iCORE ANNUAL REPORT 2011

iCORE CPE GRANT CPE45

DR. RICHARD SUTTON

REINFORCEMENT LEARNING AND ARTIFICIAL INTELLIGENCE

ICORE ANNUAL REPORT JAN 31, 2011

1. EXECUTIVE SUMMARY

The RLAI research program pursues an approach to artificial-intelligence and engineering problems in which they are formulated as large optimal-control problems and approximately solved using reinforcement-learning methods. Reinforcement learning is a new body of theory and techniques for optimal control that has been developed in the last twenty years primarily within the machine learning and operations research communities, and which have separately become important in psychology and neuroscience. Reinforcement learning researchers have developed novel methods to approximate solutions to optimal-control problems that are too large or too ill-defined for classical solution 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.

The objectives of the RLAI research program are to create new methods for reinforcement learning that remove some of the limitations on its widespread application and to develop reinforcement learning as a model of intelligence that could approach human abilities. These objectives are pursued through mathematics, through computational experiments, through the development of robotic systems, and through the development and testing of computational models of natural learning processes.

The research team consists of about 50 members, 32 of whom are graduate students and, of those, 17 of which are recipients of major scholarships. The output of the research program has remained strong, with 31 papers published or accepted for publication in archival venues during the reporting period. Two MSc students were graduated.

The primary focus of the research program has been on how intelligent machines represent their knowledge of the world. The key question here 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 machine's sensors and actuators, thereby enabling it to be compared directly to experiential data. Substantial further progress was made this year toward formalizing the core learning algorithms for doing this.

Highlights of the research program this year include: 1) the extension of off-policy learning algorithms to control, 2) the learning of thousands of simultaneous predictions in real time on a physical robot, and 3) a new initiative in robots interacting with the public at the Glenrose Rehabilitation Hospital.

2. RESEARCH PROGRAM OVERVIEW

In the project's proposal, research was divided into three main target areas. The first is extensions of existing reinforcement learning algorithms; there are many open problems in reinforcement learning, and we seek to solve them as opportunities arise. The second area is the extension of reinforcement learning ideas to address the more ambitious goals of artificial intelligence (AI). There is a natural transition from the more advanced reinforcement learning methods to mechanisms for knowledge representation, search, and human-level reasoning. A major goal for the project is to explore, implement, and illustrate these relationships. The third main area of RLAI research is a focus on applications—on designing algorithms and software that are well suited for applied research, and on several specific applications with which we are working. We discuss highlights of our research towards each of these target areas below.

In the area of core reinforcement-learning algorithms, there has been continuing progress with our recently introduced *gradient temporal-difference* (GTD) learning algorithms. Temporal-difference (TD) learning is a core technology at the heart of modern reinforcement learning; a variant of TD learning is used in almost all large-scale applications of reinforcement learning. However, an important limitation of standard algorithms is that they can only be used to learn about a way of behaving if the learning system commits to that way of behaving all the way to its completion. People, by contrast, are able to learn about a way of behaving even if they behave that way for only a limited time. We can start to do something (e.g., drive to work) and then abort it for any reason (e.g., we remember that it's a holiday) yet still learn from the incomplete trial. Or we might start behaving in a way that is consistent with many different ways of continuing. Of course we can continue only according to one of them (we can only do one thing at a time!), but we might hope to learn about all of them from the initial portion that they share. This is called "off-policy learning" and has been part of the promise of TD learning for decades; with the GTD methods we introduced two years ago it can at last be done in a scalable and efficient way, resolving the most important open problem in reinforcement learning in the last 15 years.

This year we have continued to make progress with off-policy learning by GTD algorithms. We have come to understand the algorithms better and have extended them to new settings. The most important extension has been from a prediction setting (learning the consequences of a given way of behaving) to a control setting (learning an optimal way of behaving). Last year we introduced a very general and powerful GTD prediction algorithm called *GQ(λ)*. This year we proved that if the way of behaving was continually adjusted in a local "greedy" fashion, then the resulting algorithm, called *Greedy-GQ*, will converge to an optimal policy. Greedy-GQ is a direct improvement over Q-learning, the most popular off-policy TD control method, enabling off-policy learning on large applications for which Q-learning is inapplicable, inefficient, or unreliable.

As the new GTD algorithms prove effective, they become key to our efforts in the area of extending reinforcement learning to the more ambitious goals of AI. In the last year we have focused on how TD methods can be used in robots toward this goal. Classically, reinforcement learning has focused on the relatively small goals of learning how to behave (policies) and how much reward can be obtained (value functions). A major thrust of the RLAI project is to move beyond these to be able to learn more general and abstract kinds of knowledge about the world, and to show that these things can be learned in real time for physical robots. We are now using three different kinds of robots, shown in Figure 1, including 1) the *Critterbot*, a custom-built mobile robot with many sensors, 2) several *iRobot Creates*, a commercial robot platform based on the Roomba(tm), and 3) an Aldebaran Nao torso, a stationary research robot with articulated arms and vision capabilities. In the last year we have used the Critterbot to demonstrate off-policy learning of multiple policies in parallel using the GQ(λ) algorithm.

