IN PARTNERSHIP WITH:
Université de Lille

Activity Report 2019

Project-Team SEQUEL

Sequential Learning

IN COLLABORATION WITH: Centre de Recherche en Informatique, Signal et Automatique de Lille

RESEARCH CENTER
Lille - Nord Europe

THEME
Optimization, machine learning and statistical methods

Table of contents

Project-Team SEQUEL

Creation of the Project-Team: 2007 July 01

Keywords:

Computer Science and Digital Science:

A3. - Data and knowledge
A3.1. - Data
A3.1.1. - Modeling, representation
A3.1.4. - Uncertain data
A3.3. - Data and knowledge analysis
A3.3.1. - On-line analytical processing
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4. - Machine learning and statistics
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.3. - Reinforcement learning
A3.4.4. - Optimization and learning
A3.4.5. - Bayesian methods
A3.4.6. - Neural networks
A3.4.8. - Deep learning
A3.5.2. - Recommendation systems
A5.1. - Human-Computer Interaction
A5.10.7. - Learning
A9. - Artificial intelligence
A9.2. - Machine learning
A9.3. - Signal analysis
A9.4. - Natural language processing
A9.7. - AI algorithmics

Other Research Topics and Application Domains:

B2. - Health
B3.1. - Sustainable development
B3.5. - Agronomy
B4.4. - Energy delivery
B4.4.1. - Smart grids
B5.8. - Learning and training
B7.2.1. - Smart vehicles
B9.1.1. - E-learning, MOOC
B9.5. - Sciences
B9.5.6. - Data science

1. Team, Visitors, External Collaborators

Research Scientists

Émilie Kaufmann [CNRS, Researcher]
Odalric-Ambrym Maillard [Inria, Researcher, HDR]
Michal Valko [Inria, Researcher, until Mar 2019, HDR]
Jill-Jênn Vie [Inria, Researcher, from Oct 2019]

Faculty Member

Philippe Preux [Team leader, Université de Lille, Professor, HDR]

Post-Doctoral Fellows

Pierre Ménard [Inria, from Feb 2019]
Mohammad Sadegh Talebi Mazraeh Shahi [Inria]

PhD Students

Dorian Baudry [CNRS, from Nov 2019]
Lilian Besson [Ecole Normale Supérieure de Cachan, until Sep 2019]
Nicolas Carrara [Université de Lill, until Aug 2019]
Omar Darwiche Domingues [Inria]
Johan Ferret [Google, from Sep 2019]
Yannis Flet Berliac [Université de Lille]
Ronan Fruit [Inria, until Nov 2019]
Romain Gautron [Centre de coopération internationale en recherche agronomique, from Sep 2019]
Jean-Bastien Grill [DeepMind]
Nathan Grinsztajn [Ecole polytechnique, from Oct 2019]
Édouard Leurent [Renault]
Reda Ouhamma [Ecole polytechnique, from Sep 2019]
Pierre Perrault [Inria]
Hassan Saber [Inria]
Mathieu Seurin [Université de Lille]
Julien Seznec [Le livre scolaire]
Xuedong Shang [Université de Lille]
Florian Strub [Université de Lille, until May 2019]
Jean Tarbouriech [Facebook, from Apr 2019]
Kiewan Villatel [Criteo, until Jun 2019]

Technical staff

Guillaume Gautier [CNRS, Engineer]
Franck Valentini [Inria, Engineer, from May 2019]

Interns and Apprentices

Raphael Avalos Martinez de Escobar [Inria, from Apr 2019 until Oct 2019]
Hippolyte Bourel [Inria, from Feb 2019 until Jun 2019]
Geoffrey Cideron [Université de Lille, from Apr 2019 until Sep 2019]
Alessio Della Libera [Inria, from Jul 2019 until Sep 2019]
Come Fiegel [Ecole Normale Supérieure de Paris, from Jun 2019 until Aug 2019]
Reda Ouhamma [Inria, from Apr 2019 until Jul 2019]
Nicolas Yax [Inria, from Jun 2019 until Jul 2019]

Administrative Assistant

Amelie Supervielle

Visiting Scientists

Rianne de Heide [CWI, The Netherlands, from Apr 2019 until Aug 2019]
Per Anders Jonsson [Pompeu Fabra University, Spain, from Aug 2019]
Chuan Zheng Lee [Stanford University, USA, from Jun 2019 until Oct 2019]

Arun Verma [IIT Bombay, India, from Jun 2019 until Nov 2019]
Kaige Yang [University College London, UK, from Oct 2019]

External Collaborators
Remi Bardenet [CNRS]
Olivier Pietquin [Google Brain, HDR]

2. Overall Objectives

2.1. Presentation

SEQUEL means "Sequential Learning". As such, SEQUEL focuses on the task of learning in artificial systems (either hardware, or software) that gather information along time. Such systems are named *(learning) agents* (or learning machines) in the following. These data may be used to estimate some parameters of a model, which in turn, may be used for selecting actions in order to perform some long-term optimization task.

