A Proposal to AITF iCORE by Richard Sutton to Fund

The Reinforcement Learning and Artificial Intelligence Laboratory

in connection with the University of Alberta,

the Department of Computing Science,

and the Alberta Innovates Centre for Machine Learning

July, 2012

1. Introduction

The Reinforcement Learning and Artificial Intelligence (RLAI) laboratory was founded by an iCORE chair-establishment grant to Richard Sutton in 2003, and renewed in 2008, in both cases with significant contributions from the University of Alberta, the Alberta Innovates Centre for Machine Learning (AICML), and NSERC. The research team currently includes three other principal investigators, Csaba Szepesvari, Michael Bowling, and Dale Schuurmans, and one associated faculty, Martin Müller, all tenured professors in the Department of Computing Science. The overall team, including postdoctoral fellows, students, and staff, consists of about 45 people at any given time.

Research in the RLAI laboratory 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*, a new body of theory and techniques for optimal control that has been developed in the last thirty 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. Some of the recent accomplishments of reinforcement learning include strategic decision-making in Watson[1] (the computer player of Jeopardy that defeated the best human players[2]), achieving autonomous acrobatic flight of computer controlled helicopters,[3] and widespread application to commercial advertisement placement on the internet.[4]

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, to develop reinforcement learning as a model of intelligence that could approach human abilities, and to explore areas of application of reinforcement learning technology. The major scientific achievements of the lab in the last five years include 1) creation of a new family of learning algorithms that solve two long-standing open problems in reinforcement learning, 2) demonstration of a robot that learns more about its interaction with the world, in real time, than any other physical robot, 3) publication of a new book on reinforcement learning algorithms,[5] 4) creating a super-human computer player of Heads-Up Limit Texas Hold'em Poker,[6] and 5) the first defeat of an un-handicapped top professional Go player by a computer.[7] The RLAI team's scientific accomplishments are described in more than 160 refereed publications in the archival scientific literature over the last five years.

In addition to scientific accomplishments, RLAI research has spilled over into several application areas of commercial and societal importance. The most important of these are in rehabilitative robotics, where the laboratory has been working closely with the Glenrose Rehabilitation Hospital in Edmonton. In one project, we have fielded a small autonomous mobile robot in the Courage Centre at the Glenrose that influences the behaviour of patients and visitors to the hospital. In another, we have applied our machine learning algorithms to a human amputee with a robotic prosthetic arm. Our main accomplishment in the prosthetics area is a demonstration of improved control of the robotic arm in tests with able-bodied subjects. Also in this general area, we have joined and begun collaborating with the "SMART" team, a major AIHS-funded interdisciplinary effort to develop smarter neural prostheses led by Dr Vivian Mushawar of the Department of Cell Biology at the University of Alberta, with team members from throughout Alberta and North America.

The area in which RLAI research is most likely to first have economic impact is through our collaboration with the Alberta Innovates Centre for Machine Learning (AICML), which is developing the "Nurse Rotation Tool", artificial-intelligence-based software for rapidly producing schedules for nurses that are compliant and of minimal cost. This is a problem brought to us by Alberta Health Services (AHS) that could have enormous economic impact. There are approximately 25,000 nurses in Alberta and 268,000 nationally; nursing salaries and related expenses are one of the largest costs of the health care system. AHS has provided $1M for commercialization to AICML, approximately three-quarters of which is dedicated to developing the Nurse Rotation Tool. The first version of the Tool has been submitted for evaluation to the Rotation Management Office at AHS, and parts of it are currently in use. The Nurse Rotation Tool is expected to be completed and deployed by the end of this year.

The RLAI laboratory has joined AICML in creating a new corporate entity, *10pi Incorporated*, in order to facilitate commercialization. 10pi Inc. is three-quarters owned by the principal investigators of AICML and RLAI, and one-quarter owned by the University of Alberta and Tech Edmonton. The process of creating 10pi was begun in 2010 and completed with its incorporation in February, 2012. 10pi has reduced the time to license technology from 22 months to only days. The Nurse Rotation Tool will ultimately be commercialized through 10pi.

