AITF ANNUAL REPORT 2014

DR. RICHARD SUTTON

REINFORCEMENT LEARNING AND ARTIFICIAL INTELLIGENCE

AITF ANNUAL REPORT MARCH 31, 2014

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 thirty years primarily within the machine learning and operations research communities, and which has 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 deployment, 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 55 members, 36 of whom are graduate students and, of those, 8 of which are recipients of major scholarships. The output of the research program remained strong this year, with 37 papers published or accepted for publication in refereed archival venues during the reporting period. Three PhD students and three 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) a new algorithmic idea, called *dutch traces*, that improves all algorithms for learning long-term predictions, 2) a new technique, the *interim forward view*, for deriving learning algorithms with specific desired long-term behavior, 3) the first on-line off-policy learning methods with a Monte Carlo equivalence, and 4) demonstration of real-time learning control on iRobot Create robots.

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. We discuss highlights of our research towards each of these target areas in the three paragraphs below.

Our main focus this year has been on the most classical of all reinforcement-learning algorithms, known as TD(λ), which was introduced by Sutton in 1988 and forms the foundation for most modern reinforcement learning algorithms. TD(λ) was thought to be well understood in both theory and practice, but nevertheless we have found this year two ways in which it can be significantly improved. One is in its use of *eligibility traces*, the mechanism by which reinforcement learning methods bridge small temporal gaps between events (the λ in TD(λ) refers to the rate of decay of the traces). There are classically two kinds of eligibility traces, *accumulating traces* and *replacing traces*, each of which is most effective in particular kinds of problems. This year we have developed a new form of trace, which we call *dutch traces*, which performs better than both previous forms, making them obsolete and removing the need for practitioners to choose between them based on problem characteristics.

The second way in which we have improved TD(λ) is in making it more applicable to *off-policy* learning. Recall that off-policy learning (learning about one policy from data using another) is key to our strategy for scaling reinforcement learning methods to address the more ambitious goals of AI. It turns out that a full generalization of TD(λ) to the off-policy case does not exist; even our recent GTD(λ) algorithm does not achieve natural goals such as approximation of Monte Carlo learning when λ=1. This became a significant issue for us when we sought to learn off-policy from robot data for which our state estimation is poor and thus Monte Carlo learning is appropriate. After extensive research, we finally succeeded this year in developing a new version of TD(λ), called PTD(λ), that becomes equivalent to Monte Carlo when λ=1 in off-policy as well as on-policy settings. The ideas behind PTD(λ) (and dutch traces) appear to extend to the full panoply of modern reinforcement-learning algorithms, including Sarsa(λ) and Q-learning.

In the area of applications, we have continued to make progress in robotics and in world-class computer-players of games such as Poker and Go. Our work with adaptive prosthetics has expanded rapidly to include pre-clinical trials with amputees, sensory feedback from the prosthetic robot arm, and new wearable and non-wearable robot arms.

3. RESEARCH PROJECTS

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

Dutch Traces and True Online TD(λ)

The online prediction-learning algorithm TD(λ) is the foundation for many modern reinforcement learning algorithms including Sarsa(λ), LSTD(λ), Q-learning, and the various forms of Q(λ). The "TD" refers to all these methods being Temporal Difference algorithms that "bootstrap," or update their estimates based on other estimates. The parameter λ allows these algorithms to slide smoothly between a pure bootstrapping form (λ=0), in which each estimate is updated toward the single estimate following it, and a "Monte Carlo" form involving no bootstrapping (λ=1), in which each estimate is updated toward the actual subsequent rewards rather than estimates.

In discussing updates toward future rewards and estimates, we are here adopting the *forward view* of these algorithms, which is key to understanding them theoretically. In order to obtain a causal, mechanistic algorithm, the forward view has to be converted to a *backward view* which can be implemented online without knowledge of future events. It is from the backward view that there arises the notion of eligibility traces—of short-term memories of recent events, fading according to the λ parameter, that mark particular states or features as eligible for learning should some subsequent reinforcing event occur. One of the most important results in the field establishes that the backward-view algorithms using eligibility traces are equivalent in their overall effect to their forward views. There are actually two kinds of eligibility traces, accumulating traces and replacing traces. There are equivalence results for both.

All previous equivalences between forward and backward views have been either approximate or restricted to off-line updating. Off-line updating means that updates are computed but not actually used to change estimates until the end of an episode (which resolves all intermediate computations). Updating on every step is actually better and is always used in practice, but such online algorithms are only approximately equivalent to their forward views. If the online algorithm proceeds in small steps, then the approximation can be close, but in practice larger step sizes are preferred and the discrepancy can be large; backward view algorithms could even become unstable. This has been the state of the art for about 20 years.

This year we have made a fundamental advance, creating the first reinforcement learning algorithm, *true-online TD*(λ), that is *exactly* equivalent to its forward view even under *online* updating. The key change in the algorithm is the introduction of a new kind of trace, which we call *dutch traces*. Dutch traces directly improve over both accumulating and replacing traces and remove the need for practitioners to choose a particular form of trace depending on the application.

Interim Forward Views & Off-policy Eligibility Traces

A key step in the advance described in the previous was the development of the idea of an *interim forward view*. A conventional forward view looks ahead to the end of an episode, whereas an interim forward view looks ahead to an arbitrary time horizon within the episode. The most important consequence is an interim version of the conventional forward-view/backward-view equivalence, and that the interim version can be used as a powerful tool to derive new learning algorithms. This year we have developed the interim equivalence idea and used it to answer another long-standing open question in the reinforcement learning: can one generalize eligibility traces to off-policy learning? First let us unpack the question.