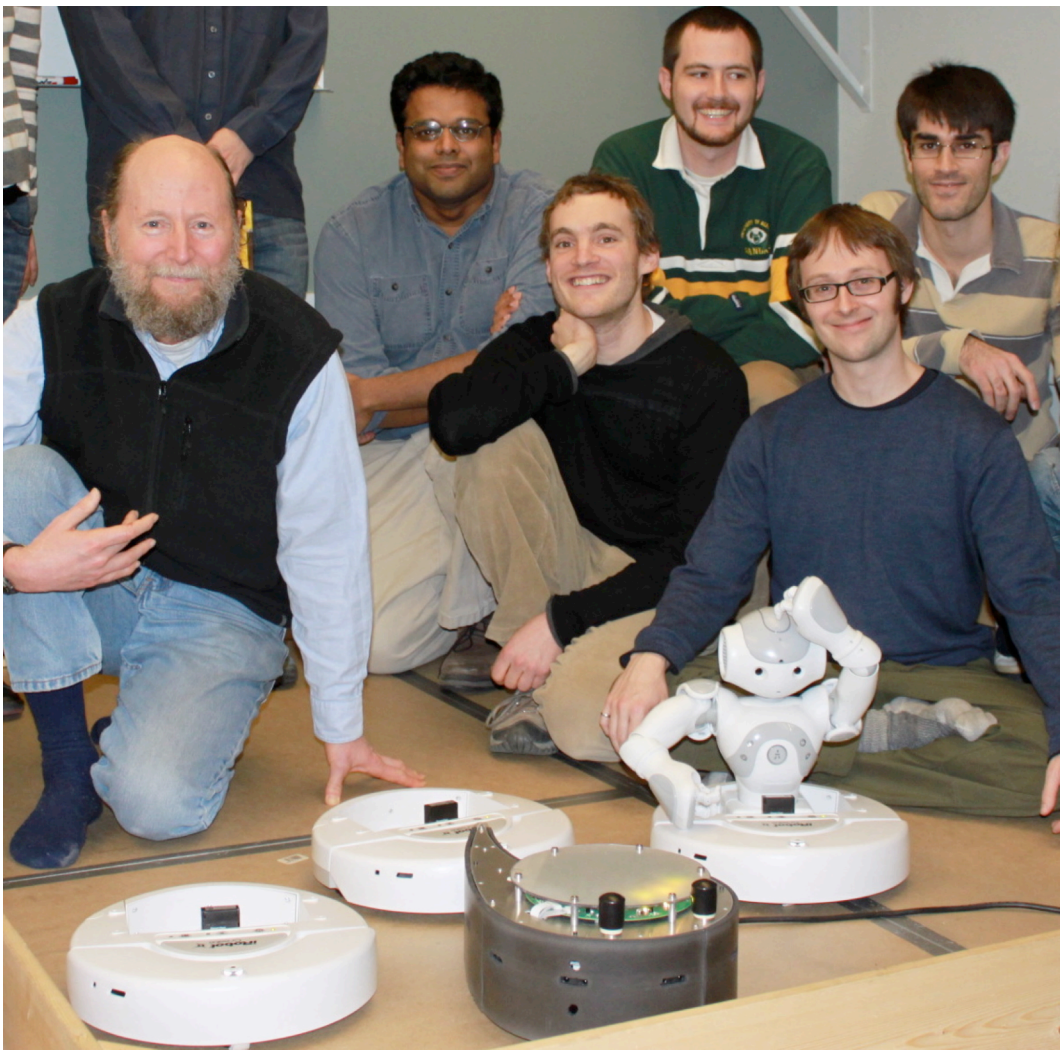

Figure 1. Team members and robot platforms used in the research.

This year we used the Critterbot to demonstrate a robot version of the psychological phenomena of nexting. Psychologists have used this term to describe the propensity of people and other animals to continually predict what will happen next to their sensory signals. The ability to next is viewed as a basic kind of awareness and knowledge of one's environment and its relationship to the body. This year, we programmed the Critterbot to learn to next in real time. We applied TD learning methods to predict 46 sensor readings at four different time scales ranging from milliseconds to tens of seconds. Figure 2 shows an example of one of these predictions. In total we learned and updated 2000 predictions ten times a second. This is the first time real-time nexting of any form has been shown on a physical robot.

