iCORE ANNUAL REPORT 2007

iCORE CPE GRANT CPE9

DR. RICHARD SUTTON

REINFORCEMENT LEARNING AND ARTIFICIAL INTELLIGENCE

ICORE ANNUAL REPORT MARCH 31, 2007

EXECUTIVE SUMMARY

A vast number of problems of economic and scientific interest involve sequences of actions where the effects of one action influence the expected utility of subsequent actions. These *sequential decision problems* include such diverse applications as inventory management, the control of robots and industrial processes, playing backgammon, and planning under uncertainty, all of which are made more challenging because of their sequential and stochastic aspects. Many problems in robotics and artificial intelligence are also of this nature, as indeed are most of the decision-making and planning problems faced by people and animals in their daily lives. Reinforcement learning is a new body of theory and techniques for solving sequential decision problems, based on classical methods such as dynamic programming and inspired by animal learning theory, that enables larger and more diverse problems to be solved.

The objectives of the reinforcement learning and artificial intelligence (RLAI) research program are to create new methods for reinforcement learning that remove some of the limitations on their widespread application, and to develop reinforcement learning as a model of intelligence that could approach human abilities.

The primary focus of the research program has been on how intelligent machines represent their knowledge of the world. The key question is how to organize the knowledge such that it can be verified, learned, and used autonomously without continual tending by human experts. This project has pursued an unusual approach in which knowledge is expressed in terms of the machines' sensors and actuators, thereby enabling it to be compared directly to experiential data. Substantial progress was made this year toward formalizing the core learning algorithms and developing planning algorithms.

Csaba Szepesvari joined the team this year as a fourth principal investigator and new associate professor in the computing science department at the University of Alberta. Dr. Szepesvari is internationally recognized as one of the foremost theorists in reinforcement learning. Overall, the RLAI team grew to about 50 members, 31 of which are graduate students, of which 12 are recipients of major scholarships. The output of the research program has remained strong, with 38 papers published or accepted for publication in archival venues during the reporting period. Four MSc students were graduated.

The project has created a new web site this year known as the RL-Library (http://rlai.cs.ualberta.ca/RLR). This site is meant to become an international repository for reinforcement learning software based on the RL-Glue standard interface we introduced last year and which has become widely used for research and education in reinforcement learning.

RESEARCH PROGRAM OVERVIEW

The iCORE research program in reinforcement learning and artificial intelligence (RLAI) pursues an approach to artificial intelligence (AI) and other engineering problems in which they are formulated as Markov Decision Processes (MDPs) and approximately solved using Reinforcement Learning. RL is a new body of theory and techniques for solving MDPs that has been developed in the last 20 years primarily within the machine learning and optimal-control research communities. RL researchers have developed novel methods to approximate solutions to MDPs that are too large or too ill-defined for classical solution methods such as dynamic programming. For example, RL methods have obtained the best known solutions in such diverse automation applications as helicopter flying, elevator scheduling, playing backgammon, and resource-constrained scheduling. The objectives of the RLAI research program are to create new methods for reinforcement learning that remove some of the limitations on their widespread application and to develop reinforcement learning as a model of intelligence that could approach human abilities.

The iCORE research proposal for the RLAI project identified four focus areas for investigation, on each of which substantial progress was made this year.

- One proposed focus was on the role of state in AI systems, in particular on predictive state representations and on planning in large state spaces.

  Predictive state representations are a new idea for modeling sequential decision problems that are not Markov, that is, for which an appropriate state representation is not available a priori but must be constructed from the stream of raw sensor data. In previous years we introduced a new formulation of predictive state representations, *temporal-difference networks*, that enables the representation of compositional predictions, a qualitative increase in abstraction abilities.

  This year we conducted further systematic studies of compositional learning systems, culminating in Eddie Rafols's MSc thesis. The most significant aspect of this work was the thorough merging of temporal-difference networks with temporal abstraction.

  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 with the kind of function approximation needed for large state spaces is largely an open problem. One appealing approach, called *Dyna*, is to create hypothetical experience using the world model and then learn from it using RL algorithms just as if it had actually occurred.

  This year we developed the first Dyna-style planning methods extended to use linear function approximation. We completed an initial set of experiments with the new methods, culminating in Cosmin Paduraru's MSc thesis.

- A second proposed focus of the research program was on building powerful, expressive representations of the world, in particular, on the extension of RL beyond

the flat and low-level representations commonly used with MDPs to the more flexible, structured, and higher-level representations used by AI systems. The ability to represent knowledge about possible courses of action at a multiplicity of inter-related temporal scales vastly increases the generality and range of application of RL methods. Last year we combined the theory of options (one simple way of formalizing temporal abstraction) with temporal-difference networks and explored corresponding learning algorithms.

We have already mentioned our completion this year of a thorough merging of temporal-difference networks with temporal abstraction (Rafols's MSc thesis). This year we have begun a more ambitious effort, known as the PEAK project (predictive empirical abstract knowledge), in which high-level knowledge is related more explicitly to low-level sensations and actions.