Off-policy learning refers to learning about one decision-making policy from data obtained while following a different policy. The policy learned about is called the *target policy*, and the policy used to generate the data is called the *behaviour policy*. The special case in which the two policies are the same, called the *on-policy* case, is an easier learning problem, but the general, off-policy case must be addressed to obtain the full power of reinforcement learning. For example, suppose you are observing a chemical plant while it is controlled by a person (the behaviour policy) but hoping to learn a better policy (the target policy) that you would use if you were allowed to take control of the plant. This is a case of off-policy learning. In general, off-policy learning is important whenever you want to take full advantage of the available data without interfering with an existing controller—a very common case indeed. Off-policy learning arises even in Q-learning (Watkins 1989), perhaps the most popular of all reinforcement learning methods. Much of the RLAI project's research in recent years has been devoted to finding off-policy learning methods that are compatible with function approximation.

Eligibility traces, as previously discussed, refers to the ability of an on-policy algorithm like TD(λ) to slide smoothly between a bootstrapping form (λ=0), which has advantages in terms of learning rate, and a non-bootstrapping, Monte Carlo form (λ=1), which has advantages in terms of asymptotic accuracy and in dealing with non-Markov state representations. There are many off-policy algorithms that use eligibility traces, such as GTD(λ), GQ(λ), LSTD(λ), and various forms of Q(λ), but surprisingly none of them achieve a non-bootstrapping, Monte Carlo form at λ=1.

The open question can thus be stated simply as: Can there be an off-policy Monte Carlo algorithm? Can that algorithm be as computationally efficient as TD(λ)? Can it operate online like TD(λ)? By applying our interim equivalence technique we have been able to answer all these questions in the affirmative. Our new algorithms have a distinctive mechanism: They maintain an additional set of learned weights, called the *provisional* weights, that are used to hold weight changes that might need to be retracted should the actions selected deviate from the target policy. In reference to this distinctive mechanism, the main new algorithms are called PTD(λ) and PQ(λ), off-policy generalizations of TD(λ) and Watkins's Q(λ) respectively. They are the first off-policy learning algorithms with an exact equivalence to Monte Carlo algorithms when λ=1.

Reinforcement Learning in Robot Microworlds

This year we have carried our application of reinforcement learning to robotics an important step further, to include learning and decision-making in real time that affects the behavior of the robot. In past work we have used TD algorithms to learn and update thousands of predictions ten times every second, but the learned predictions did not affect the behavior of the robot in any way. This year we have demonstrated learning control in several settings using a new robotic platform, the iRobot Create, a version of the Roomba vacuum-cleaning robot with the vacuum removed and designed for use by researchers and hobbyists. We have developed software and hardware to make this an robust and inexpensive platform for research.

The figure at right shows a Create robot and a deliberately simple "microworld" environment. The black circular region forms a virtual pen. The robot detects when any of the four downward-looking sensors across its front bumper leave the black region, in which case forward motion is prohibited; the edges of the black region thus form a virtual wall. This and other constraints ensure that the robot never loses contact with the pen, which may be circular, as here, or can take any shape desired. As one example of learning, we tasked the robot to find any way of moving forward as often as possible. Within a few minutes it learned to travel rapidly around the perimeter of the circle with minimal turning.

In another demonstration of control learning, we connected a small camera to a Create robot and to an onboard Raspberry Pi computer serving as its brain. Our GTD(λ) algorithm was used to learn to predict when the robot's front bumper would be activated, and the robot was placed in a physically constrained area, such as a pen with solid walls on four sides. As the robot approached a wall, information from the camera was used to anticipate collision with the wall (and the resultant bumper activation) by about half a second. The initial behavior was to move forward by default, and to turn a random amount after running into a wall; this resulted in experience from which collision predictions could be learned. When the prediction exceeded a threshold, the robot was programmed to turn immediately rather than continuing forward. The result was that, after just a few collisions, the robot learned to turn prior to bumping, thus avoiding collisions. The key benefit of this learning strategy is that it requires no teacher and no prior knowledge about the color of the pen's walls. Once we ran it with a "pen" consisting of the legs of young visiting students seated on the floor. Learning was slower, but the robot eventually learned to avoid bumping into any of the student's legs.

Microworld experiments like these parallel a long tradition in artificial intelligence of studying small simulated worlds, such as games and blocks worlds, that include important issues in a simplified form.

4. OBJECTIVES FOR NEXT YEAR

With regard to core reinforcement learning algorithms, we have an ambitious agenda for the coming year. The first step will be to attempt combine this year's two innovations—dutch eligibility traces and provisional weights—with each other and with the gradient-TD idea, to produce an off-policy learning algorithm with function approximation, convergence guarantees, and on exact online forward-view/backward view equivalence. This will require further development of theory and expertise in interim forward views. Even if successful, this will be far from the end of algorithm development in this area. We also plan further developments in 1) normalization of updates to get the benefits associated with *weighted* importance sampling, 2) variance reduction by taking into account the expected magnitude of the importance-sampling ratios on a step-by-step basis, 3) step-by-step variation in step sizes to shift emphasis in approximation to the states of greatest importance, and 4) *hybrid* gradient-TD methods that incorporate all of these advances.

With regard to the larger ambitions of AI, work will proceed on several fronts. Perhaps most pressing is that we will begin to directly address what might generically be described as "planning"—the use of a model of the world to influence behavior. This has long been a focus of AI research, but it is still rare to do it with models that are learned from data. We now have extensive (albeit still incomplete) experience with off-policy prediction learning methods that could learn models of the appropriate form, and thus it seems time to start trying to put them together with planning methods. We also expect next year to obtain new results using off-policy learning to shape behavior toward the purpose of faster learning, thereby producing a computational analog of the psychological notion of curiosity. It would be natural to do this using an overall measure of learning progress in our *Horde architecture* (a parallel network of learning agents). We also continue to plan to use our new algorithms in support of the automated construction of appropriate representations from and for sensorimotor interaction. Progress in this area has been slow in AI, but we hope to be more successful by leveraging the rich stream of data provided by our robots' sensorimotor experience. Our robots, we believe, will make it easy to automatically create large numbers of naturally interrelated prediction and control tasks. It may still be too early to directly take on this goal, but we will be reaching for it as soon as possible.