Because of the many areas in which machine learning can have commercial impact, the RLAI team has developed substantive collaborations with a number of external companies, including 1) Causata, a London-based company based on adaptive real-time customer interaction, which uses learning algorithms based on those developed in the RLAI laboratory, and where former RLAI PhD student David Silver consults and has worked part time, 2) Deepmind Technologies, a singularity-oriented startup company, also in London, where Sutton is on the Advisory Board, and former RLAI postdoc Joel Veness is an employee, 3) Toyota, where colleague Michael James is a principal in their research laboratory, and former RLAI MSc student Michael Delp is an employee, 4) Google, where former RLAI team members Istvan Szita, David Pal, Martin Zinkevich and Finnegan Southey are employees, 5) IBM, where Gerry Tesauro, principal researcher and one of the creators of Watson, is a colleague with whom we collaborated in our work with Computer Go. We have also placed RLAI alumni at Microsoft (Lihong Li and Jiayuan Huang, research scientists, and Neda Mirian, software developer), Fire Plan Strategies (Winnipeg, Brian Tanner, principal), AiLive (Palo Alto, Dana Wilkinson, research scientist), TheFind.com (Mountain View, Qin Iris Wang, research scientist), and Taobao.com (China, Feng Jiao, senior research scientist).

Finally, one of the informal goals of the RLAI laboratory has been to become one of the strongest research groups in the field of reinforcement learning, and to become known as such. This is a challenging goal because the field is multi-disciplinary, large, and has been growing rapidly. Evidence of the fields multi-disciplinary nature is provided by the many special issues of scientific journals devoted to reinforcement learning, not just in machine learning but also in engineering, neuroscience, and psychology.[8] Evidence of the size and growth of the field is provided by Google scholar, which shows a rapid and sustained increase in the number of scholarly articles per year, roughly tripling since the RLAI laboratory's founding in 2003.[9] Evidence of the scientific prominence of the RLAI laboratory is more subjective, but still strong. Principal investigator Csaba Szepesvari is perhaps the most well known theorist in the reinforcement learning, and the chair, Sutton, is perhaps the single most well known researcher in the field. Sutton's prominence is in part due to his being the first author of the field's most popular textbook. The second author of the textbook, Andrew Barto, has also co-led a large group, at the University of Massachusetts, but has retired this year. Other groups conducting reinforcement learning research around the world, although numerous, all have either a single prominent researcher, or are only weakly associated with reinforcement learning. Szepesvari and Sutton both have continued to receive invitations to give keynote speeches because of their work in reinforcement learning.[10] Overall, the evidence is consistent with the RLAI laboratory being the largest single group devoted to computational reinforcement learning research in the world.

2. Research Program Planned

This section contains an overview of the proposed research of the RLAI project, dividing it into three main interrelated areas. The first is extensions of core reinforcement learning algorithms; there are many open problems in reinforcement learning, and the RLAI project seeks 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. We feel 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 particularly suited for applications, and on several specific application areas in which we are working or planning to work.

2.1. Core Reinforcement Learning Algorithms

In the last few years the RLAI group has introduced a new family of learning algorithms that solve two of the most important long-standing open problems in reinforcement learning. The distinctive feature of the new algorithms, called gradient-TD algorithms, is that they are temporal-difference (TD) algorithms whose updates are, in expected value, in the direction of the gradient of a scalar objective function. As true stochastic-gradient-descent algorithms, they converge robustly under a much wider range of conditions than the previous reinforcement learning algorithms. The first open problem they solve is convergence with non-linear function approximators (such as neural networks). The second open problem is convergence under off-policy training. Off-policy training is training on data generated from one policy (for selecting actions) while learning about another. Off-policy learning is key to managing the classic tradeoff between exploration and exploration, and to efficient large-scale learning in general.

Gradient-TD algorithms were an important breakthrough, but their development is not complete. Existing gradient-TD methods appear to learn more slowly than conventional TD methods under the restricted settings where the conventional methods are guaranteed to converge. In current and future work we propose to explore “hybrid” methods that gain the best of both classes of methods. Existing gradient-TD methods are also restricted to action-value methods. Extending their benefits to the second large class of reinforcement learning algorithms—those based on explicitly learning the parameters of the action-selection policy—is also proposed as a second key area of research on core reinforcement learning algorithms.

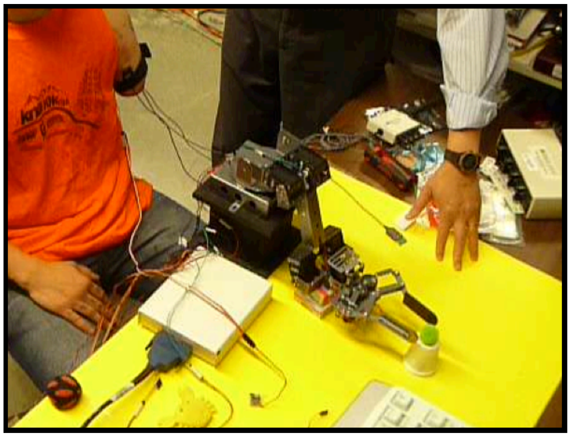
2.2. Addressing the More Ambitious goals of Artificial Intelligence