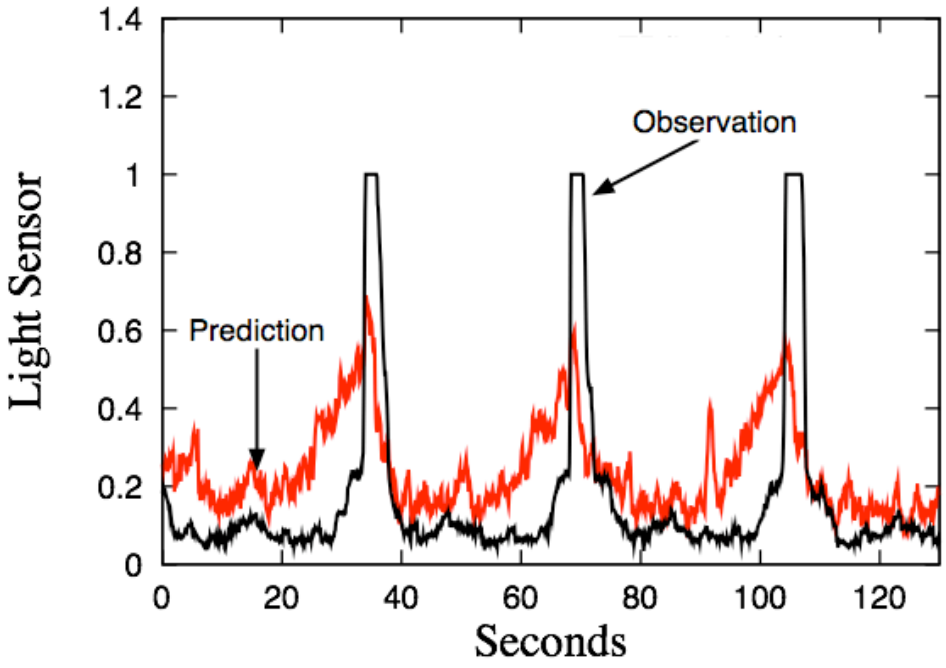

Figure 2. "Nexting" in a reinforcement-learning robot. The black line shows readings from a light sensor peaking every 30 seconds as the robot passes a bright light once on each trip around its pen. The red line shows a learned prediction, based on other sensor readings, rising several seconds in advance of the actual light.

In other applications, associated faculty member Martin Mueller has continued work on high-performance Computer Go. His program, Fuego, took second place in 9x9 Go and 13x13 Go at the 2010 Computer Olympiad, and first place in full 19x19 Go at the 2010 Computer Go UEC Cup (University of Electro-Communications, Japan).

The project has started a new collaboration with the Glenrose Rehabilitation Hospital in Edmonton. Funded by MITACS, the objective of this new two-year initiative is to explore social interaction between robots and members of the public, including patients. The project will program and field a mobile robot based on the iRobot Create(tm) that interacts with visitors to the Courage Center, a new facility at the Glenrose intended to showcase the use of new technologies, such as robotics, in rehabilitative medicine.

3. RESEARCH PROJECTS

This section describes in more detail some of the research projects making up the research program

Nexting on a Reinforcement-Learning Robot

The term "nexting" has been used by psychologists to refer to the propensity of people and many other animals to continually predict what will happen next in an immediate, local, and personal sense. When we hear a melody we predict what the next note will be or when the next downbeat will occur, and are surprised and interested when our predictions are disconfirmed. When we read a sentence we guess what the next word will be, or how the sentence will end. When we see a bird in flight, hear our own footsteps, or handle an object, we continually make and confirm multiple predictions about our sensory input. In all these examples, we continually predict what will happen to us *next*. When nexting, an individual may be predicting many or all of their sensory inputs, and at multiple time scales. When we read, for example, it seems likely that we next at the letter, word, and sentence levels, each involving a substantially different time scale. Nexting can be seen as the most basic kind of prediction, preceding and possibly underlying all the others. That people and a wide variety of animals learn and make simple predictions at a range of short time scales was established so long ago in psychology that it is known as "classical conditioning." Animals seem wired to learn the predictive relationships of their world.

To be able to next is to have a basic kind of knowledge about how the world works in interaction with your body. To be able to learn to next—to notice any disconfirmed predictions and continually adjust your nexting—is to be aware of your world in a significant way. To build a robot that can do both of these things is a natural goal, which we have pursued. Prior attempts to do this can be grouped in two approaches. The first approach is to build a very short-term model of the world's dynamics, either in terms of differential equations or state-transition probabilities. This approach usually ends up being very different from nexting. The second approach, which we pursued on the Critterbot, is to use TD methods to learn long-term predictions directly. The prior work pursuing this approach has almost all been in simulation, and has used table-lookup representations and a small number of predictions. Our work is the first to demonstrate real-time nexting on a physical robot. We showed that thousands of anticipatory predictions at various time-scales can be learned in parallel on a physical robot in real-

time using a reinforcement learning methodology. We used a large feature representation and a standard TD learning algorithm to make real-time predictions about the short-term future of the robot's sensor readings using a consumer laptop. An example is shown in Figure 2. The learning was entirely on-policy and used conventional TD algorithms. In future work we will extend it to off-policy learning and GTD algorithms.

The Horde Architecture for Learning Knowledge Grounded in Sensorimotor Experience

How to learn, represent, and use knowledge of the world in a general sense remains a key open problem in artificial intelligence. There are high-level representation languages based on first-order predicate logic and Bayes networks that are very expressive, but in these languages knowledge is difficult to learn and computationally expensive to use. There are also low-level languages such as differential equations and state-transition matrices that can be learned from data without supervision, but these are much less expressive. Knowledge that is even slightly forward looking, such as 'If I keep moving, I will bump into something within a few seconds' cannot be expressed directly with differential equations and may be expensive to compute from them. There remains room for exploring alternate formats for knowledge that are expressive yet learnable from unsupervised sensorimotor data.

In this project we are pursuing a novel approach to knowledge representation based on the notion of value functions and on other ideas and algorithms from reinforcement learning. In our approach, knowledge is represented as a large number of approximate value functions learned in parallel, each with its own policy, pseudo-reward function, pseudo-termination function, and pseudo-terminal-reward function. Our architecture, called *Horde*, consists of a large set of independent reinforcement learning agents, which we call *demons*. The approach here is similar to that which we have taken in previous years with TD networks and options. Horde differs from TD networks in its more straightforward handling of state and function approximation (no predictive state representations) and in its use of GTD algorithms for off-policy learning, which are considerably more efficient than those used in prior work with TD networks. We have deployed Horde on the Critterbot to learn a variety of predictions and behaviors off-policy and in real-time.