With regard to the adaptive prosthetics effort, our primary goal for next year will be to learn from people in real-time and make the benefits of the learning available immediately. Recall that we have algorithms that learn to predict which joints of their prosthetic arm users will want to control next. To date this learning has been done off-line from recorded data. In the coming year we plan to offer the predicted joint to the user as their first choice, or perhaps, if it goes well, even giving it to them without their asking. We also look forward to Patrick Pilarski taking up a position in the Faculty of Medicine and Dentistry and becoming a full principal investigator of the RLAI project.

With regard to Poker, we draw closer to completely solving two-person Limit Poker by computing the optimal minimax Poker strategy.

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, NSERC
Martin Müller	Associated faculty member	DARPA, NSERC
András György	Adjunct Faculty Member; Research Associate	Computer and Research Institute of the Hungarian Academy of Sciences

c. Post Doctoral Fellows & Research Associates

Name	Role	Awards / Special Info
Joseph Modayil	Research Associate	
Patrick Pilarski	Post doctoral fellow, adjunct faculty member	
Harm van Seijen	Post doctoral fellow	
Hado van Hasselt	Post doctoral fellow	
Tor Lattimore	Post doctoral fellow	
Karim Abou-Moustafa	Post doctoral fellow	
Mohamed Elgendi	Post doctoral fellow	
Christopher Archibald	Post doctoral fellow	Now researcher at the Computer Science and Engineering Group Mississippi State University

d. PhD Candidates

Name	Role	Scholarships / Awards / Special Info
Gabor Balazs	PhD candidate	GPA Award
Nolan Bard	PhD candidate	
Neil Burch	PhD candidate	AIF, NSERC CGS, CS GPA Award
Katherine Chen	PhD candidate	
Kenneth Dwyer	PhD candidate	
Pooria Joulani	PhD candidate	2014 Computing Science GPA Award, Department of Computing Science GPA Award
Marc Gendron-Bellemare	PhD candidate	
Ruitong Huang	PhD candidate	
Michael Johanson	PhD candidate	
Anna Koop	PhD candidate	
Ashique Mahmood	PhD candidate	2014 Computing Science GPA Award
Aslan Ozlem	PhD candidate	Provost Doctoral Entrance Award
Bernardo Ávila Pires	PhD candidate	2014 Computing Science GPA Award
James Neufeld	PhD candidate	
Chris Rayner	PhD candidate	Queen Elizabeth II Scholarship
Adam White	PhD candidate	
Martha White	PhD candidate	NSERC CGS D, Honorary Izaak Walton Killam Memorial Scholarship
Farzaneh Mirzazadeh	PhD candidate	NSERC PGSD2
Hengshuai Yao	PhD candidate	
Yaoliang Yu	PhD candidate	
Mohammad Ajallooeian	PhD candidate	
Gergely Neu	PhD candidate	
Junfeng Wen	PhD candidate	

e. MSc Candidates

Name	Role	Scholarships / Awards / Special Info
Hao Cheng	MSc candidate	
Kavosh Asadi Atui	MSc candidate	2014 Computing Science GPA Award
Bing Xu	MSc candidate	
Ujjwal Das Gupta	MSc candidate	GSA Professional Development Award
Roshan Shariff	MSc candidate	Alberta Innovates Graduate Student Scholarship
Mostafa Vafadost	MSc candidate	
Travis Dick	MSc candidate	Queen Elizabeth II Graduate Scholarship Masters Level, Alexander Graham Bell Scholarship M
Joshua Davidson	MSc candidate	
Shun Jie Lau	MSc candidate	
Ryan Kiros	MSc candidate	
Ann Edwards	MSc candidate	
Craig Sherstan	MSc candidate	NSERC
Sanjeev Sharma	MSc candidate	

f. Other Members

Name	Role
Beverly Balaski	Program administrator
Alexandra Kearney	Undergrad researcher
Adam Parker	Undergrad researcher
Brendan Bennett	Undergrad researcher
Darlinton Prauchner	Undergrad researcher
Jacob Denson	High school summer student, July-August 2013
Nicholas Westbury	High school summer student, July-August 2013
Pavlo Malynin	High school summer student, July-August 2013

g. Visitors

Name	Institution
David Pal	Google
Gabor Bartok	ETH Zurich
Erik Talvitie	Franklin and Marshall College
Tor Lattimore	Australian National University
David Taralla	University of Liège

Susan Murphy	University of Michigan
Zhirong Yang	Aalto University
Ruth Urner	University of Waterloo
Xinhua Zhang	NICTA
Dominik Meyer	Technical University of Munich
Johannes Gunther	Technical University of Munich

6. Collaborations

Provincial	
Participants	Nature of Collaboration
Alberta Ingenuity Center for Machine Learning	R. Sutton, D. Schuurmans, Cs. Szepesvári, and M. Bowling are among the nine principle investigators for this center at the University of Alberta. Total annual funding for AICML is $2M
Patrick M. Pilarski and Richard Sutton (RLAI), Linda Pilarski (U of A, Faculty of Medicine), and Carina Debes Marun (Cross Cancer Institute)	Collaboration with researchers from the AHFMR Team Microfluidics, Alberta Innovates Health Solutions, University of Alberta, and Cross Cancer Institute to explore intelligent biomedical image analysis methods to facilitate rapid lab-on-a-chip diagnostics. Part of the technology transfer focused "nanoBridge Research Grant RES-NAN-07-G10: Development of Fluorescence In Situ Hybridization (FISH) Platform, Chip, and Analysis Software" ($145K)
Patrick Pilarski and Richard Sutton (RLAI) and Dr. Kim Adams University of Alberta Rehab Med and Dr. Mahdi Tavakoli Department of Electrical and Computer Engineering University of Alberta	Collaborative work on the use of assistive robots to facilitate the development, skill acquisition, and cognitive assessment of young children and infants with severe motor impairments (e.g., children with cerebral palsy); work involves the use of machine intelligence to enable the automatic adaptation of a robot control system to match the needs and abilities of individual children. Pilarski and Adams co-supervise visiting graduate students from a partner institution in Mexico city.