- A third proposed focus of the research program was broadly concerned with approximation and generalization in RL. Approximation is required in all large-scale applications, yet is incompletely understood in both theory and practice. Most of our work to date has focused on developing software to improve practice. Our RL toolkit was designed to facilitate the use of approximation and other RL techniques in applications. Last year we extended the RL toolkit with RL-Glue, a new interface standard that has become a de-facto international standard for RL competitions.

- This year we created the RL-Library, a central site for storing and organizing RL code based on standards such as RL-Glue. Such sites exist for other branches of machine learning, and the need for one in reinforcement learning has been felt for some time, but it has not been possible to create it without a standard such as RL-Glue. At the Neural Information Processing Systems Conference in 2006 we achieved a consensus within the reinforcement learning community for the site to be created and hosted at the University of Alberta. The RL-Library is currently up and available at "http://rlai.cs.ualberta.ca/RLR."

- The fourth proposed area of investigation was the demonstration of advances in robotics applications. Our robotics laboratory has a variety of robots including a robot Segway, *Kato*. Last year we used Kato to demonstrate robot *geo-caching*—the locating, usually by people, of a hidden cache through knowledge only of its GPS coordinates.

  This year we have begun replacing key components of Kato with more robust learning-based approaches. In particular, we are exploring extending the idea of *sensor bootstrapping*—defining the meaning of one sensor in terms of its dynamic, causal relationship to another sensor. We are working toward a robot system capable of completing the Turkey Trot, a four-kilometer walk for charity that happens every fall at the University of Alberta.

RESEARCH PROJECTS

This section describes in more detail a few of the research projects making up the research program.

Very Large or Continuous Action Spaces

Large action spaces are common in practical problems, where actions may have continuous values. Despite this, no existing work has addressed the theory of efficient learning in MDPs with large action spaces. Within the realm of on-line learning, large action spaces have been studied in a simpler model in which the MDP is degenerate in the sense that it has a single state. Such problems are called stochastic bandit problems and despite their simplicity have numerous applications, ranging from experiment design for clinical trials to finding the correct setting of the parameters of an industrial plant. We have proposed a new variation of existing methods for bandit problems with large numbers of actions and shown that its rate of convergence is better that that of previous method and in fact that the rates are the best possible. However, these results are not completely satisfactory as they require that the user of the algorithm know some properties (such as smoothness) which rarely hold in practice. We plan to explore if these results can be extended to the case where such knowledge is not available.

We have also been further developing the *upper confidence-bounds based tree* (UCT) algorithm due to Levente and Szepesvari (before coming to Alberta). UCT is a search algorithm for large, discrete action spaces with a hierarchical structure. We proved that UCT's rate of convergence for deterministic environments depends on the difficulty of the search problem and not on the size of the tree to be searched. To the best of our knowledge, UCT is the first search algorithm that adapts to the difficulty of the search problem. UCT has proven to be a very practical algorithm. For example, in the most active area of research into Computer Go, the 9 x 9 board competitions, the top 10 programs are all based on UCT. These results were discussed in *Scientific America*.

In a related project within the RLAI research program, we have combined RL algorithms with the UCT-based Computer Go program MoGo to produce what is currently the world's best Go playing program. This is currently a very active area of research for us and for the international research community.

Computational Models of Animal Learning

This year we added an experimental component to our research program by teaming with Professor Jim Kehoe of the University of New South Wales. Dr. Kehoe is one of the world's foremost experts in animal learning psychology. His area of specialization is in the temporal structure of prediction processes in animal learning (classical conditioning), which makes his work particularly relevant to this project. There are striking relationships between animal learning models and computational reinforcement learning algorithms; we are exploring ways in which animal learning behavior may yield new

insights leading to better algorithms. This has happened several times in the past, leading to some of the most effective modern algorithms such as TD(λ) and Q-learning. There are animal experiments that can significantly help the computational work and that Kehoe can do quickly and inexpensively in Australia. The project is funding this work through a small subcontract. The first experiments are ongoing at the end of this reporting period.

Reinforcement Learning for Hybrid Vehicles

In collaboration with Toyota Motor Corporation, we have been exploring the use of RL technology for improving the fuel efficiency of gas-electric hybrid cars. Whenever developing a major application, new issues arise that feed into, inform, and ultimately direct ongoing research. In this application we seek to minimize fuel consumption without interfering with the car's performance. In fact, we would like the driver to be unable to tell that the learning system is operating. The car should drive just the same whether or not the learning system is engaged; the gas mileage should just be better. The special challenge is that the driver will impact performance in ways outside of the learning algorithm's control. The driver may ask for accelerations that can be delivered only with poor fuel efficiency. The driver may take the kind of trips that are not fuel efficient (e.g., very short trips or fast highway driving). The challenge for us is to design a learning agent that can separate the effects of its decisions from the effects of the driver's decisions.

We have formulated this challenge as an instance of the more general problem of control learning with disturbances. Disturbances are defined as aspects of the world that impact performance but which cannot be affected by the choices of the learning agent. The driver is one such disturbance (assuming the controller is achieving its goal of unobtrusiveness). For another example, consider controlling the heating system for a building to maximize comfort while minimizing costs. A major disturbance here is the weather; it will have a large impact on heating costs, but is outside the control of the learning agent.