For the purpose of model building, the agent needs to represent information collected so far in some compact form and use it to process newly available data.

The acquired data may result from an observation process of an agent in interaction with its environment (the data thus represent a perception). This is the case when the agent makes decisions (in order to attain a certain objective) that impact the environment, and thus the observation process itself.

Hence, in SEQUEL, the term **sequential** refers to two aspects:

- The **sequential acquisition of data**, from which a model is learned (supervised and non supervised learning),
- the **sequential decision making task**, based on the learned model (reinforcement learning).

Examples of sequential learning problems include:

Supervised learning tasks deal with the prediction of some response given a certain set of observations of input variables and responses. New sample points keep on being observed.

Unsupervised learning tasks deal with clustering objects, these latter making a flow of objects. The (unknown) number of clusters typically evolves during time, as new objects are observed.

Reinforcement learning tasks deal with the control (a policy) of some system which has to be optimized (see [63]). We do not assume the availability of a model of the system to be controlled.

In all these cases, we mostly assume that the process can be considered stationary for at least a certain amount of time, and slowly evolving.

We wish to have any-time algorithms, that is, at any moment, a prediction may be required/an action may be selected making full use, and hopefully, the best use, of the experience already gathered by the learning agent.

The perception of the environment by the learning agent (using its sensors) is generally not the best one either to make a prediction, or to take a decision (we deal with Partially Observable Markov Decision Problem). So, the perception has to be mapped in some way to a better, and relevant, state (or input) space.

Finally, an important issue of prediction regards its evaluation: how wrong may we be when we perform a prediction? For real systems to be controlled, this issue can not be simply left unanswered.

To sum-up, in SEQUEL, the main issues regard:

- the learning of a model: we focus on models that map some input space $\mathbb{R}^P$ to $\mathbb{R}$,
- the observation to state mapping,
- the choice of the action to perform (in the case of sequential decision problem),
- the performance guarantees,
- the implementation of usable algorithms,

all that being understood in a *sequential* framework.

3. Research Program

3.1. In Short

SEQUEL is primarily grounded on two domains:

- the problem of decision under uncertainty,
- statistical analysis and statistical learning, which provide the general concepts and tools to solve this problem.

To help the reader who is unfamiliar with these questions, we briefly present key ideas below.

3.2. Decision-making Under Uncertainty

The phrase "Decision under uncertainty" refers to the problem of taking decisions when we do not have a full knowledge neither of the situation, nor of the consequences of the decisions, as well as when the consequences of decision are non deterministic.

We introduce two specific sub-domains, namely the Markov decision processes which model sequential decision problems, and bandit problems.

3.2.1. Reinforcement Learning

Sequential decision processes occupy the heart of the SEQUEL project; a detailed presentation of this problem may be found in Puterman's book [61].

A Markov Decision Process (MDP) is defined as the tuple $(\mathcal{X}, \mathcal{A}, P, r)$ where $\mathcal{X}$ is the state space, $\mathcal{A}$ is the action space, P is the probabilistic transition kernel, and $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to \mathbb{R}$ is the reward function. For the sake of simplicity, we assume in this introduction that the state and action spaces are finite. If the current state (at time t) is $x \in \mathcal{X}$ and the chosen action is $a \in \mathcal{A}$, then the Markov assumption means that the transition probability to a new state $x' \in \mathcal{X}$ (at time $t+1$) only depends on (x, a). We write $p(x'|x, a)$ the corresponding transition probability. During a transition $(x, a) \to x'$, a reward $r(x, a, x')$ is incurred.

In the MDP $(\mathcal{X}, \mathcal{A}, P, r)$, each initial state x_0 and action sequence $a_0, a_1, ...$ gives rise to a sequence of states $x_1, x_2, ...$, satisfying $\mathbb{P}(x_{t+1} = x'|x_t = x, a_t = a) = p(x'|x, a)$, and rewards [1] $r_1, r_2, ...$ defined by $r_t = r(x_t, a_t, x_{t+1})$.