Patrick Pilarski and Richard Sutton (RLAI), Martin Ferguson-Pell (Dean, Faculty of Rehabilitation Medicine, University of Alberta), Liping Qi (University of Alberta, Rehabilitation Medicine), Simon Grange (University of Calgary / University of Alberta / Alberta Health Services)	Collaboration between the RLAI and the Rehabilitation Robotics Sandbox Laboratory (Rehabilitation Medicine, U. Alberta); this collaboration focuses on using new machine learning methods to predict fatigue in wheelchair users and enable novel muscle stimulation paradigms.
Patrick Pilarski and Richard Sutton (RLAI), and AHFMR SMART Neuroprosthetics Team, University of Alberta	Preliminary work on the use of machine learning to adapt and optimize neural interfaces and assistive robotic devices for people with motor disabilities
L.M. Pilarski (Dept. Oncology, University of Alberta), L.M. McMullen (Agriculture, University of Alberta), M. Gaenzle (Agriculture, University of Alberta), Patrick Pilarski (RLAI), X. Yang (Lacombe Research Centre)	The meat processing industry needs a portable and rapid platform to detect pathogens during meat processing. We work with end users and colleagues to develop inexpensive computer systems that enable the automation of the testing system. This work includes software and hardware automation, biomedical pattern analysis, and machine intelligence. A key part of our system is data analysis to detect E. Coli from the data coming in from the device's sensors---and the presentation of these results in a form that is useful and intuitive to both staff at the point of use.
Dr. Ian Adatia, Director of the Pediatric Cardiac Critical and Intermediate Care Program, Stollery Children's Hospital and Roger Zemp, Department of Electrical and Computer Engineering University of Alberta	Joint research with Dale Schuurmans on adaptive signal processing techniques for heart disease diagnosis, focusing on predicting events with ECG and PCG signals.
Joseph Modayil and Gary Faulkner, Jim Raso, Michael Cimolini from the Glenrose Rehabilitation Hospital	Collaboration with the “Robotics in the Glenrose Courage Centre” project is studying the interaction between people and autonomous robots in a rehabilitative medicine setting.

National	
Participants	Nature of Collaboration
D. Precup, McGill University	Richard Sutton Co-organized the 8th Bellairs Workshop on Reinforcement Learning. Also joint research on off-policy eligibility traces
Daniel Lizotte, University of Waterloo	Joint research with Michael Bowling on reinforcement learning for medical applications

International	
Participants	Nature of Collaboration
Zinhua Zhang, NICTA	Joint research with Dale Schuurmans on representation learning and optimization.
Yuhong Guo, Temple University	Joint research with Dale Schuurmans on machine learning in bioinformatics.
Yaoliang Yu, Carnegie Mellon University	Joint research with Dale Schuurmans on efficient optimization methods for machine learning, convex reformulations of representation learning problems, and algorithmic techniques for exploiting structured sparsity.
Shie Mannor, Technion, Haifa, Israel	Joint research with Csaba Szepesvári on online collaborative filtering.
Andras Antos, MTA SZTAKI, Hungary	Joint research with Csaba Szepesvári on active learning.
Lihong Li, Microsoft Research, Redmond	Joint research with Csaba Szepesvári on off-policy learning.
Branislav Kveton, Technicolor, Palo Alto	Joint research with Csaba Szepesvári on bandit combinatorial optimization
Koby Crammer, Technion, Haifa, Israel	Joint research with Csaba Szepesvári on bandit resource allocation
Mohammad Ghavamzadeh, Adobe Research, San Jose	Joint research with Csaba Szepesvári on classification calibration.
David Pal, Google	Joint research with Csaba Szepesvári on online learning.
Yasin Abbasi-Yadkori, Queensland University of Technology, Brisbane, Australia	Joint research with Csaba Szepesvári on online learning in control.
Erik Talvitie, Franklin and Marshall College	Joint research and supervision of graduate students with Michael Bowling, developing new algorithms for intelligent exploration in domains with sparse rewards.
Nathan Sturtevant, University of Denver	Joint research with Michael Bowling on multiplayer game theory and automating heuristic construction for search.
Marc Lanctot, Maastricht University	Joint research with Michael Bowling on computational game theory.

Patrick Pilarski and Richard Sutton (RLAI), Jason P. Carey (Department of Mechanical Engineering, University of Alberta), Michael R. Dawson, Jacqueline S. Hebert, K. Ming Chan (Glenrose Rehabilitation Hospital)	"Reinforcement Learning for Adaptive Prosthetics;" this collaboration investigates the use of reinforcement learning and real-time machine learning to enable adaptive, intuitive control of myoelectric prostheses and other assistive robotic devices.
Gergely Neu, INRIA Lille	Joint research with Andras Gyorgy on online learning, Markov decision problems, and adaptive source coding.
Tamas Linder, Queen's University	Joint research with Andras Gyorgy prediction in changing environments, structural problems of quantization
Gabor Lugosi, Pompeu Fabra University	Joint research with Andras Gyorgy prediction in changing environments, developing a family of efficient algorithms
Levente Kocsis, Computer and Automation Research Institute of the Hungarian Academy of Sciences	Joint research with Andras Gyorgy, efficient boosting algorithms for ranking
Gergely Neu and Andras Antos of Budapest University of Technology and Economics	Joint research with Andras Gyorgy, Markov Decision Processes with adversarial reward models