We have developed a new algorithm for RL with disturbances and have shown in simplified cases that it can significantly improve performance. This work is still in its initial stages and we expect to explore several different algorithms before determining which are most effective in general and in the hybrid-car domain in particular.

Dual Representations for Reinforcement Learning

Algorithms for dynamic programming (DP) and reinforcement learning are usually formulated in terms of value functions—representations of the long run expected value of a state or state-action pair. The concept of value is so pervasive in DP and RL, in fact, that it can be hard to imagine that a value function representation is not a necessary component of any solution approach. Nevertheless, we have this year begun a systematic exploration of such an alternative approach.

The dual approach to dynamic programming and reinforcement learning is based on maintaining an explicit representation of a stationary distribution instead of a value function. A significant advantage of the dual approach is that it allows one to exploit well developed techniques for representing, approximating and estimating probability distributions without running the risks associated with divergent value function estimation. A second advantage is that some distinct algorithms for the average reward and discounted reward cases become unified under the dual representation. We have developed a modified dual of the standard linear program that guarantees a globally normalized state visit distribution is obtained. With this reformulation, we have derived novel dual forms of dynamic programming, including policy evaluation, policy iteration and value iteration. We have also derived dual formulations of temporal-difference learning to obtain new forms of Sarsa and Q-learning. Finally, we have explored the scaling of these techniques to large domains by introducing function approximation. The dual view seems to yield a viable alternative to standard value-function-based techniques.

OBJECTIVES FOR NEXT YEAR

A major focus for next year will be on planning with linear function approximation. Planning is a core topic for the entire research program, corresponding roughly to reasoning in people and to optimal decision making in control theory. Much of the 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. There are a number of possible approaches arising out of our previous work in iLSTD and in Dyna-style planning systems. We will be developing these both theoretically and experimentally over the next year. The primary objective for the year will be a general and sound planning algorithm for worlds with linearly approximated dynamics at a single time scale. If this is achieved then extensions to multiple time scales and predictive state representations will be attempted.

A new focus for next year will be on treating reinforcement learning as a tracking problem. Conventionally in machine learning we assume the goal of converging to a single optimal solution. In contrast to this, our preliminary results suggest that in large problems a time-varying solution may perform better than any fixed solution, even if the underlying problem being solved is stationary. Our objective for next year is to demonstrate this phenomena in a state of the art Computer Go program.

Creating compelling examples and expositions of our experience-based approach to AI remains a priority. We seek to publish at least two papers on this topic next year. One of these will be an experiential exploration of the commonsense notion of physical objects. Sutton will be on sabbatical for much of next year and intends to use the time to write a monograph on this subject.

We continue to seek software and algorithms which make RL easier to apply without expert knowledge. All of the software systems we have made – the RL-Toolkit, RL-Glue,

and the RL-Library – are towards this goal. All these software systems need to be strengthened and promoted. Our goal for the next year is to increase their user-friendliness and increase their use in education.

In concert with the software effort we will develop a new focus on "black box" RL algorithms. These are algorithms that have no parameters or settings of any kind and can be used without knowledge of what is going on inside. This is a long-term goal that will probably be the topic of several theses. Our goal for next year is to develop RL algorithms that require only meta-parameters – parameters for setting the other parameters. Such systems will inevitably pay a penalty in initial learning rate, but would be far easier to use.

Over the next year we seek to develop and publish a new model of animal learning (classical conditioning) based on the temporal-difference learning algorithm and an extended temporal representation which we call *micro-stimuli*. Micro-stimuli turn a simple sensory event into a temporally extended and multi-component representation which enables precise timing of responses. Our objective for the next year is to develop the model through comparisons with empirical data from animal learning and neuroscience.

In robotics, our focus over the next year or more will be on route following. We plan to integrate ideas from ongoing research into a robot system capable of completing the Turkey Trot, a four-kilometer walk for charity that happens every fall at the University of Alberta. The system will require robust positioning, local and global navigation, and obstacle avoidance, all over a long challenging route. It will be further complicated by crowds of other walkers. We hope to enter a preliminary system in the fall of 2007, with the aim to successfully complete the course at an average walking pace in the fall of 2008.

RESEARCH TEAM MEMBERS AND CONTRIBUTIONS

Team Leader

Name	Role / Topic	Awards / Special Info
Richard Sutton	Principal investigator	NSERC, AICML

Faculty Team Members

Name	Role / Topic	Awards / Special Info
Michael Bowling	Principal investigator Robotics, games, reinforcement learning	NSERC, AIF New Faculty, AICML
Dale Schuurmans	Principal investigator Probabilistic methods in artificial intelligence, machine learning	CRC II Chair, NSERC, AICML
Csaba Szepesvari	Principal investigator Nonparametric learning, statistical techniques	NSERC, AICML

Mark Ring	Visiting professor, Chapman University Continual learning, neural networks, sequence learning.	