The history of the process up to time t is defined to be $H_t = (x_0, a_0, ..., x_{t-1}, a_{t-1}, x_t)$. A policy π is a sequence of functions $\pi_0, \pi_1, ...$, where π_t maps the space of possible histories at time t to the space of probability distributions over the space of actions $\mathcal{A}$. To follow a policy means that, in each time step, we assume that the process history up to time t is $x_0, a_0, ..., x_t$ and the probability of selecting an action a is equal to $\pi_t(x_0, a_0, ..., x_t)(a)$. A policy is called stationary (or Markovian) if π_t depends only on the last visited state. In other words, a policy $\pi = (\pi_0, \pi_1, ...)$ is called stationary if $\pi_t(x_0, a_0, ..., x_t) = \pi_0(x_t)$ holds for all $t \geq 0$. A policy is called deterministic if the probability distribution prescribed by the policy for any history is concentrated on a single action. Otherwise it is called a stochastic policy.

[1] Note that for simplicity, we considered the case of a deterministic reward function, but in many applications, the reward r_t itself is a random variable.

We move from an MD process to an MD problem by formulating the goal of the agent, that is what the sought policy π has to optimize. It is very often formulated as maximizing (or minimizing), in expectation, some functional of the sequence of future rewards. For example, an usual functional is the infinite-time horizon sum of discounted rewards. For a given (stationary) policy π, we define the value function $V^\pi(x)$ of that policy π at a state $x \in \mathcal{X}$ as the expected sum of discounted future rewards given that we state from the initial state x and follow the policy π:

$$V^\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t | x_0 = x, \pi\right], \tag{1}$$

where $\mathbb{E}$ is the expectation operator and $\gamma \in (0,1)$ is the discount factor. This value function V^π gives an evaluation of the performance of a given policy π. Other functionals of the sequence of future rewards may be considered, such as the undiscounted reward (see the stochastic shortest path problems [60]) and average reward settings. Note also that, here, we consider the problem of maximizing a reward functional, but a formulation in terms of minimizing some cost or risk functional would be equivalent.

In order to maximize a given functional in a sequential framework, one usually applies Dynamic Programming (DP) [58], which introduces the optimal value function $V^*(x)$, defined as the optimal expected sum of rewards when the agent starts from a state x. We have $V^*(x) = \sup_\pi V^\pi(x)$. Now, let us give two definitions about policies:

- We say that a policy π is optimal, if it attains the optimal values $V^*(x)$ for any state $x \in \mathcal{X}$, *i.e.*, if $V^\pi(x) = V^*(x)$ for all $x \in \mathcal{X}$. Under mild conditions, deterministic stationary optimal policies exist [59]. Such an optimal policy is written π^*.
- We say that a (deterministic stationary) policy π is greedy with respect to (w.r.t.) some function V (defined on $\mathcal{X}$) if, for all $x \in \mathcal{X}$,

$$\pi(x) \in \arg\max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V(x')\right].$$

  where $\arg\max_{a \in \mathcal{A}} f(a)$ is the set of $a \in \mathcal{A}$ that maximizes $f(a)$. For any function V, such a greedy policy always exists because $\mathcal{A}$ is finite.

The goal of Reinforcement Learning (RL), as well as that of dynamic programming, is to design an optimal policy (or a good approximation of it).

The well-known Dynamic Programming equation (also called the Bellman equation) provides a relation between the optimal value function at a state x and the optimal value function at the successors states x' when choosing an optimal action: for all $x \in \mathcal{X}$,

$$V^*(x) = \max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V^*(x')\right]. \tag{2}$$

The benefit of introducing this concept of optimal value function relies on the property that, from the optimal value function V^*, it is easy to derive an optimal behavior by choosing the actions according to a policy greedy w.r.t. V^*. Indeed, we have the property that a policy greedy w.r.t. the optimal value function is an optimal policy:

$$\pi^*(x) \in \arg\max_{a \in \mathcal{A}} \sum_{x' \in \mathcal{X}} p(x'|x,a)\left[r(x,a,x') + \gamma V^*(x')\right]. \tag{3}$$

In short, we would like to mention that most of the reinforcement learning methods developed so far are built on one (or both) of the two following approaches ([64]):