General Atari 2600 Game Playing

Although the artificial intelligence community has achieved superb performance in narrow domains such as Chess or Backgammon, developing agents that can perform well across a wide-variety of tasks has remained challenging. One of the main reasons for this difficulty is the lack of an appropriate and widely agreed upon performance metric for measuring general agent ability. This project uses the large number of Atari 2600 video games as a platform for this purpose. The Atari 2600 was the dominant video game console during the late 1970s and early 1980s. With the prevalence of high quality open-

source Atari emulators, as well as 25 years of Moore's Law, it is now feasible to use the set of Atari games as an AI testbed. With so many different games, they can be split into training and test sets. Because each Atari game was designed to be interesting to humans, there is little risk of employing techniques that are so general as to be vacuous and, as any technique needs to work across a space of Atari games, there is minimal risk of overfitting to the specific aspects of any particular game. Any technique that improves the performance of an agent across a large set of Atari games is more likely to be relevant to the broader goals of Artificial Intelligence.

We focused on methods that can play arbitrary Atari 2600 console games without game-specific assumptions or prior knowledge. Two main approaches were considered: reinforcement-learning methods and search methods. The reinforcement-learning methods used feature vectors generated from the game screen as well as the console RAM to learn to play a given game. The search-based methods used the emulator to simulate the consequence of actions into the future, aiming to play as well as possible by only exploring a very small fraction of the state-space. To insure the generic nature of our methods, all agents were designed and tuned using four specific games. Once the development and parameter selection was complete, the performance of the agents was evaluated on a set of 50 randomly selected games. Significant learning was found for the reinforcement-learning methods on most games. Additionally, some instances of human-level performance was achieved by the search-based methods.

Autostep

Many of the learning algorithms used in the project have parameters that must be tuned manually for good performance; we are always looking for ways that they can be set automatically. Foremost among these parameters is the step-size parameter of stochastic gradient-descent algorithms. Several methods have been proposed for automatically setting step-size parameters, but unfortunately all of them have at least one parameter of their own, and this meta-parameter must generally be tuned manually to the particular problem, thereby limiting the benefit of these methods. This year we carried out a major study of algorithms for automatic step-size adaptation, and ultimately produced a new algorithm, called *Autostep*, with no parameters or meta-parameters, that is, that can be applied with no domain knowledge other than what is needed to formulate a problem as gradient descent. We stress tested this algorithm by applying it to Critterbot data. We asked it to make simple short-term predictions of all of the robot's the sensory variables. Autostep was able to find step-size parameters for all these predictions with no tuning.

Autostep and its predecessor algorithms are all designed for conventional supervised learning, not reinforcement learning. A goal for next year will be to extend them to TD and GTD learning, and then to use them to improve nexting and learning in Horde. If this is successful, we then hope to use Autostep to evaluate the utility of representational components such as the features of linear function approximation. This seems a promising method for directing representation discovery.

4. OBJECTIVES FOR NEXT YEAR

With GQ(λ), our development of gradient TD methods is almost but not quite complete. There is some evidence that GTD methods are still not as efficient as classical methods on problems where classical methods are applicable. Some of this perhaps cannot be helped. Nevertheless, we would like an algorithm that can learn like classical methods when those methods are sound and like GTD methods for off-policy situations, and we continue to seek such a hybrid algorithm.

We expect an important focus for our research in the coming year will be on discovering state representations from sensorimotor interaction. The problem of discovering useful representations for learning systems has been recognized as crucial for half a century. Progress has been slow, but we hope to be more successful by leveraging off of the rich stream of data provided by the Critterbot's sensorimotor experience. This data provides many subtasks in the form of sensory signals to predict and control, all of which are due to the same physical system. A multi-task setting such as this may be key to making progress on representation learning. In past work the construction of tasks and their interrelatedness has been largely artificial and therefore suspect. Our robot, we believe, will make it easy to automatically create large numbers of naturally interrelated prediction and control tasks.

In January 2011, the RLAI project began a new initiative, funded by MITACS, the Alberta Ingenuity Centre for Machine Learning, and the Glenrose Rehabilitation Hospital, to introduce an interactive mobile robot component to the new Building Trades of Alberta Courage Center in the Glenrose Rehabilitation Hospital. The Courage Center is a convergence point of rehabilitation therapy and new technologies, including robotics, virtual reality, simulation, and bionics. Robotic devices are becoming more affordable and more widely used in physiotherapy. Robots can significantly enhance conventional therapy by compensating for physical disabilities of the patient, by providing additional feedback to the patient and therapist, and by motivating the patient to continue the therapy by himself (e.g., at home). Robotic technology includes mobility aids for locomotion and navigation as well as prostheses and other manipulation aids.

The primary objective of the Glenrose initiative is to design, build, and program a mobile robot that can interact successfully, without supervision, with members of the public visiting the Courage Center. The robot is expected to interact with and engage visitors through its motion, sounds, or light displays. The robot's means of sensing people and other aspects of its environment will include a camera and a microphone. Visitors will be able to play with the robot and to punish or reward it to change its behavior. As the robot hardware needs to be extremely robust and stable, it will be based on the iRobot Create, a commercially-manufactured research platform, enhanced with additional sensors, computer, and learning software. A docking station will enable the robot to autonomously recharge itself. This robot will help investigate the medical, technological and social challenges in robot rehabilitation therapy.