Postdoctoral Fellows

Name	Topic/notes	Awards / Special Info
Yaakov Engel	Gaussian process reinforcement learning	Alberta Ingenuity Fellowship
Mohammad Ghavamzadeh	Hierarchical reinforcement learning, Bayesian policy-gradient reinforcement learning	
Yuxi Li	Computational Finance	
Elliot Ludvig	Computational models of animal learning Integrating approaches to natural and artificial intelligence	
Martin Zinkevich	Minimal regret algorithms	

PhD Students

Name	Role / Topic	Scholarships / Awards / Special Info
Arash Afkanpour	Function approximation and feature discovery in reinforcement learning	iCORE
Amir massoud Farahmand	Manifold learning methods for reinforcement learning problems	PhD Academic Achievement Award, Provost Doctoral Entrance Award
Alborz Geramifard	Least-squares reinforcement learning methods and planning	
Yuhong Guo	Learning bayesian networks	
Jiayuan Huang	Spectral clustering and semi-supervised learning	University of Waterloo
Feng Jiao	Bayesian methods	University of Waterloo until Sept 2006 then employed by Google
Daniel Lizotte	Bayesian global optimization	Killam, iCORE
Adam Milstein	Localization and mapping	University of Waterloo
David Silver	RL in computer GO	iCORE, FS Chia, AIF
Ivett Szabo	Bandit algorithms	Budapest University of Technology and Economics
Brian Tanner	Experience oriented reinforcement learning	NSERC CGS-D, Killam, AIF, AIF Bill Bridger iCORE, Walter H. Johns
Qin Wang	Learning structured classifiers for statistical dependency parsing	

Tao Wang	New representation and approximations for sequential decision making	SIGART/AAAI Doctoral Consortium Scholarship
Adam White	A benchmarking system for reinforcement learning	AIF, iCORE
Dana Wilkinson	Subjective mapping	Ontario Graduate Scholarship, President's Scholarship from University of Waterloo
Linli Xu	Convex large margin training techniques	University of Waterloo

MSc Students

Name	Role / Topic	Scholarships / Awards / Special Info
Yasin Abbasi-Yadkori	Bandit problem in the continuous action space	
Andrew Albert	Machine learning and games	NSERC CGS-M, Walter H. Johns
Nolan Bard	Agent modeling using state estimation	
Kate Davison	Generating bidding strategies in bridge	
Varun Grover	Applying shaping and transfer leaning to reinforcement learning	
Michael Johanson	Opponent modeling in poker	
Brad Joyce	Reinforcement learning with continuous action	iCORE, NSERC PGS-M, Walter H. Johns, Dept Academic Award
Armita Kaboli	Bayesian calibration for robot localization	ECE student
Anna Koop	Predictive features for knowledge abstraction	AIF, iCORE, NSERC CGS-M, Walter H. Johns
Volodymyr Mnih	Efficient stopping rules	NSERC CGS-M, iCORE, Walter H. Johns
Gergely Neu	Inverse reinforcement learning	Budapest University of Technology and Economics
James Neufeld	Autonomous outdoor navigation	
Cosmin Paduraru	Planning with approximate and learned MDP models	iCORE
Eddie Rafols	Temporal abstraction in temporal-difference networks	NSERC PGS-M, Walter H. Johns
Masoud Shahamiri	Disturbance in reinforcement learning	

Other team members

Name	Role
Andrew Butcher	Graphics programming
Steve Fraser	Segway hardware
Akiko Green	Administrative assistant
Jacqueline Jean	WISEST summer student
Mark Lee	Programmer

Nelson Loyola	AIBO programming
Jason Roberts	Segway programming
Christian Smith	Simulation programming
Mike Sokolsky	Robot engineer
Lori Troop	Program administrator
Stephen Walsh	Segway hardware

COLLABORATIONS

Provincial	
Participants	Nature of Collaboration
Alberta Ingenuity Center for Machine Learning	R. Sutton, D. Schuurmans, Cs. Szespesvari and M. Bowling are among the eight principle investigators for this center at the U. of Alberta. Total annual funding for AICML is 2.2M.

National	
Participants	Nature of Collaboration
Doina Precup, McGill University Prakash Panangaden, McGill University Yoshua Bengio, U. Montreal Shie Mannor, McGill University	NSERC Collaborative Research and Development Grants-Project, "Learning and prediction in high-dimensional stochastic domains," with Nortel Networks, Bell Canada, to R. Sutton and the listed collaborators $186,523 total (Sept 1/06 – Aug 31/09).
Jonathan Schaeffer, Robert Holte, Duane Szafron, and Michael Buro University of Alberta	NSERC Strategic Grant, "Intelligent Agents for Interactive Entertainment" to M. Bowling, and D. Schuurmans, and the listed collaborators $480K total (Oct 1/04-Sept 30/07).

International	
Participants	Nature of Collaboration
Shalabh Bhatnagar, Indian Institute of Science, Bangalore	Joint research with R. Sutton, M. Ghavamzadeh, and M. Lee on actor-critic RL algorithms.
Sylvain Gelly, University of Paris South	Joint research with D. Silver on the use of RL in Computer Go.
Jean-Yves Audibert	Joint research with Cs. Szepesvari on variance estimation in bandit problems.
E. James Kehoe, University of New South Wales, Australia	Joint research with R. Sutton, E. Ludvig and J. Neufeld on the relationship between reinforcement learning and learning in animals.
Satinder Singh, University of Michigan	Joint research with R. Sutton on predictive state representations and temporal-difference networks.