7. GRADUATES

Name	Degree	Research topic	Current Position
Ryan Kiros	MSc	Learning Deep Representations, Embeddings and Codes from the Pixel Level of Natural and Medical Images	PhD student at the University of Toronto
Mostafa Vafadost	MSc	Temporal Abstraction in Monte Carlo Tree Search	Software engineer at EMC Corp in Edmonton
Hao Cheng	MSc	Bregman Divergence Clustering: A Convex Approach	PhD student at University of Washington
Yaoliang Yu	PhD	Fast Gradient Algorithms for Structured Sparsity	Postdoctoral researcher at Carnegie Mellon University
Gergely Neu	PhD	Online Learning in Non-Stationary Markov Decision Processes	INRIA Lille - Nord Europe
Marc Gendron-Bellemare	PhD	Fast, Scalable Algorithms for Reinforcement Learning in High Dimensional Domains	Research Scientist, Google DeepMind

8. INTELLECTUAL PROPERTY

Intellectual Property	Status	Short Description
PATENTS	none	
LICENSES		
Spinoff Companies	none	

9. PUBLICATIONS

REFEREED JOURNAL PUBLICATIONS

K. Abou-Moustafa, F. De La Torre, and F. Ferrie, "Pareto Discriminant Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Aug. 2013.

K. Abou-Moustafa, Y. Yasui, J. Scott, D. Guttman, and A. Kozyrskyj, "The Divergence as a diversity measure. Application to gut microbiome analysis," *Statistics in Medicine*, Jan. 2014.

S. Adamia, M. Bar-Natan, B. Haibe-Kains, P. M. Pilarski, S. Pevzner, H. Avet-Loiseau, L. Lode, S. Verselis, E. A. Fox, I. Galinsky, S. Mathews, I. Dagogo-Jack, M. Wadleigh, D. P. Steensma, G. Motyckova, D. J. Deangelo, J. Quackenbush, R. Stone and J. D. Griffin, "NOTCH2 and FLT3 gene mis-splicing are common events in patients with acute myeloid leukemia (AML): New potential targets in AML," *Blood*, Feb. 2014.

S. Adamia, B. Haibe-Kains, P. M. Pilarski, M. Bar-Natan, S. Pevzner, H. Avet-Loiseau, L. Lode, S. Verselis, E. A. Fox, J. Burke, I. Galinsky, I. Dagogo-Jack, M. Wadleigh, D. P. Steensma, G. Motyckova, D. J. DeAngelo, J. Quackenbush, R. Stone, and J. D. Griffin, "A Genome-wide aberrant RNA splicing in patients with acute myeloid leukemia identifies novel potential disease markers and therapeutic targets," *Clinical Cancer Research*, Vol 20(5): 1135-45, Mar. 2014.

A. Afkanpour, Cs. Szepesvári, and M. Bowling, "Alignment Based Kernel Learning with a Continuous Set of Base Kernels," *Machine Learning*, Jun. 2013, pp. 305-324.

M. Elgendi, "Fast QRS detection with an optimized knowledge-based method: Evaluation on 11 standard ECG databases," *PLoS One*, Sep. 2013.

M. Elgendi, P. Bobhate, J. Rutledge, Y. Coe, R. Zemp, D. Schuurmans, and I. Adatia, "Spectral analysis of the heart sounds in children with and without pulmonary hypertension," *Int'l Journal of Cardiology*, to appear

M. Elgendi, I. Norton, M. Brearley, D. Abbott, and D. Schuurmans, "Systolic Peak Detection in Acceleration Photoplethysmograms Measured from Emergency Responders in Tropical Conditions," *PLoS One*, Oct. 2013.

E. J. Kehoe, E. A. Ludvig, R. S. Sutton, "Timing and cue competition in conditioning of the nictitating membrane response of the rabbit (*Oryctolagus cuniculus*)," *Learning and Memory 20*:97-102, 2013. Cold Spring Harbor Press.

L. Kocsis, A. Gyorgy, and A. N. Ban, "BoostingTree: Parallel Selection of Weak Learners in Boosting, with Application to Ranking," *Machine Learning*, Jun. 2013, pp. 293-320.

T. Lattimore, and M. Hutter, "General time consistent discounting," *Theoretical Computer Science*, Jan. 2014, pp. 140-154.

T. L. MacKay, N. Bard, M. Bowling, and D. C. Hodgins, "Do pokers players know how good they are? Accuracy of poker skill estimation in online and offline players," *Computers in Human Behavior*, Feb. 2014, pp. 419-424.

J. Modayil, A. White, R. S. Sutton, "Multi-timescale Nexting in a Reinforcement Learning Robot," *Adaptive Behavior 22*(2):146-160, April 2014.

H. Nakhost and M. Müller, "Towards a Theory of Random Walk Planning: Regress Factors, Fair Homogeneous Graphs, and Extensions," *AI Communications*, accepted.

L. M. Pilarski, C. Debes Marun, L. Martin, C.P. Venner, P. M. Pilarski, and A. R. Belch, "B lymphocytes as potential cancer stem cells in multiple myeloma," *Journal of Oncopathology*, Jul. 2013, pp. 11-22.

P. M. Pilarski, "Aligning homeostatic and heterostatic perspectives," *Constructivist Foundations 9*(2):213–215, Mar. 2014.

Z. Shiller, S. Sharma, I. Stern, and A. Stern, "On-Line Obstacle Avoidance at High Speeds," *Int'l Journal of Robotics Research*, Sep. 2013, pp. 1030-1047.

H. van Seijen, S. Whiteson, and L. Kester, "Efficient Abstraction Selection in Reinforcement Learning," *Computational Intelligence*, Aug. 2013.