The creation of robustly converging gradient-TD methods is a key step enabling the learning of large quantities of knowledge, which is widely perceived as key to strong artificial intelligence (AI). Using these algorithms, we propose to learn vast quantities of facts about the interaction between a robot and its environment using gradient-TD learning algorithms and a large-scale parallel architecture we call *Horde*. To date, we have used Horde to learn about 6000 sensorimotor predictive facts ranging from about 0.1 to 8 seconds of anticipation. All 6000 predictions are made and learned every 0.1 seconds. We believe this is the largest amount of interactive sensorimotor knowledge that has ever been learned by a physical robot.

We propose to extend this preliminary work in many directions, large and small, over the next five years. One important step will be to change the behavior of the robot to maximize learning progress, thereby producing a computational version of curiosity. Another will be to use the prediction learning to direct the selection of features and ultimately the construction of all the representational elements of the AI. These are long-standing open problems for AI, and we expect that our interactive setting with the robot will enable us to make new progress on them.

2.3. Applications

In a collaboration with the Glenrose Rehabilitation Hospital, we have fielded a small mobile robot in their Building Trades of Alberta Courage Centre. (This project receives additional funding directly from the Hospital, from AICML, and from Mprime.) The Courage Center is a new facility at the Hospital intended to showcase the use of new technologies, such as robotics, in rehabilitative medicine. Robots are becoming more affordable and more widely used in physiotherapy. They can 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 themselves (e.g., at home). The objective of this project is to explore social interaction between robots and members of the public, including patients. Our robot runs autonomously for months at a time, moving about within the entry hall of the Courage Center. Using the Horde architecture, it learns to predict its sensations and expresses surprise by emitting sounds when exposed to novel sensory experiences. (To minimize disruption, the robot remained quiet on evenings, nights and weekends.) Over a period of time, several patients brought chairs and sat to watch the robot during the day. Although the robot affected people's behaviour in this way, people do not yet affect the robot. We propose to explore a number of new sensors for the robot, including a depth-sensing camera, to make the robot responsive to people.



We are also applying the Horde architecture to learn about a more intimate interaction between people and robots—that between an amputee and their prosthetic limb. The figure to the right shows an amputee at the Glenrose wired up to our Horde learning architecture and to a simple multi-joint robot arm designed and built by our collaborators in the Mechanical Engineering Department. In this common setup, the amputee uses a myoelectric signal from a stump muscle to control one of the joints of the robot arm, switching between joints using a separate toggle switch. The time spent switching is wasted and can be reduced by toggling through the robot arm's joints in an intelligent, user- and task-specific order. We applied Horde to this problem to anticipate which joint the amputee would want next. In a controlled preliminary study with an able-bodied subject, Horde reduced the switching-time overhead by 14%. We propose to extend this study to clinical testing with amputee patients, and to explore other uses of Horde-learned predictions.

The application to Poker is led by principal investigator Michael Bowling. This is a very active, competitive area academically, and also of great interest to the public as the computers have begun to challenge the best human players. As mentioned earlier, the greatest achievement in this area was when Polaris, a program from Bowling's research group, defeated both human teams at the Man-Machine Poker Competition in Las Vegas in 2008.[6] That competition restricted to twoperson Poker with a limited, fixed set of possible bids. For this restricted case, we expect to completely solve the game by computing the optimal minimax Poker strategy by the end of this year. Once that is done, attention will switch to removing the restrictions on the game, considering first No Limit Poker, and then Poker with three or more players. The extension to No Limit is expected to be fairly straightforward, but multi-player Poker involves a host of new issues and is expected to take longer.

The application to Computer Go is led by associated faculty Martin Müller. Müller has developed a software research platform called *Fuego* that has been used to produce reinforcement-learning-based and other competitive Go players. As mentioned earlier, a version of Fuego was the first computer program to defeat a top professional (9-dan) Go player in an even game.[7] That was in 2009 on a 9x9 board, the smallest version of Go that is played commonly by people. It is now generally accepted that computers are probable stronger than people at 9x9 Go, and attention has switched to the full 19x19 Go where people are still stronger but computers are improving rapidly. Fuego's best recent result was winning the 2010 UEC Cup, the biggest Computer Go tournament of that year. The objective of this project is to develop the artificial intelligence technology needed to build the world's best Go player. As in the Poker project, the technology to achieve super-human skill is expected to also be useful for problems of commercial and societal importance.