Industrial	
Participants	Nature of Collaboration
Michael James, Toyota Motor Corporation	With R. Sutton and M. Shahamiri exploring the use of reinforcement learning technology for improving the fuel efficiency of gas-electric hybrid cars.

The following table lists some of the collaborative activities of RLAI team members within the research community.

Collaborative activities in the research community		
Name	Conference/Journal/Activity	Role
B.Tanner	Grounding Perception, Knowledge, and Cognition in Sensorimotor Experience, a workshop at the 20th Annual Conference on Advances in Neural Information Processing Systems	Organizer and co-chair
Adam White	The First Annual Reinforcement Learning Competition, a workshop at the 20th Annual Conference on Advances in Neural Information Processing Systems	Organizer and co-chair
Cs. Szepesvari	AI Communications	Associate editor
D. Schuurmans	Artificial Intelligence Journal	Associate editor
M. Bowling	23rd International Conference on Machine Learning	Volunteer chair
R. Sutton	23rd International Conference on Machine Learning	Senior program committee
M. Bowling	22nd National Conference on Artificial Intelligence	Senior program committee

D. Schuurmans	22nd Conference on Artificial Intelligence	Senior program committee
D. Schuurmans	21st Annual Conference on Neural Information Processing Systems	Senior program committee
Cs. Szepesvari	17th European Conference on Machine Learning	Senior program committee
R. Sutton	20th Annual Conference on Neural Information Processing Systems	Program committee
M. Bowling	20th Annual Conference on Neural Information Processing Systems	Program committee
M. Bowling	20th International Joint Conference on Artificial Intelligence	Program committee
M. Bowling	Journal of Field Robotics	Program committee
M. Bowling	24th Annual International Conference on Machine Learning	Program committee
M. Bowling	2007 Robotics: Science and Systems Conference	Program committee
D. Schuurmans	20th International Joint Conference on Artificial Intelligence	Program committee
D. Schuurmans	Conference on Empirical Methods in Natural Language Processing Conference on Computational Natural Language Learning	Program committee
Cs. Szepesvari	20th Annual Conference on Neural Information Processing Systems	Program committee
Cs. Szepesvari	20th International Joint Conference on Artificial Intelligence	Program committee
Cs. Szepesvari	23rd International Conference on Machine Learning	Program committee
Cs. Szepesvari	24th International Conference on Machine Learning	Program committee
Cs. Szepesvari	21st National Conference on Artificial Intelligence	Program committee
Cs. Szepesvari	22nd National Conference on Artificial Intelligence	Program committee
Cs. Szepesvari	20th Annual Conference on Learning Theory	Program committee
M. Ghavamzadeh	20th Annual Conference on Neural Information Processing Systems	Program committee
M. Ghavamzadeh	20th International Joint Conference on Artificial Intelligence	Program committee
M. Ghavamzadeh	6th International Joint Conference on Autonomous Agents and Multiagent Systems	Program committee
M. Ghavamzadeh	24th International Conference on Machine Learning	Program committee
M. Ghavamzadeh	22nd National Conference on Artificial Intelligence	Program committee
B. Tanner	5th International Joint Conference on Autonomous Agents and Multiagent Systems	Program committee

RLAI team members hosted many prominent short-term visitors during the reporting period, including: Scott Sanner (University of Toronto), Pascal Poupart (University of Waterloo), Doron Tal (Research Institute for Advanced Computer Science at NASA Ames Research Center), Koby Crammer (University of Pennsylvania), James Kehoe (University of New South Wales), Kocsis Levente (Computer and Automation Research Institute, Hungarian Academy of Sciences), Eva Czabarka (University of South Carolina), Laszlo Szekely (University of South Carolina), Liang Huang (University of Pennsylvania), Alex Strehl (Rutgers University), Sylvain Gelly (University of Paris South), Peter Auer (University of Leoben), Wenye Li (The Chinese University of Hong Kong), Jean-Yves Audibert (Centre d'Enseignement et de Recherche en Technologies de l'Information et Systèmes, Ecole Nationale des Ponts et Chaussées)

GRADUATES

Student	Degree	Research topic	Current Position
Cosmin Paduraru	MSc	Planning with Approximate and Learned MDP Models	PhD student at McGill with Doina Precup
Eddie Rafols	MSc	Temporal Abstraction in Temporal-difference Networks	
Alborz Geramifard	MSc	iSLTD	PhD student at University of Alberta
AdamWhite	MSc	A Benchmarking System for Reinforcement Learning	PhD student at University of Alberta

INTELLECTUAL PROPERTY

None.

PUBLICATIONS

REFEREED JOURNAL PUBLICATIONS

A. Antos, Cs. Szepesvári and R. Munos, "Learning Near-optimal Policies with Bellman-Residual Minimization Based Fitted Policy Iteration and a Single Sample Path," *Machine Learning Journal,* to appear.