HIGHLY REFEREED ARCHIVAL CONFERENCE PROCEEDINGS

Y. Abbasi-Yadkori, P. Bartlett, V. Kanade, Y. Seldin, and Cs. Szepesvári, "Online Learning in Markov Decision Processes with Adversarially Chosen Transition Probability Distributions," *Proc. Conf. Advances in Neural Information Processing Systems* (NIPS 2013), Dec. 2013, 25% acceptance

O. Aslan, H. Cheng, X. Zhang, and D. Schuurmans, "Convex Two-Layer Modeling," *Proc. Conf. Advances in Neural Information Processing Systems* (NIPS 2013) Dec. 2013, 25% acceptance.

N. Bard, M. Johanson, and M. Bowling, "Asymmetric Abstractions for Adversarial Settings," *Proc. 13th Int'l Conf. on Autonomous Agents and Multiagent Systems* (AAMAS 2014), to appear, 24% acceptance.

N. Burch, M. Johanson, and M. Bowling, "Solving Imperfect Information Games Using Decomposition," *Proc. 28th Conf. on Artificial Intelligence* (AAAI 2014), to appear, 28% acceptance.

H. Cheng, X. Zhang, and D. Schuurmans, "Convex Relaxation of Bregman Divergence Clustering," *Proc. Conf. on Uncertainty in Artificial Intelligence* (UAI 2013), Jul. 2013, pp. 162, 31% acceptance.

J. Dauwels, P. Kumar, F. Mahmood, K. Wong, A. Agrawal, M. Elgendi, S. Kannan, D. Menoth, R. Shukla, and A. Chan, "A study on the effect of subliminal priming on subjective perception of images: A machine learning approach," *Proc. 15th Int'l Conf. on Biomedical Engineering* (ICBME 2013), Dec. 2013.

T. Davis, N. Burch, and M. Bowling, "Using Response Functions to Measure Strategy Strength," *Proc. 28th Conf. on Artificial Intelligence* (AAAI 2014), to appear, 28% acceptance.

T. Dick, A. Gyorgy, and Cs. Szepesvári, "Online Learning in Markov Decision Processes with Changing Cost Sequences," *Proc. Int'l Conf. on Machine Learning* (ICML 2014), Jan. 2014, pp. 512-520, 15% acceptance.

M. Gendron-Bellemare, J. Veness, and M. Bowling, "Bayesian Learning of Recursively Factored Environments," *Proc. Int'l Conf. on Machine Learning* (ICML 2013), Jun. 2013, 25% acceptance.

Y. Guo, and D. Schuurmans, "Multi-label Classification with Output Kernels," *Proc. European Conf. on Machine Learning* (ECML 2013), Sep. 2013, pp. 16, 25% acceptance.

R. Huang, and Cs. Szepesvári, "A Finite-Sample Generalization Bound for Semiparametric Regression: Partially Linear Models," *Proc. Int'l Conf. on Artificial Intelligence and Statistics* (AISTATS), 2014.

P. Mazrooei, C. Archibald, and M. Bowling, "Automating Collusion Detection in Sequential Games," *Proc. 27th Conf. on Artificial Intelligence* (AAAI 2013), Jul. 2013.

F. Mirzazadeh, Y. Guo, and D. Schuurmans, "Convex Co-embedding," *Proc. 28th Conf. on Artificial Intelligence* (AAAI 2014), Feb. 2014.

H. Nakhost and M. Müller, "Towards a second generation random walk planner: An experimental exploration," *Proc. Int'l Joint Conf. on Artificial Intelligence* (IJCAI 2013), pp. 2336-2342, 2013. Acceptance rate 28%.

J. Neufeld, A. Gyorgy, D. Schuurmans, and Cs. Szepesvári, "Adaptive Monte-Carlo via Bandit Allocation," *Proc. Int'l Conf. on Machine Learning* (ICML 2014), to appear, 25% acceptance.

R. S. Sutton, R. A. Mahmood, D. Precup, and H. van Hasselt, "A new Q(λ) with interim forward view and Monte Carlo equivalence," *Proc. Int'l Conf. on Machine Learning* (ICML 2014), to appear.

H. van Seijen, and R. S. Sutton, "True Online TD(λ)," *Proc. Int'l Conf. on Machine Learning* (ICML 2014), to appear.

X. Zhang, Y. Yu, and D. Schuurmans, "Polar Operators for Structured Sparse Estimation," *Proc. Conf. Advances in Neural Information Processing Systems* (NIPS 2013), Dec. 2013, 25% acceptance.

N. Zolghadr, Cs. Szepesvári, A. Gyorgy, B. Gábor, and R. Greiner, "Online Learning with Costly Features and Labels," *Proc. Conf. Advances in Neural Information Processing Systems* (NIPS 2013), Dec. 2013, 25% acceptance.

OTHER CONFERENCE AND WORKSHOP PROCEEDINGS

K. Abou-Moustafa, F. Ferrie and D. Schuurmans, "Divergence Based Graph Estimation for Manifold Learning," *Proc. IEEE Global Conf. on Signal and Information Processing* (IEEE GlobalSIP 2013), Dec. 2013, pp. 4.

K. Abou-Moustafa, D. Schuurmans, and F. Ferrie, "Learning a Metric Space for Neighbourhood Topology Estimation. Application to Manifold Lear," *Proc. 5th Asian Conf. on Machine Learning* (ACML 2013), Nov. 2013, pp. 16.

C. Archibald, M. Bowling, and J. Davidson, "The Baseline Approach to Agent Evaluation," *Workshop on Computer Poker and Imperfect Information* (AAAI 2013), Workshop Paper, Jul. 2013.

U. Das Gupta, S. Sriram, S. Sharma, and R. Greiner, "Learning Markov Networks with Bounded Inference Complexity," *Int'l Conf. on Machine Learning* (ICML 2013), Workshop Paper, Jun. 2013.