3. Research Team

The research team is led by the chair, Richard Sutton, three other principal investigators, Csaba Szepesvari, Michael Bowling, and Dale Schuurmans, and Martin Müller, all tenured professors in the Department of Computing Science at the University of Alberta. Currently, they are joined by research associate András György, five postdocs, 25 PhD students, 9 MSc students, and Beverly Balaski, a part-time administrative assistant.

4. Budget

The RLAI laboratory receives support from several other funding sources, including the Alberta Innovates Centre for Machine Learning, NSERC, the Glenrose Hospital, and the University of Alberta, such that approximately 70% of its funding currently comes from sources other than AITF.

It is proposed to continue the core AITF funding for the RLAI laboratory at current levels. This includes funding for the chair/director, three postdocs, eight graduate students, a part-time administrative assistant, computational and robotic equipment, totaling approximately $600K/year. We also include $40K to enable a new multi-disciplinary international meeting focused on reinforcement learning to be held in Alberta in 2015. A detailed budget spreadsheet is attached.

End notes

[1] “Simulation, learning, and optimization techniques in Watson's game strategies,” by Tesauro, G., Gondek, D. C., Lenchner, J., Fan, J., Prager, J. M. *IBM Journal of Research and Development 56*(3.4):1–11, 2012.

[2] “Smarter Than You Think: What Is I.B.M.’s Watson?,” by Clive Thompson. *The New York Times Magazine,* June 16, 2010.

[3] “An Application of Reinforcement Learning to Aerobatic Helicopter Flight,” by Pieter Abbeel, Adam Coates, Morgan Quigley, and Andrew Y. Ng. In: *Proceedings of the Conference on Neural Information Processing Systems 19*, 2007. See http://heli.stanford.edu.

[4] E.g., see “Reinforcement Learning for Online Optimization of Banner Format and Delivery” by Benoit Baccot. In *Online Multimedia Advertising: Techniques and Technologies*, edited by Xian-Sheng Hua, Tao Mei, and Alan Hanjalic. IGI Global 2010.

[5] *Algorithms for Reinforcement Learning*, by Csaba Szepesvari. Morgan and Claypool 2010.

[6] “AI beats human poker champions,” EETimes, News and Analysis, July 7, 2008. http://www.eetimes.com/electronics-news/4077803/AI-beats-human-poker-champions.

[7] E.g., see http://senseis.xmp.net/?Fuego and http://oase.nutn.edu.tw/FUZZ_IEEE_2009/result.htm.

[8] For example:

- 1992, special double issue of *Machine Learning* on “Reinforcement Learning”
- 1996, special triple issue of *Machine Learning* on “Reinforcement Learning”
- 2002, special double issue of *Machine Learning* on “Reinforcement Learning”
- 2010, special issue of *Machine Learning* on “Empirical Evaluations in Reinforcement Learning”
- 2008, special issue of *IEEE Transactions on Systems, Man, and Cybernetic–Part B, Cybernetics* on “Adaptive Dynamic Programming and Reinforcement Learning in Feedback Control”
- July, 1995, special issue of *Robotics and Autonomous Systems* on Reinforcement Learning and Robotics
- 2008, special issue of *Neural Networks* on “Reinforcement Learning of Motor Skills with Policy Gradients”
- 2012, special issue of the *European Journal of Neuroscience* on "Beyond simple reinforcement learning: The computational neurobiology of reward-learning and valuation"
- 2009, special issue of *Cognition* on “Reinforcement learning and higher level cognition”

[9] This is based on the number of google scholar hits to the phrase "reinforcement learning," which (as verified by sampling inspection) tends to be proportional to the number of scholarly papers published in that year. When the RLAI lab was founded in August, 2003, there were about 3000/year, whereas, in July, 2012, there are about 9000/year. There was another tripling from 1996, when there were about 1000/year. For comparison, the phrase "machine learning" received about 50,000 hits/year in recent years, while the phrase "cold fusion" received about 700/year.

[10] In 2011, for example, Sutton gave invited keynote speeches at the

- 25th International Conference on Neural Information Processing Systems, Granada, Spain
- 21st Annual Conference of the Japanese Neural Network Society, Okinawa, Japan
- 21st International Conference on Inductive Logic Programming, Windsor Great Park, UK
- Workshop on Life-long Learning from Sensorimotor Experience, at the 25th International Conference of the Association for the Advancement of Artificial Intelligence, San Francisco, USA

In 2011, for example, Szepesvari gave an invited keynote speech at the

- 9th European Workshop on Reinforcement Learning (the largest regular meeting for the reinforcement learning research community), Athens, Greece

and was the program chair of the

- 22nd International Conference on Algorithmic Learning Theory, Espoo, Finland