- Bellman's dynamic programming approach, based on the introduction of the value function. It consists in learning a "good" approximation of the optimal value function, and then using it to derive a greedy policy w.r.t. this approximation. The hope (well justified in several cases) is that the performance V^π of the policy π greedy w.r.t. an approximation V of V^* will be close to optimality. This approximation issue of the optimal value function is one of the major challenges inherent to the reinforcement learning problem. **Approximate dynamic programming** addresses the problem of estimating performance bounds (*e.g.* the loss in performance $||V^* - V^\pi||$ resulting from using a policy π-greedy w.r.t. some approximation V- instead of an optimal policy) in terms of the approximation error $||V^* - V||$ of the optimal value function V^* by V. Approximation theory and Statistical Learning theory provide us with bounds in terms of the number of sample data used to represent the functions, and the capacity and approximation power of the considered function spaces.
- Pontryagin's maximum principle approach, based on sensitivity analysis of the performance measure w.r.t. some control parameters. This approach, also called **direct policy search** in the Reinforcement Learning community aims at directly finding a good feedback control law in a parameterized policy space without trying to approximate the value function. The method consists in estimating the so-called **policy gradient**, *i.e.* the sensitivity of the performance measure (the value function) w.r.t. some parameters of the current policy. The idea being that an optimal control problem is replaced by a parametric optimization problem in the space of parameterized policies. As such, deriving a policy gradient estimate would lead to performing a stochastic gradient method in order to search for a local optimal parametric policy.

Finally, many extensions of the Markov decision processes exist, among which the Partially Observable MDPs (POMDPs) is the case where the current state does not contain all the necessary information required to decide for sure of the best action.

3.2.2. Multi-arm Bandit Theory

Bandit problems illustrate the fundamental difficulty of decision making in the face of uncertainty: A decision maker must choose between what seems to be the best choice ("exploit"), or to test ("explore") some alternative, hoping to discover a choice that beats the current best choice.

The classical example of a bandit problem is deciding what treatment to give each patient in a clinical trial when the effectiveness of the treatments are initially unknown and the patients arrive sequentially. These bandit problems became popular with the seminal paper [62], after which they have found applications in diverse fields, such as control, economics, statistics, or learning theory.

Formally, a K-armed bandit problem ($K \geq 2$) is specified by K real-valued distributions. In each time step a decision maker can select one of the distributions to obtain a sample from it. The samples obtained are considered as rewards. The distributions are initially unknown to the decision maker, whose goal is to maximize the sum of the rewards received, or equivalently, to minimize the regret which is defined as the loss compared to the total payoff that can be achieved given full knowledge of the problem, *i.e.*, when the arm giving the highest expected reward is pulled all the time.

The name "bandit" comes from imagining a gambler playing with K slot machines. The gambler can pull the arm of any of the machines, which produces a random payoff as a result: When arm k is pulled, the random payoff is drawn from the distribution associated to k. Since the payoff distributions are initially unknown, the gambler must use exploratory actions to learn the utility of the individual arms. However, exploration has to be carefully controlled since excessive exploration may lead to unnecessary losses. Hence, to play well, the gambler must carefully balance exploration and exploitation. Auer *et al.* [57] introduced the algorithm UCB (Upper Confidence Bounds) that follows what is now called the "optimism in the face of uncertainty principle". Their algorithm works by computing upper confidence bounds for all the arms and then choosing the arm with the highest such bound. They proved that the expected regret of their algorithm increases at most

at a logarithmic rate with the number of trials, and that the algorithm achieves the smallest possible regret up to some sub-logarithmic factor (for the considered family of distributions).

3.3. Statistical analysis of time series

Many of the problems of machine learning can be seen as extensions of classical problems of mathematical statistics to their (extremely) non-parametric and model-free cases. Other machine learning problems are founded on such statistical problems. Statistical problems of sequential learning are mainly those that are concerned with the analysis of time series. These problems are as follows.

3.3.1. Prediction of Sequences of Structured and Unstructured Data

Given a series of observations $x_1, ..., x_n$ it is required to give forecasts concerning the distribution of the future observations $x_{n+1}, x_{n+2}, \cdots$; in the simplest case, that of the next outcome x_{n+1}. Then x_{n+1} is revealed and the process continues. Different goals can be formulated in this setting. One can either make some assumptions on the probability measure that generates the sequence $x_1, \cdots, x_n, \cdots$, such as that the outcomes are independent and identically distributed (i.i.d.), or that the sequence is a Markov chain, that it is a stationary process, etc. More generally, one can assume that the data is generated by a probability measure that belongs to a certain set $\mathcal{C}$. In these cases the goal is to have the discrepancy between the predicted and the "true" probabilities to go to zero, if possible, with guarantees on the speed of convergence.