5. RESEARCH TEAM MEMBERS AND CONTRIBUTIONS

a. Team Leader

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

b. Faculty Team Members

Michael Bowling	Faculty team member	AICML, NSERC
Dale Schuurmans	Faculty team member	AICML, CRC II Chair, NSERC, MITACS
Csaba Szepesvári	Faculty team member	AICML, AIF New Faculty, NSERC
Martin Mueller	Associated faculty member	DARPA, NSERC

c. Post Doctoral Fellows & Research Associates

Name	Role	Awards / Special Info
Elliot Ludvig	Research associate	Until January 2011, now an Associate Research Scholar at Princeton
Thomas Degris-Dard	Post doctoral fellow	MITACS
Joseph Modayil	Post doctoral fellow	
David Pal	Post doctoral fellow	
Patrick Pilarski	Post doctoral fellow	NSERC
Istvan Szita	Post doctoral fellow	
Joel Veness	Post doctoral fellow	
Xinhua Zhang	Post doctoral fellow	

d. PhD Candidates

Name	Role	Scholarships / Awards / Special Info
Yasin Abbasi-Yadkori	PhD candidate	
Arash Afkanpour	PhD candidate	iCORE ICT
Gabor Balazs	PhD candidate	
Nolan Bard	PhD candidate	iCORE ICT, NSERC PGS, President's Doctoral Prize of Distinction
Gabor Bartok	PhD candidate	
Katherine Chen	PhD candidate	
Kenneth Dwyer	PhD candidate	
Amir massoud Farahmand	PhD candidate	
Alireza Farhangfar	PhD candidate	iCORE ICT
Marc Gendron-Bellemare	PhD candidate	iCORE ICT
Ruitong Huang	PhD candidate	Provost Doctoral Entrance Award
Michael Johanson	PhD candidate	iCORE ICT
Anna Koop	PhD candidate	NSERC PGS
Marc Lanctot	PhD candidate	iCORE ICT
Hamid Reza Maei	PhD candidate	iCORE ICT
Ashique Mahmood	PhD candidate	Provost Doctoral Entrance Award
Gergely Neu	PhD candidate	
Aslan Ozlem	PhD candidate	
James Neufeld	PhD candidate	iCORE ICT
Chris Rayner	PhD candidate	iCORE ICT, NSERC PGS
Yi Shi	PhD candidate	
Brian Tanner	PhD candidate	
Adam White	PhD candidate	NSERC CGS
Martha White	PhD candidate	AIF
Min Yang	PhD candidate	
Hengshuai Yao	PhD candidate	
Yaoliang Yu	PhD candidate	

e. MSc Candidates

Name	Role	Scholarships / Awards / Special Info
Michael Delp	MSc candidate	iCORE ICT
Leah Hackman	MSc candidate	
Michael Joya	MSc candidate	
Andre Luzardo	MSc candidate	ELAP Scholarship
Bernardo Ávila Pires	MSc candidate	

f. Other Members

Name	Role
Beverly Balaski	Administrative assistant at Oct 2010
Akiko Green	Administrative assistant until Aug 2010
Lori Troop	Program administrator
Michael Shirt	High school summer student, July-August 2010
Jason Ye	High school summer student, July-August 2010

g. Visitors

Name	Institution
Mathukumalli Vidyasagar	University of Texas at Dallas
Ihab Ilyas University of Waterloo	University of Waterloo
Andras Antos	MTA Sztaki
Abdeslam Boularias	Laval University
Mike Sokolsky	Stanford University
Xin Xu	National University of Defense Tech, China
Nando de Freitas	UBC
Yoshua Bengio	Universite de Montreal
Raymond Hsu	Stanford University

6. COLLABORATIONS

Provincial	
Participants	Nature of Collaboration
Alberta Ingenuity Center for Machine Learning	R. Sutton, D. Schuurmans, Cs. and Szepesvári M. Bowling are among the ten principle investigators for this center at the University of Alberta. Total annual funding for AICML is $2M.
Elliot Ludvig, Marcia Spetch, Matt Brown, Brian Buck and Chris Madan all of University of Alberta	Studies of Choice Behavior in Pigeons and Humans. Funded by Grant from Alberta Gaming Research Institute. 10K.
Patrick Pilarski, and Jason Carey, Department of Computing Science, University of Alberta,	"Reinforcement Learning for Adaptive Prosthetics" This project is a collaboration between AICML, the Glenrose Rehabilitation Hospital, and the Composite & Biomedical Materials Research Group.
Patrick Pilarski and Wilsun Xu, Department of Electrical and Computer Engineering, University of Alberta	Preliminary collaboration in the application of reinforcement learning to specific challenges in power systems engineering, both on a household scale and on the scale of provincial power generation and distribution.
Thomas Degris-Dard, MITACS, Glenrose Rehabilitation Hospital	A collaboration to introduce an interactive mobile robot component to the new Building Trades of Alberta Courage Center in the Glenrose Rehabilitation Hospital

National	
Participants	Nature of Collaboration
Doina Precup (PI), McGill University Prakash Panangaden, McGill University Yoshua Bengio, Montreal University Shie Mannor, McGill University Richard Sutton, University of Alberta	NSERC Collaborative Research and Development Grants-Project, "Learning and Prediction in High Dimensional Stochastic Domains," with Nortel Networks and Bell Canada. $186,523 total funding (Sept 1/06 – Aug 31/10).
Shai Ben-David, University of Waterloo Yoshua Bengio, McGill University Hugh Chipman, University of Waterloo Dale Schuurmans, University of Alberta Richard Sutton, University of Alberta Pascal Vincent, Montreal University	MITACS Grant, "Statistical Learning of Complex Data with Complex Distributions" to Dale Schuurmans, and the listed collaborators. $91K total to Dale Schuurmans (Apr 1/09-Mar 31/11).