A. Blum, T. Sandholm, and M. Zinkevich, "Online Algorithms for Market Clearing," *Journal of the ACM*, Sept. 2006, Vol. 53, No. 5, pp. 845-879.

E.A. Ludvig, K. Conover, P. Shizgal, "Effects of Reward Magnitude on Timing in Rats," *Journal of Experimental Analysis of Behavior*, Mar. 2007, 87, pp. 201-218.

M. Zinkevich, A. Greenwald, M. Littman, "A Hierarchy of Prescriptive Goals for Multiagent Learning," *Artificial Intelligence*, to appear.

HIGHLY REFEREED ARCHIVAL CONFERENCE PROCEEDINGS

P.Auer, R. Ortner and C. Szepesvári, "Improved Rates for the Stochastic Continuum-Armed Bandit Problem Time Associative Bandit Problems," *Proc. of the 20th Annual Conf. on Learning Theory*, (COLT-07), to appear. 43% acceptance rate.

N. Bard, M. Bowling, "Particle Filtering for Dynamic Agent Modelling in Simplified Poker," *Proc. of the 22nd Conf. on Artificial Intelligence* (AAAI-07), 2007, to appear. 27% acceptance rate.

S. Bergsma and Q. Wang, "Learning Noun Phrase Query Segmentation," *Proc. of EMNLP-CoNLL 2007*, to appear. 17% acceptance rate of oral presentations.

I. Biro, Z. Szamonek and C. Szepesvári, "Sequence Prediction Exploiting Similarity Information, *Proc. of the 20th Int'l Joint Conf. of Artificial Intelligence*, (IJCAI-07), 2007, to appear. 35% acceptance rate.

L.Cheng, SVN. Vishwanathan, D. Schuurmans, S.Wang, and T. Caelli, "Implicit Online Learning with Kernels," *Proc. Neural Information Processing Systems 20* (NIPS-06), 2006. 24% acceptance rate.

S. Gelly, and D. Silver, "Combining Online and Offline Knowledge in UCT," *Proc. of the 24th Int'l Conf. on Machine Learning*, (ICML-07), 2007, to appear. 29% acceptance rate.

A. Geramifard, M. Bowling, M. Zinkevich, and R.S. Sutton, "iLSTD: Eligibility Traces and Convergence Analysis," *Proc. Neural Information Processing Systems 20* (NIPS-06), 2006. 24% acceptance

M. Ghavamzadeh, and Y. Engel, "Bayesian Policy Gradient Algorithms," *Proc. Neural Information Processing Systems 20* (NIPS-06), 2006. 24% acceptance rate.

M. Ghavamzadeh and Y. Engel, "Bayesian Actor-Critic Algorithms," *Proc. of the 24th Int'l Conf. on Machine Learning* (ICML-07), 2007, to appear. 29% acceptance rate.

A. Ghodsi, F. Southey, and D. Wilkinson, "Improving Embeddings by Flexible Exploitation of Side Information," *Proc. of the 20th Intl. Joint Conf. on Artificial Intelligence* (IJCAI-07), 2007, to appear. 35% acceptance rate.

Y. Guo, and D. Schuurmans, "Convex Structure Learning for Bayesian Networks: Polynomial Feature Selection and Approximate Ordering," *Proc. of the 22nd Conf. on Uncertainty in Artificial Intelligence* (UAI-06), 2006. 31% acceptance rate.

A. György, L. Kocsis, I. Szabó, and Cs. Szepesvári, "Continuous Time Associative Bandit Problems," *Proc. of the 20th Int'l Joint Conf. of Artificial Intelligence* (IJCAI-07),

2007, to appear. 35% acceptance rate.

J. Huang, A. Smola, A. Gretton, K. Borgwardt, and B. Schoelkopf, "Correcting Sample Selection Bias by Unlabeled Data," *Proc. Neural Information Processing Systems 20* (NIPS-06), 2006. 24% acceptance rate.

J. Huang and D. Schuurmans, "Information Marginalization on Subgraphs," *Proc. 17th European Conf. on Machine Learning and the 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases* (ECML/PKDD-06), 2006. 14% acceptance rate.

J. Huang, T. Zhu, and D. Schuurmans, "Web Community Identification From Random Walks," *Proc. 17th European Conf. on Machine Learning and the 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases* (ECML/PKDD-06), 2006. 14% acceptance rate

F. Jiao, S. Wang, C. Lee, R. Greiner, and D. Schuurmans, "Semi-Supervised Conditional Random Fields for Improved Sequence Segmentation and Labeling," *Proc. Joint Conf. of the Int'l Committee on Computational Linguistics and the Association for Computational Linguistics* (COLING/ACL-06), 2006. 23% acceptance rate.

F. Jiao, J. Xu, L. Yu, and D. Schuurmans, "Protein Fold Recognition Using the Gradient Boost Algorithm," *Computational Systems Bioinformatics Conf.* (CSB-06). 19% acceptance rate.