Alternatively, rather than making some assumptions on the data, one can change the goal: the predicted probabilities should be asymptotically as good as those given by the best reference predictor from a certain pre-defined set.

Another dimension of complexity in this problem concerns the nature of observations x_i. In the simplest case, they come from a finite space, but already basic applications often require real-valued observations. Moreover, function or even graph-valued observations often arise in practice, in particular in applications concerning Web data. In these settings estimating even simple characteristics of probability distributions of the future outcomes becomes non-trivial, and new learning algorithms for solving these problems are in order.

3.3.2. Hypothesis testing

Given a series of observations of $x_1, \cdots, x_n, \cdots$ generated by some unknown probability measure μ, the problem is to test a certain given hypothesis H_0 about μ, versus a given alternative hypothesis H_1. There are many different examples of this problem. Perhaps the simplest one is testing a simple hypothesis "μ is Bernoulli i.i.d. measure with probability of 0 equals 1/2" versus "μ is Bernoulli i.i.d. with the parameter different from 1/2". More interesting cases include the problems of model verification: for example, testing that μ is a Markov chain, versus that it is a stationary ergodic process but not a Markov chain. In the case when we have not one but several series of observations, we may wish to test the hypothesis that they are independent, or that they are generated by the same distribution. Applications of these problems to a more general class of machine learning tasks include the problem of feature selection, the problem of testing that a certain behavior (such as pulling a certain arm of a bandit, or using a certain policy) is better (in terms of achieving some goal, or collecting some rewards) than another behavior, or than a class of other behaviors.

The problem of hypothesis testing can also be studied in its general formulations: given two (abstract) hypothesis H_0 and H_1 about the unknown measure that generates the data, find out whether it is possible to test H_0 against H_1 (with confidence), and if so, how can one do it.

3.3.3. Change Point Analysis

A stochastic process is generating the data. At some point, the process distribution changes. In the "offline" situation, the statistician observes the resulting sequence of outcomes and has to estimate the point or the points at which the change(s) occurred. In online setting, the goal is to detect the change as quickly as possible.

These are the classical problems in mathematical statistics, and probably among the last remaining statistical problems not adequately addressed by machine learning methods. The reason for the latter is perhaps in that the problem is rather challenging. Thus, most methods available so far are parametric methods concerning piece-wise constant distributions, and the change in distribution is associated with the change in the mean. However, many applications, including DNA analysis, the analysis of (user) behavior data, etc., fail to comply with this kind of assumptions. Thus, our goal here is to provide completely non-parametric methods allowing for any kind of changes in the time-series distribution.

3.3.4. Clustering Time Series, Online and Offline

The problem of clustering, while being a classical problem of mathematical statistics, belongs to the realm of unsupervised learning. For time series, this problem can be formulated as follows: given several samples $x^1 = (x_1^1, \cdots, x_{n_1}^1), \cdots, x^N = (x_N^1, \cdots, x_{n_N}^N)$, we wish to group similar objects together. While this is of course not a precise formulation, it can be made precise if we assume that the samples were generated by k different distributions.

The online version of the problem allows for the number of observed time series to grow with time, in general, in an arbitrary manner.

3.3.5. Online Semi-Supervised Learning

Semi-supervised learning (SSL) is a field of machine learning that studies learning from both labeled and unlabeled examples. This learning paradigm is extremely useful for solving real-world problems, where data is often abundant but the resources to label them are limited.

Furthermore, *online* SSL is suitable for adaptive machine learning systems. In the classification case, learning is viewed as a repeated game against a potentially adversarial nature. At each step t of this game, we observe an example $\mathbf{x_t}$, and then predict its label $\widehat{y}_t$.

The challenge of the game is that we only exceptionally observe the true label y_t. In the extreme case, which we also study, only a handful of labeled examples are provided in advance and set the initial bias of the system while unlabeled examples are gathered online and update the bias continuously. Thus, if we want to adapt to changes in the environment, we have to rely on indirect forms of feedback, such as the structure of data.