International	
Participants	Nature of Collaboration
E. James Kehoe, University of New South Wales, Australia	Joint research with Elliot Ludvig on creating and testing new reinforcement learning models of classical conditioning in animals.
Fuat Balci, Princeton University	Joint research with Elliot Ludvig on empirical and computational studies of timing in non-human animals.
Rong Zheng, University of Houston	Joint research with Csaba Szepesvári on application of online learning techniques in networking
Tiberio Caetano, NICTA	Joint research with Dale Schuurmans on machine learning in structured-output prediction problems.
Li Cheng, TTI-Chicago	Joint research with Dale Schuurmans on discriminative structured learning models for computer vision.
Yuhong Guo, Temple University	Joint research with Dale Schuurmans on machine learning in bioinformatics.
Andras gyorgy, MTA SZTAKE, Hungary	Joint research with Csaba Szepesvári on the theory of online learning in MDPs3
G. Tesauro, R. Segal, J. Marecki, M. Mueller, Csaba Szepesvari, A. MacGovern, A. Barto, W. Dabney, M. Boddy, C. Rosin	DARPA, via IBM subcontract/GALE program, total funding: $868,942
Andras Antos, MTA SZTAKI, Hungary	Joint research with Csaba Szepesvári on active learning.
Olivier Cappé, CNRS	Joint research with Csaba Szepesvári on scaling up online learning algorithms and applications to telecommunication

7. GRADUATES

Name	Degree	Research topic	Current Position
Ashique Mahmood	MSc	Automatic Step-Size Adaptation in Incremental Supervised Learning	PhD candidate with the group
Michael Delp	MSc	Experiments in Off-policy Reinforcement Learning with the GQ(λ) Algorithm	Researcher with Toyota Technical Center, Ann Arbor, MI, USA

8. INTELLECTUAL PROPERTY

None.

Intellectual Property	Status	Short Description
PATENTS		
LICENSES		
Rebel Entertainment	New	Michael Bowling entered into an option agreement to license the media rights to Polaris, our computer program for playing heads-up limit Texas hold'em

9. PUBLICATIONS

REFEREED JOURNAL PUBLICATIONS

F. Balci, E. A. Ludvig, and D. Brunner, "Within-session Modulation of Timed Anticipatory Responding: When to Start Responding," *Behavioural Processes*, Jun. 2010, pp. 204-206

P. Beeson, J. Modayil, and B. Kuipers, "Factoring the Mapping Problem: Mobile Robot Map-Building in the Hybrid Spatial Semantic Hierarchy," *International Journal of Robotics Research*, Apr. 2010, 4, pp. 428—459

E. J. Kehoe, E. A. Ludvig, and R.S. Sutton, "Timing in Trace Conditioning of the Nictitating Membrane Response of the Rabbit (Oryctolagus cuniculus): Scalar, Nonscalar, and Adaptive Features," *Learning & Memory*, Dec. 2010, pp.600-604

E. A. Ludvig, F. Balci, and M. L. Spetch, "Reward Magnitude and Timing in Pigeons," *Behavioural Processses*, to appear

L. Mukhergee, L. Xu, D. Schuurmans, J. Peng, V. Singh, L. Xu, L. Mukherjee, J. Peng, V. Singh, and D. Schuurmans, "An Efficient Algorithm for Maximal Margin Clustering," *Journal of Global Optimization* (JOGO), to appear

P. M. Pilarski, and C. J. Backhouse, "Towards Robust Cellular Image Classification: Theoretical Foundations for Wide-angle Scattering Pattern Analysis," *Biomedical Optics Express*, Oct. 2010, 4, pp. 1225-1233

D. Schuurmans, T. Caelli, L. Cheng, and M. Gong, "Real-time Discriminative Background Subtraction," *IEEE Transactions on Image Processing*, to appear

REFEREED ARCHIVAL CONFERENCE PROCEEDINGS

Y. Abbasi-Yadkori, J. Modayil, and Cs. Szepesvari, "Extending Rapidly-Exploring Random Trees for Asymptotically Optimal Anytime Motion Planning," *IEEE International Conference on Intelligent Robots and Systems*, Oct. 2010, 50% acceptance

G. Bartok, D. Pal, and Cs. Szepesvari, "Toward a Classification of Finite Partial-Monitoring Games," *Algorithmic Learning Theory* (ALT), Oct. 2010, 60% acceptance

S. Bergsma, D. Lin, and D. Schuurmans, "Improved Natural Language Learning via Variance-Regularization Support Vector Machines," *Conf. on Computational Natural Language Learning* (CoNLL), Jul. 2010, pp. 172-181. 25% acceptance

A. M. Farahmand, R. Munos, and Cs. Szepesvari, "Error Propagation for Approximate Policy and Value Iteration," *Neural Information Processing Systems* (NIPS), Dec. 2010, to appear, 24% acceptance

W. Li, L. Xu, and D. Schuurmans, "Facility Locations Revisited: An Efficient Belief Propagation Approach," *IEEE Int'l Conf. on Automation and Logistics* (ICAL), Aug. 2010, pp. 408-413. 52% acceptance