T. Dick, A. Gyorgy, and Cs. Szepesvári, "Online Learning in Markov Decision Processes with Changing Reward Sequences," *Proc. 1st Multidisciplinary Conf. on Reinforcement Learning and Decision Making* (RLDM 2013), Oct. 2013, pp. 198-201.

A. Edwards, A. Kearney, M.R. Dawson, R.S. Sutton, and P. M. Pilarski, "Temporal-Difference Learning to Assist Human Decision Making during the Control of an Artificial Limb," *Proc. 1st Multidisciplinary Conf. on Reinforcement Learning and Decision Making* (RLDM 2013), Oct. 2013, pp. 1-5.

T. Everitt, T. Lattimore, and M. Hutter, "Free Lunch for Optimisation under the Universal Distribution," *Proc. IEEE Congress on Evolutionary Computation* (IEEE CEC 2014), to appear.

S. Fernando and M. Müller, "Analyzing Simulations in Monte Carlo Tree Search for the Game of Go," *Computers and Games* (CG 2013), 2013.

J. Gunther, P. M. Pilarski, G. Helfrich, H. Shen, and K. Diepold, "First steps toward an intelligent laser welding architecture using deep neural networks and reinforcement learning," *Proc. 2nd Joint Int'l Conf. on System-integrated Intelligence: New Challenges for Product and Production Engineering* (SysInt 2014), to appear.

S.C. Huang, B. Arneson, R. Hayward, M. Müller and J. Pawlewicz, "MoHex 2.0: A pattern-based MCTS Hex player," *Computers and Games* (CG 2013), 2013.

S.C. Huang, and M. Müller, "Investigating the Limits of Monte Carlo Tree Search Methods in Computer Go," *Computers and Games* (CG 2013), 2013.

P. Kumar, F. Mahmood, K. Wong, M. Elgendi, A. Agrawal, S. Kannan, D. Menoth, R. Shukla, J. Dauwels, and A. Chan, "Inferring subliminal primes from EEG through machine learning," *Proc. 35th Annual Int'l* Conf. *IEEE Engineering in Medicine and Biology Society* (EMBC 2013), Jul. 2013.

X. Liu, B. Xu, Y. Zhang, Q. Yan, L. Pang, Q. Li, H. Sun, and B. Wang, "Combination of Diverse Ranking Models for Personalized Expedia Hotel Searches," *IEEE Int'l Conf. on Data Mining* (ICDM 2013), Workshop Paper, Dec. 2013.

A. Mahmood, and R.S. Sutton, "Representation Search through Generate and Test," *Workshop on Learning Representations* (AAAI 2013), Workshop Paper, Jul. 2013.

J. Modayil, "Two Perspectives on Learning Rich Representations from Robot Experience," *Workshop on Learning Rich Representations from Low-Level Sensors* (AAAI 2013), Workshop Paper, Jul. 2013.

P. M. Pilarski, L. Qi, M. Ferguson-Pell, and S. Grange, "Determining the Time until Muscle Fatigue using Temporally Extended Prediction Learning," *Proc. 18th Int'l Functional Electrical Stimulation Society Conf.* (IFESS 2013), Jun. 2013, pp. 37–40.

R. Shariff, and T. Dick, "Lunar Lander: A Continuous-Action Case Study for Policy Gradient Actor-Critic Algorithms," *Proc. 1st Multidisciplinary Conf. on Reinforcement Learning and Decision Making* (RLDM 2013), Oct. 2013.

H. van Seijen, S. Whiteson, and L. Kester, "Efficient Abstraction Selection in Reinforcement Learning (Extended Abstract)," *Proc. 10th Symposium on Abstraction, Reformulation, and Approximation* (SARA 2013), Jun. 2013.

D. Silver, R. S. Sutton, M. Mueller, "Temporal-difference search in computer go," *Proceedings of the ICAPS-13 Workshop on Planning and Learning*, 2013.

MAGAZINE ARTICLES

N. Bard, J. Hawkin, J. Rubin, and M. Zinkevich, "The Annual Computer Poker Competition," *AI Magazine*, Jun. 2013.

SPECIAL/INVITED PRESENTATIONS

Person	**Title**	**Venue**
R. Sutton	Knowledge and Sensorimotor Data	Invited talk at workshop at the International Conference on Robotics and Automation, Karlsruhe, Germany
R. Sutton	The Quest for the Ultimate TD(λ) Prediction-Learning Algorithm	Invited talk at the 2013 European Workshop on Reinforcement Learning, Dagstuhl, Germany
R. Sutton	There's Still Plenty to Do at the Bottom	Closing Talk at the First Multi-disciplinary Conference on Reinforcement Learning and Decision Making, Princeton, USA
R. Sutton	Planning in Reinforcement Learning	Opening talk at the 8th Bellairs Workshop on Reinforcement Learning, Barbados
R. Sutton	Mind and Data: Learning to Predict Long-term Consequences Efficiently	Dept. of Computer Science, University of Michigan, USA
R. Sutton	Reinforcement Learning and Psychology: A Personal Story	Dept. of Psychology, University of Michigan, USA
D. Schuurmans	Learning with Output Kernels and Latent Kernels	Nanjing University
D. Schuurmans	Convex Methods for Representation Learning	Nanjing University

