VUB SCIENCES & BIOENGINEERING SCIENCES

The Research Group

Artificial Intelligence Lab

has the honor to invite you to the public defense of the PhD thesis of

Anna HARUTYUNYAN

to obtain the degree of Doctor of Sciences

Title of the PhD thesis:

Beyond Single-Step Temporal Difference Learning

Promotor:

Prof. Ann Nowé

The defense will take place on

Friday March 30 2018 at 17:00 h

in Auditorium D.2.01 at the campus Humanities, Sciences and Engineering of the Vrije Universiteit Brussel, Pleinlaan 2 - 1050 Elsene, and will be followed by a reception.

Members of the jury:

Prof. Wolfgang De Meuter (chairman)
Prof. Bernard Manderick (secretary)
Prof. Peter Vrancx (co-promotor)
Prof. Ann Dooms
Prof. Emma Brunskill (Stanford Univ., USA)
Dr. Bruno Scherrer (INRIA, Nancy, Fr.)

Curriculum vitae

Anna received her bachelor's degree in computer science and mathematics from Utah State University in 2010, and her master's degree in theoretical computer science from Oregon State University in 2012. In 2013, seeking a field that offers an empirical challenge along with a theoretical one, she was fortunate to receive support from IWT for a PhD at the AI lab of VUB. Since then, her work has spanned several areas of reinforcement learning, as well as practical contributions to the MIRAD exoskeleton project. She has published at several top-tier international peer-reviewed venues, and has received two best paper awards. Currently, Anna is a research scientist at DeepMind, contributing to their ambitious mission of solving intelligence.

Abstract of the PhD research

An intelligent agent performs an action. Exactly one time step later the agent experiences the consequence of its action. It observes the feedback from the environment, and finds itself in a new state. The agent, then, is able to revise its belief about the quality of his action based only on prior beliefs and this momentary feedback. This idea, of updating estimates based on other estimates, roughly explains *temporal difference learning*, perhaps the most fundamental concept of reinforcement learning. It is a powerful idea, and its uncanny analogues are observed in the activity of the dopamine-carrying neurons of the human brain. The naive computational form of this idea, however, is understandably limited. In particular: *a single, primitive time-step experience is not sufficiently rich to learn with sophistication.* In this dissertation, we consider this problem from several partially related and partially complementary directions that all aim to enrich the single step.

Richer updates. Allowing the agent take multiple actions before making the update makes it easier to judge their quality correctly. Unfortunately, doing so is fundamentally difficult when the desired updates are *off-policy*, that is: concerning a course of actions different from the agent's behavior. We devise novel off-policy multi-step algorithms that rely on the idea of correcting off- policy actions in the value, rather than the conventional probability space, and analyze their convergence,

Richer actions. Consider all of the micro-actions involved in doing something as simple as walking. Yet, we are able to collapse them in a single complex *macro*-action. As the agent gathers experience, it must likewise be able to ascend the levels of acting hierarchy. The *options framework* is a general model for such temporal abstraction in reinforcement learning. We first make a novel link between the options framework, and multi-step temporal difference planning. Using this link, we devise an algorithm that is able to learn about options that *terminate* off-policy, that is: irrespectively of the behavior options. We then consider a novel scheme for discounting when using options, and show that it offers benefits from the perspective of both optimization and representation.

Richer feedback. Learning is rarely completely insular: when learning a new task, we assimilate multiple sources of feedback. Likewise, in reinforcement learning, the environment feedback alone may be too infrequent to be efficiently exploitable. *Potential-based reward shaping* is a paradigm for augmenting it with additional feedback, notable for its attractive theoretical guarantees. However, it requires the additional feedback to be presented through a specific abstraction, which may be cumbersome to obtain. We devise a framework that allows to provide the additional feedback in the natural form directly, while maintaining the desired theoretical guarantees.