H. Nakhost and M. Müller, "Action Elimination and Plan Neighborhood Graph Search: Two Algorithms for Plan Improvement," *20th Intl Conf. on Automated Planning and Scheduling* (ICAPS 2010), 34% acceptance

G. Neu, A. György, A. Antos, and Cs. Szepesvári, "Online Markov Decision Processes under Bandit Feedback," *Neural Information Processing Systems* (NIPS), Dec. 2010, to appear, 6% acceptance

G. Neu, A. Gyorgy, and Cs. Szepesvari, "The Online Loop-free Stochastic Shortest-Path Problem," *23rd Annual Conf. on Learning Theory* (COLT), Jun. 2010, pp. 231—243, 32% acceptance

D. Pal, B. Poczos, and Cs. Szepesvari, "Estimation of Renyi Entropy and Mutual Information Based on Generalized Nearest-Neighbor Graphs," *Neural Information Processing Systems* (NIPS), to appear. 24% acceptance

Y. Shi, Y. Guo, G. Lin, D. Schuurmans, "Kernel-based Gene Regulatory Network Inference," *Int'l Conf. on Computational Systems Bioinformatics* (CSB), Aug. 2010, pp. 156-165. 26% acceptance

R. S. Sutton, J. Modayil, M. Delp, T. Degris, P. M. Pilarski and A.White, "Horde: A Scalable Real-time Architecture for Learning Knowledge from Unsupervised Sensorimotor Interaction," *10th Int'l Conf. on Autonomous Agents and Multiagent Systems* (AAMAS), to appear, 22% acceptance

Cs. Szepesvari, R. Zheng, and A. Pallavi, "Sequential Learning for Optimal Monitoring of Multi-channel Wireless Networks," *INFOCOMM*, to appear, 16% acceptance

P. Torma, A. Gyorgy, and Cs. Szepesvari, "A Markov-Chain Monte Carlo Approach to Simultaneous Localization and Mapping," *Proc. of the 13th Int'l Conf. on Artificial Intelligence and Statistics* (AISTAT), May. 2010, pp. 605–612, 41% acceptance

S. V. N. Vishwanathan, A. Saha, and X. Zhang, "New Approximation Algorithms for Minimum Enclosing Convex Shapes," *ACM-SIAM Syposium on Discrete Algorithms* (SODA), Jan. 2011, 30% acceptance

M. White, and A. White, "Interval Estimation for Reinforcement-Learning Algorithms in Continuous-State Domains," *Neural Information Processing Systems* (NIPS), Jan 2011, pp. 2532-2540, 24% acceptance

Y. Yu, M. Yang, L. Xu, M. White, and D. Schuurmans, "Relaxed Clipping: A Global Training Method for Robust Regression and Classification," *Neural Information Processing Systems* (NIPS), Jan. 2011, pp. 9, 24% acceptance

X. Zhang, T. Graepel, and R. Herbrich, "Bayesian Online Learning for Multi-label and Multi-variate Performance Measures," *Int'l Conf. on Artificial Intelligence and Statistics* (AISTATS), May. 2010, pp. 956--963, 40% acceptance

X. Zhang, A. Saha, and S. V. N. Vishwanathan, "Lower Bounds on Rate of Convergence of Cutting Plane Methods," *Neural Information Processing Systems* (NIPS), Dec. 2010, 24% acceptance

S. Whiteson, B. Tanner, M. E. Taylor, and P. Stone, "Protecting Against Evaluation Overfitting in Empirical Reinforcement Learning," *2011 IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning* (ADPRL), to appear, 70% acceptance

J. Modayil, "Discovering Sensor Space: Constructing Spatial Embeddings That Explain Sensor Correlations," *Int'l Conf. on Development and Learning* (ICDL), Aug. 2010, pp. 120—125. 30% acceptance

BOOKS and CHAPTERS

E. A. Ludvig, M. Gendron-Bellemare and K. G. Pearson, "A Primer on Reinforcement Learning in the Brain: Psychological, Computational, and Neural perspectives," in E. Alonso, E. Mondragon (Eds.), *Computational Neuroscience for Advancing Artificial Intelligence: Models, Methods and Applications* (pp. 111-144). Hershey, PA: IGI Global.

Elliot A. Ludvig, (in press) "Reinforcement Learning in Animals," to appear in N. Steel (Ed.), *The Encyclopedia of the Sciences of Learning*, New York: Springer

Csaba Szepesvari, *Algorithms for Reinforcement Learning*, Jul. 2010, edited by Morgan and Claypool Publishers

Csaba Szepesvari, "Reinforcement Learning Algorithms for MDPs," *Wiley Encyclopedia of Operations Research*, edited by Wiley, to appear

OTHER PUBLICATIONS

M. Bowling, "Bet on the bot - Will Polaris do for Poker what Deep Blue did for Chess?" *The Economist*

S. Jiampojamarn, K. Dwyer, S. Bergsma, A. Bhargava, Q. Dou, M. Kim, and G. Kondrak, "Transliteration Generation and Mining with Limited Training Resources," *ACL 2010 Named Entities Workshop* (NEWS), Jul. 2010

E. A. Ludvig, M. S. Mirian, R. S. Sutton, and E. J. Kehoe, "Associative Learning from Replayed Experience," *Annual Conference of the Pavlovian Society*, to appear