3.3.6. Online Kernel and Graph-Based Methods

Large-scale kernel ridge regression is limited by the need to store a large kernel matrix. Similarly, large-scale graph-based learning is limited by storing the graph Laplacian. Furthermore, if the data come online, at some point no finite storage is sufficient and per step operations become slow.

Our challenge is to design sparsification methods that give guaranteed approximate solutions with a reduced storage requirements.

4. Application Domains

4.1. Sequential decision making under uncertainty and prediction

The spectrum of applications of our research is very wide: it ranges from the core of our research, that is sequential decision making under uncertainty, to the application of components used to solve this decision making problem.

To be more specific, we work on computational advertising and recommendation systems; these problems are considered as a sequential matching problem in which resources available in a limited amount have to be matched to meet some users' expectations. The sequential approach we advocate paves the way to better tackle the cold-start problem, and non stationary environments. More generally, these approaches are applied to the optimization of budgeted resources under uncertainty, in a time-varying environment, including constraints on computational times (typically, a decision has to be made in less than 1 ms in a recommendation system). An other field of application of our research is related to education which we consider as a sequential matching problem between a student, and educational contents.

The algorithms to solve these tasks heavily rely on tools from machine learning, statistics, and optimization. Henceforth, we also apply our work to more classical supervised learning, and prediction tasks, as well as unsupervised learning tasks. The whole range of methods is used, from decision forests, to kernel methods, to deep learning. For instance, we have recently used deep learning on images. We also have a line of work related to software development studying how machine learning can improve the quality of software being developed. More generally, we apply our research to data science.

5. Highlights of the Year

5.1. Highlights of the Year

- Organization of the 1st Reinforcement Learning Summer Scool: 2 weeks of lectures, keynotes, and practical sessions fully dedicated to bandits and reinforcement learning. We received about 300 applications from all around the world and selected 110 participants.
- Julien Seznec and Michal Valko have obtained an oral at AI&Stats (2,5% acceptance rate) [32].
- This is the ultimate SEQUEL highlight: after 12 years, following Inria's policy, SEQUEL comes to an end. We have designed a new team-project which will be named SCOOL.

5.1.1. Awards

BEST PAPER AWARD:

[16]
M. ASADI, M. S. TALEBI, H. BOUREL, O.-A. MAILLARD. *Model-Based Reinforcement Learning Exploiting State-Action Equivalence*, in "ACML 2019, Proceedings of Machine Learning Research", Nagoya, Japan, 2019, vol. 101, pp. 204 - 219, https://hal.archives-ouvertes.fr/hal-02378887

6. New Software and Platforms

6.1. gym-backgammon

Backgammon environment

KEYWORD: Artificial intelligence

FUNCTIONAL DESCRIPTION: This software program follows the openai gym API (https://gym.openai.com/), that is the interaction loop of reinforcement learning: the game is in a certain state, the agent selects an action, this action is simulated in the game, and the next state of the game as well as the return are returned to the agent. All these notions follows backgammon rules and should be understood as pertaining to the reinforcement learning vocabulary. As far as we are aware of, gym-backgammon is the only existing software of this type, and it is available in open source. Great care has been put into the debugging and the efficiency. This software program is developped in python. The inetraction is made according to a client-server model.

- Author: Alessio Della Libera
- Contact: Philippe Preux
- URL: https://github.com/dellalibera/gym-backgammon

6.2. gym-rubik

Rubik's cube environment

KEYWORD: Artificial intelligence

FUNCTIONAL DESCRIPTION: This software program follows the openai gym API (https://gym.openai.com/), that is the interaction loop of reinforcement learning: the game is in a certain state, the agent selects an action, this action is simulated in the game, and the next state of the game as well as the return are returned to the agent. All these notions follows Rubik's cube rules and should be understood as pertaining to the reinforcement learning vocabulary. Great care has been put into the debugging and the efficiency. This software program is developped in python.

- Author: Raphaël Avalos Martinez De Escobar
- Contact: Philippe Preux
- URL: https://github.com/raphaelavalos/gym-rubikscube

6.3. highway-env

An environment for autonomous driving decision-making

KEYWORDS: Generic modeling environment - Simulation - Autonomous Cars - Artificial intelligence

FUNCTIONAL DESCRIPTION: The environment is composed of several variants, each of which corresponds to driving scenes: highway, roundabout, intersection, merge, parking, etc. The road network is described by a graph, and is then populated with simulated vehicles. Vehicle kinematics follows a simple Bicycle model, and their behavior is determined by models derived from road traffic simulation literature. The ego-vehicle has access to a description of the scene through several types of observations, and its behavior is controlled through an action space, either discrete (change of lanes, of cruising speed) or continuous (accelerator pedal, steering wheel angle). The objective function to maximize is also described by the environment and may vary depending on the task to be solved. The interface of the library is inherited from the standard defined by OpenAI Gym, consisting of four main methods: gym.make(id), env.step(action), env.reset(), and env.render().