L. Kocsis and Cs. Szepesvári, "Bandit Based Monte-Carlo Planning," *Proc. of the 17th European Conference on Machine Learning* (ECML-06), 2006, LNCS/LNAI 4212, pp. 282-293. 14.5% acceptance rate.

R.S. Sutton, A. Koop, and D. Silver, "On the Role of Tracking in Stationary Environments," *Proc. of the 24th Int'l Conf. on Machine Learning*, (ICML-07), 2007. 29% acceptance rate.

C. Lee, S. Wang, F. Jiao, D. Schuurmans, and R. Greiner, "Learning to Model Spatial Dependency: Semi-Supervised Discriminative Random Fields," *Proc. Neural Information Processing Systems 20* (NIPS-06), 2006. 24% acceptance rate.

M. Littman, N. Ravi, A. Talwar, and M. Zinkevich, "An Efficient Optimal-Equilibrium Algorithm for Two-Player Game Trees," *Proc. of the 22nd Conf. on Uncertainty in Artificial Intelligence* (UAI-06), 2006. 31% acceptance rate.

D. Lizotte, T. Wang, M. Bowling, and D. Schuurmans, "Automatic Gait Optimization with Gaussian Process Regression," *Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence* (IJCAI-2007), 2007, to appear. 35% acceptance rate.

A. Milstein and T. Wang, "Localization With Dynamic Motion Models: Motion Model Parameters Dynamically in Monte Carlo Localization," *Proc. of the 3rd Intl. Conf. on Informatics in Control, Automation and Robotics* (ICINCO 2006), 2006. 10% acceptance rate.

N. Ratliff, J. Bagnell, and M. Zinkevich, "Maximum Margin Planning," *Proc. of the 23rd Int'l Conf. on Machine Learning* (ICML-2006), 2006. 20% acceptance rate.

N. Ratliff, D.Bagnell, M. Zinkevich, "Subgradient Methods for Structured Prediction," *Artificial Intelligence and Statistics 11* (AISTATS-07), 2007.

D. Silver, R.S. Sutton, and M. Mueller, "Reinforcement Learning of Local Shape in the Game of Go," *Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence*, (IJCAI-07), 2007. 35% acceptance rate.

F. Southey, W. Loh, and D. Wilkinson, "Inferring Complex Agent Motions from Partial Trajectory Observations," *Proc. of the 20th Int'l Joint Conf. on Artificial Intelligence* (IJCAI-2007), 2007. 35% acceptance rate.

B. Tanner, V. Bulitko, A. Koop, and C. Paduraru, "Grounding Abstraction in Predictive State Representations," *Proc. of the 20th Int'l Joint Conf. of Artificial Intelligence* (IJCAI-07), 2007. 35% acceptance rate.

Q. Wang, D. Lin, and D. Schuurmans, "Simple Training of Dependency Parsers via Structured Boosting," *Proc. Neural Information Processing Systems 20* (NIPS-06), 2006. 24% acceptance rate.

Q. Wang, D. Lin and D. Schuurmans, Simple Training of Dependency Parsers via Structured Boosting, *Proc. of the 20th Int'l Joint Conf. of Artificial Intelligence* (IJCAI-07), 2007, pp. 1756-1762. 35% acceptance rate.

L. Xu, D. Wilkinson, F. Southey and D. Schuurmans, "Discriminative Unsupervised Learning of Structured Predictors," *Proc. of the 23rd Int'l Conf. on Machine Learning* (ICML-06), pp. 1057-1064. 20% acceptance rate.

D. Zhou, J. Huang, and B. Schoelkopf, "Learning with Hypergraphs: Clustering, Classification, and Embedding, *Proc. Neural Information Processing Systems 20* (NIPS-06), 2006. 24% acceptance rate.

M. Zinkevich, M. Bowling, N. Burch, "A New Algorithm for Generating Equilibria in Massive Zero-Sum Games," *Proc. of the 22nd National Conf. on Artificial Intelligence* (AAAI-07), 2007, to appear. 27% acceptance rate.

BOOKS AND CHAPTERS
None

SPECIAL INVITED PRESENTATIONS

May, 2006, "Learning to Cooperate using Hierarchical Reinforcement Learning," presented at the workshop on Hierarchical Autonomous Agents and Multi-Agent Systems (H-AAMAS), at the fifth Int'l Joint Conf. on Autonomous Agents and Multiagent Systems (AAMAS-2006), Hakodate, Japan, Mohammad Ghavamzadeh.

June 13, 2006, "Learning to Control an Octopus Arm with Gaussian Process Temporal Difference Methods," Gaussian Processes in Practice workshop, Bletchley Park, UK, Yakov Engel.

June 29, 2006 "Gaussian Process Temporal Difference Learning - Theory and Practice," Kernel machines and Reinforcement Learning workshop, ICML'06, Pittsburgh, PA., Yakov Engel.

July 10, 2006, "Incremental Least-Squares Temporal Difference Learning," AI seminar at Carnegie Mellon University, Michael Bowling.

July 12, 2006, "Convex Training Algorithms for Hard Machine Learning Problems," Plenary Lecture at the Center for Language and Speech Processing, summer workshop, Johns Hopkins University, Dale Schuurmans.

