R. S., **Rafols, E. J.**, **Koop, A.**, “Temporal abstraction in temporal-difference networks,” *Advances in Neural Information Processing Systems 18*, 2006.
18. **Tanner, B.**, Sutton, R. S., “TD(λ) networks: Temporal-difference networks with eligibility traces,” *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, Bonn, Germany, 2005. 8 large typeset pages. 27% acceptance rate.
19. **Rafols, E. J.**, Ring, M. B., Sutton, R. S., **Tanner, B.**, “Using predictive representations to improve generalization in reinforcement learning,” *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, 2005. 6 large typeset pages. 18% acceptance rate.
20. **Tanner, B.**, Sutton, R. S., “Temporal-difference networks with history,” *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI)*, Edinburgh, Scotland, 2005. 6 large typeset pages. 18% acceptance rate.
21. Sutton, R. S., **Tanner, B.**, “Temporal-difference networks,” *Advances in Neural Information Processing Systems 17*. MIT Press, 2005.