Joseph Modayil, "Learning Grounded Communicative Intent from Human-Robot Dialog," Robotics Science and Systems Workshop (Towards Closing the Loop: Active Learning for Robotics), *AAAI Fall Symposium on Dialog with Robots*, Nov. 2010

M. Ponsen, M. Lanctot, and S. de Jong, "MCRNR: Fast Computing of Restricted Nash Responses by Means of Sampling," *2010 AAAI Interactive Decision Theory and Game Theory (IDTGT) Workshop*, Jul. 2010

N. Quadrianto, D. Schuurmans, and A. Smola, "Distributed flow Algorithms for Scalable Similarity Visualization," 2010 *Workshop on Optimization Based Methods for Emerging Data Mining Problems*, Dec. 2010

E. A. Ludvig, M. S. Mirian, R.S. Sutton, and E. J. Kehoe, "Associative Learning from Replayed Experience," *Annual Conference of the Pavlovian Society*, to appear

SPECIAL/INVITED PRESENTATIONS

Date	Title	Venue
4/6/10	Keynote at the 5th Bellairs workshop on reinforcement learning: Representational and perceptual learning	Barbados Workshop on Reinforcement Learning
4/6/10	Perceptual Learning in Humans and Other Animals	Barbados Workshop on Reinforcement Learning
4/4/10	Off Policy Actor Critic for Low Level Control with the Critterbot	Barbados Workshop on Reinforcement Learning
4/6/10	Toward Adaptive Reinforcement Learning Algorithms	Barbados Workshop on Reinforcement Learning
4/6/10	Feature and Variable Selection: Context for Representation Learning in RL	Barbados Workshop on Reinforcement Learning
4/7/10	Computational Curiosity	Barbados Workshop on Reinforcement Learning
4/18/10	Toward Adaptive Reinforcement Learning Algorithms	MIT
5/26/10	Understanding the Sensorimotor World	MITACS 2010
5/26/10	Efficient model estimation in the presence of outliers	MITACS 2010
5/26/10	A Real Time Prediction Machine: Timing, Reinforcement Learning, and the Brain	Princeton University
6/6/10	Lemonade Stand Game: Panel Discussion	ACM EC Trading Agent Design and Analysis
6/8/10	AI After Dark: Computers Playing Poker	MIT
6/25/10	Local Regret	Yahoo! Research
7/17/10	Algorithms of Reinforcement Learning	AAAI Conference
7/19/10	A New Family of Stable and Efficient Methods for Temporal-Difference Learning with Function Approximation Based on Gradient Descent	MIT
7/22/10	A New Family of Stable and Efficient Methods for Temporal-Difference Learning with Function Approximation Based on Gradient Descent	University of California at Berkeley
8/18/10	Kernel-based Gene Regulatory Network Inference	Stanford University
8/20/10	Reinforcement Learning and artificial Intelligence	AITF Banff Summit
8/26/10	Data Mining for Smart Medical Logic	Bosch Healthcare, Palo Alto, California.
9/13/10	New Algorithms for Off Policy Reinforcement	Stochastic approximation:

	Learning and Beyond	methodology, theory and applications in statistics (Workshop, Bristol)
9/15/10	Data Mining for Smart Medical Logic	Bosch Research and Technology Center
9/16/10	New Algorithms for Off-policy Reinforcement Learning	University of Southampton, School of Electronics and Computer Science
10/11/10	How to Choose Cakes (if you must?)-- Advice from Statistics	Vanderbilt University
10/26/10	A Real-time Prediction Machine: Timing, Reinforcement Learning, and the Brain	University of Lethbridge
11/3/10	Reinforcement Learning Artificial Intelligence for Robotic Adaptation to New Environments	ICE: The Tech Conference Alberta Council of Technologies Machine Learning Conference
11/05/10	AI After Dark: Computers Playing Poker	University of Texas at Austin
11/10/10	How to Choose Cakes (if you must?)-- Advice from Statistics	University of Texas at Austin
11/12/10	Toward the Classification of Finite Partial Monitoring Games	University of Houston

AWARDS

Wenye Li, Linli Xu, and Dale Schuurmans received the Best Paper in Logistics (IEEE ICAL 2010) award for the following paper:

W. Li, L. Xu, and D. Schuurmans, "Facility Locations Revisited: An Efficient Belief Propagation Approach," *IEEE Int'l Conf. on Automation and Logistics* (ICAL), Aug. 2010, pp. 408-413, 52% acceptance

THESES

Ashique Mahmood, "Automatic Step-Size Adaptation in Incremental Supervised Learning," September 30, 2010.

Michael Delp, "Experiments in Off-policy Reinforcement Learning with the GQ(λ) Algorithm," January 2011.

10. OUTREACH

Csaba Szepesvari supervised two high school students in July and August as part of the High School Internship Program at the Department of Computing Science

Patrick Pilarski gave a hands-on robotics and artificial intelligence demonstration and lecture for grade 8-12 students participating in the University of Alberta DiscoverE Leadership and Engineering Applications, Projects and Possibilities (LEAPP) camp.

Patrick Pilarski served as a panel member for "Off-loading Intelligence—When Machines Decide," Alberta Council of Technologies / Alberta ICT Council Machine Learning Forum, Edmonton, Alberta, June 1, 2010.

Michael Bowling, "Microchips" on CBC Radio
http://www.cbc.ca/ideas/episodes/2010/10/27/microchips/#