August 24, 2006, "Predictive Action Descriptions from Experience," presented at Inductive Logical Programming 2006 in Santiago de Compostela, Spain, Brian Tanner.

November 8, 2006, "Games, Optimization, and Online Algorithms," Session on Online Convex Optimization, INFORMS 2006, Pittsburgh, PA, Martin Zinkevich.

December 8, 2006, "What's Wrong with Reinforcement Learning," at the workshop: Towards a New Reinforcement Learning at the Twentieth Annual Conference on Neural Information Processing Systems (NIPS-2006), Richard Sutton.

December 8, 2006, "Grounding Artificial Cognition," at the workshop: Grounding Perception, Knowledge and Cognition in Sensori-Motor Experience at the Twentieth Annual Conference on Neural Information Processing Systems (NIPS-2006)

December 8, 2006, "Bayesian Reinforcement Learning with Gaussian Processes," at the workshop: Towards a New Reinforcement Learning at the Twentieth Annual Conference on Neural Information Processing Systems (NIPS-2006), Yakov Engel.

December 8, 2006, "Bayesian Policy Gradient Algorithms" at the workshop: Towards a New Reinforcement Learning, at the Twentieth Annual Conference on Neural Information Processing Systems (NIPS-2006), Mohammad Ghaamzadeh.

December 8, 2006, "Using Upper Confidence Bounds to Control Exploration and Exploitation," at the workshop: On-line Trading of Exploration and Exploitation, at the 20th Annual Conference on Neural Information Processing Systems (NIPS-2006), Csaba Szepesvari.

December 8, 2006, "Sample Complexity Results for Reinforcement Learning in Large State Spaces," at the workshop: Towards a New Reinforcement Learning at the Twentieth Annual Conference on Neural Information Processing Systems (NIPS-2006), Csaba Szepesvari.

February 5, 2007, "A Reinforcement Learning Model of Response Timing in Classical Conditioning" part of Focus Session on "Modeling Data" at Winter Conference on Animal Learning & Behavior, Winter Park, Co, Elliot Ludvig.

AWARDS

The Alberta Ingenuity Center for Machine Learning, four of whose eight principal investigators are the four principal investigators of this RLAI project, received the Alberta Science and Technology Leadership Foundation (ASTech) 2006 award for Outstanding Leadership in Alberta Technology.

Michael Bowling was awarded an Alberta Ingenuity New Faculty grant for 100k/year for three years. This grant is to pursue "Subjective Models for Autonomous Robots."

The following paper won the best student paper award at 2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning: T. Wang, M. Bowling, and D. Schuurmans, "Dual Representations for Dynamic Programming and Reinforcement Learning."

THESES

Alborz Geramifard, (MSc thesis), "iLSTD," December 12, 2006.

Cosmin Paduraru, (MSc thesis), "Planning with Approximate and Learned MDP Models," December 12, 2006.

Eddie Rafols, (MSc thesis), "Temporal Abstraction in Temporal-difference Networks," September 29, 2006.

Adam White, (MSc thesis), "A Benchmarking System for Reinforcement Learning,"

September, 2006.

OUTREACH

Dan Lizotte exhibited a joint research project with Michael Bowling on automatically tuning robotic gaits using experience at the Smithsonian Institute's Folklife Festival in Washington, D.C., highlighting Alberta.

Dan Lizotte presented an introductory robotics lecture and demo to the grade eight science classes of Richard S. Fowler Catholic Junior High School.

Anna Koop conducted a session for "Choices - A Conference for Grade Six Girls," giving them hands-on RL experience.

Anna Koop conducted a session for "Women in Technology," a program for grade nine female students to participate in IT workshops.

James Neufeld conducted Segway Robot demonstrations for the students from Augustana University, Brander Gardens Elementary school, and Edmonton high school teachers.

Jacqueline Jean was a WISEST (Women in Scholarship, Engineering, Science and Technology) summer student.

FUNDING

Summary of funding from other sources:

NSERC discovery grant to Richard Sutton	50,000/yr
NSERC CRDPJ to co-principal investigator Richard Sutton	186,523 total
Canada research chair to Dale Schuurmans	100,000/yr
NSERC discovery grant to Dale Schuurmans	47,000/yr
NSERC discovery grant to Michael Bowling	26,000/yr
NSERC discovery grant to Michael Bowling	28,000/yr
AIF New Faculty Grant to Michael Bowling	328,455 total
AIF postdoctoral fellowship to Yaakov Engel	110,000 total
The Alberta Ingenuity Center for Machine Learning (AICML), PIs Sutton, Bowling, Schuurmans, Holte, Greiner, Schaeffer, Goebel, and Szepesvari, provided support for programming staff, software management, visitors, robot lab renovation, robot hardware, software support, postdocs and students	~2.2M/total
Student awards and scholarships	~342,000/yr
NSERC Strategic Grant, "Intelligent Agents for Interactive Entertainment" to M. Bowling, D. Schuurmans, with Jonathan Schaeffer, Robert Holte, Duane Szafron, and Michael Buro	480,000 total