D. Schuurmans	Convex Methods for Representation Learning	University of Science and Technology China
D. Schuurmans	Convex Two-layer Modeling	NIPS Workshop on Output Representation Learning
D. Schuurmans	Multi-label Classification with Output Kernels	European Conference on Machine Learning, Prague
Cs. Szepesvári	Towards robust reinforcement learning algorithms	Stanford University
Cs. Szepesvári	Multiple Kernel Learning with Many Kernels	University of Berkeley, California
Cs. Szepesvári	Optimistic Optimization Algorithms for Planning in MDPs	8th Barbados Workshop on Reinforcement Learning
Cs. Szepesvári	A Randomized Mirror Descent Algorithm for Large Scale Multiple Kernel Learning	Technion, Machine Learning Seminar
Cs. Szepesvári	Online Learning with Costly Features and Labels	ICML 2013 Workshop on Prediction
Cs. Szepesvári	Online-to-confidence-set conversions and application to sparse stochastic bandits	Microsoft Research, Redmond
Cs. Szepesvári	Online Markov Decision Processes: State of the art and some new results	European Workshop on Reinforcement Learning
Cs. Szepesvári	Rate-Optimal Strategies for Partial Monitoring Games	Technion
Cs. Szepesvári	Online Learning in Markov Decision Processes under Bandit Feedback	Tel-Aviv University
Cs. Szepesvári	Adaptive Monte-Carlo via Bandit Allocation	University of California
M. Bowling	Artificial Intelligence Going “All-In”	University of Alberta (IEEE Tech Night)
M. Bowling	The real world is much more like a poker game..	The President's Society (University of Alberta event in Calgary)
M. Bowling	Games and Intelligence of the Artificial Kind	Telus World of Science
J. Modayil	Planning with learned models on robots	8th Barbados Workshop on Reinforcement Learning
A. Gyorgy	Online Learning in Markov Decision Processes with Changing Reward Sequences	ETH Zurich, Switzerland
A. Gyorgy	Universal Prediction in Changing Environments	41st Annual Meeting of the Statistical Society of Canada
A. Gyorgy	Optimistic optimization algorithms for planning in MDPs II	8th Barbados Workshop on Reinforcement Learning
A. Gyorgy	Universal Prediction in Changing Environments	University of Alberta, Dept. of Mathematics and Statistics, Statistics Seminar

P. M. Pilarski	Real-time Control with Temporally Extended Predictions: A Sensorimotor Approach to Planning?	8th Barbados Workshop on Reinforcement Learning
P. M. Pilarski	Enhancing Human Abilities and Health with Machine Learning	Extreme Science: The Next Frontier (Exploring New and Extreme Science), The Centre for Global Education
P. M. Pilarski	Learning and Using Contextual Information in the Control of Assistive Devices	2013 International Conference on Rehabilitation Robotics, 1st Workshop on Present and Future of Non-invasive Peripheral-Nervous-System-Machine Interfaces
H. van Seijen	True Online TD(λ)	University of Maastricht
H. van Seijen	True Online TD(λ)	Delft University
H. van Seijen	True Online TD(λ)	University of Amsterdam
H. van Seijen	True Online TD(λ)	DeepMind
K. Abou-Moustafa	Data Dependent Distance Metrics	Qatar Computing Research Institute, Doha, Qatar

AWARDS

R.S. Sutton: University of Massachusetts Amherst Computer Science Outstanding Achievement and Advocacy Award

THESES

Ryan Kiros, MSc, "Learning Deep Representations, Embeddings and Codes from the Pixel Level of Natural and Medical Images," November 2013.

Mostafa Vafadost, MSc, "Temporal Abstraction in Monte Carlo Tree Search," November 2013.

Hao Cheng, MSc, "Bregman Divergence Clustering: A Convex Approach," November 2013.

Yaoliang Yu, PhD, "Fast Gradient Algorithms for Structured Sparsity," November 2013.

Gergely Neu, PhD, "Online Learning in Non-Stationary Markov Decision Processes," July 2013.

Marc Gendron-Bellemare. PhD, "Fast, Scalable Algorithms for Reinforcement Learning in High Dimensional Domains," August 2013

10. OUTREACH

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

Katherine Chen led three one-hour long sessions on robotics. Each session had 11-17 male and female students aged 11-14 years. The Young Scientist Conference is a program put on by Elk Island Public Schools.

Katherine Chen led an Arduino activity for 6th grade girls for the WISEST SET conference and for 9th grade girls for the Women in Technology (WIT) and Canadian Information Processing Society (CIPS) conference.

Chris Rayner assistant supervisor for two sessions where grade six girls learned to work with and do light programming with Arduinos (to control servos attached to cat ears).

Farzaneh Mirzazadeh WISEST's SET (Science, Engineering and Technology) Conference is a one day event for all high school girls in grades 10-12. This full-day event gives over 200 students the opportunity to engage with U of A students, staff and faculty and to learn what an education and career in SET could offer them through hands-on activities and face-to-face conversation with female role models who work in the field. I was a volunteer activity assistant on a session held about social network analysis.

Farzaneh Mirzazadeh volunteered as a high school students' guide for the Iverson exam and then toured and lead groups of students to watch demos in different labs.

Farzaneh Mirzazadeh volunteered as an activity assistant in one three sessions that contained around 15 grade 9 students to make female students interested in computing.

Joseph Modayil mentored the High School Internship student, Nicholas Westbury.

Joseph Modayil demonstration of lab research, for visitors from the provincial government (IAE)

Joseph Modayil demonstration of robot learning for young girls (perhaps grades 3-7), as part of an girls in engineering (GEM/DiscoverE) camp. (The visit was co-ordinated by Patrick Pilarski)

Patrick Pilarski presented a demonstration on robotics and artificial intelligence for ESL. Exchange students from Nigata University, Japan.

Patrick Pilarski presented a demonstration and science outreach for grades 5-8. Hands-on robotics and artificial intelligence demonstration and lecture for 20 students in the Dept. Computing Science Summer Camps (Riveting Robotics).

Patrick Pilarski presented a hands-on robotics and artificial intelligence demonstration and lecture for 6 students participating in the WISEST program at UoA (MecEng).

Alexandra Kearney presented at the Glenrose Rehabilitation Center and Spinal Function Lab, “Prediction for Robotic Prosthetic Control” for students currently participating in rehabilitation research.