- Author: Edouard Leurent
- Contact: Edouard Leurent

7. New Results

7.1. Decision-making Under Uncertainty

7.1.1. Reinforcement Learning

Model-Based Reinforcement Learning Exploiting State-Action Equivalence, [16]

Leveraging an equivalence property in the state-space of a Markov Decision Process (MDP) has been investigated in several studies. This paper studies equivalence structure in the reinforcement learning (RL) setup, where transition distributions are no longer assumed to be known. We present a notion of similarity between transition probabilities of various state-action pairs of an MDP, which naturally defines an equivalence structure in the state-action space. We present equivalence-aware confidence sets for the case where the learner knows the underlying structure in advance. These sets are provably smaller than their corresponding equivalence-oblivious counterparts. In the more challenging case of an unknown equivalence structure, we present an algorithm called ApproxEquivalence that seeks to find an (approximate) equivalence structure, and define confidence sets using the approximate equivalence. To illustrate the efficacy of the presented confidence sets, we present C-UCRL, as a natural modification of UCRL2 for RL in undiscounted MDPs. In the case of a known equivalence structure, we show that C-UCRL improves over UCRL2 in terms of regret by a factor of SA/C, in any communicating MDP with S states, A actions, and C classes, which corresponds to a massive improvement when C SA. To the best of our knowledge, this is the first work providing regret bounds for RL when an equivalence structure in the MDP is efficiently exploited. In the case of an unknown equivalence structure, we show through numerical experiments that C-UCRL combined with ApproxEquivalence outperforms UCRL2 in ergodic MDPs.

Practical Open-Loop Optimistic Planning, [25]

We consider the problem of online planning in a Markov Decision Process when given only access to a generative model, restricted to open-loop policies-i.e. sequences of actions-and under budget constraint. In this setting, the Open-Loop Optimistic Planning (OLOP) algorithm enjoys good theoretical guarantees but is overly conservative in practice, as we show in numerical experiments. We propose a modified version of the algorithm with tighter upper-confidence bounds, KL-OLOP, that leads to better practical performances while retaining the sample complexity bound. Finally, we propose an efficient implementation that significantly improves the time complexity of both algorithms.

Budgeted Reinforcement Learning in Continuous State Space, [20]

A Budgeted Markov Decision Process (BMDP) is an extension of a Markov Decision Process to critical applications requiring safety constraints. It relies on a notion of risk implemented in the shape of a cost signal constrained to lie below an-adjustable-threshold. So far, BMDPs could only be solved in the case of finite state spaces with known dynamics. This work extends the state-of-the-art to continuous spaces environments and unknown dynamics. We show that the solution to a BMDP is a fixed point of a novel Budgeted Bellman Optimality operator. This observation allows us to introduce natural extensions of Deep Reinforcement Learning algorithms to address large-scale BMDPs. We validate our approach on two simulated applications: spoken dialogue and autonomous driving.

Regret Bounds for Learning State Representations in Reinforcement Learning, [29]

We consider the problem of online reinforcement learning when several state representations (mapping histories to a discrete state space) are available to the learning agent. At least one of these representations is assumed to induce a Markov decision process (MDP), and the performance of the agent is measured in terms of cumulative regret against the optimal policy giving the highest average reward in this MDP representation. We propose an algorithm (UCB-MS) with $O(\sqrt{T})$ regret in any communicating MDP. The regret bound shows that UCB-MS automatically adapts to the Markov model and improves over the currently known best bound of order O(T 2/3).

Planning in entropy-regularized Markov decision processes and games, [24]

We propose SmoothCruiser, a new planning algorithm for estimating the value function in entropy-regularized Markov decision processes and two-player games, given a generative model of the environment. SmoothCruiser makes use of the smoothness of the Bellman operator promoted by the regularization to achieve problem-independent sample complexity of order $O(1/\epsilon\ 4)$ for a desired accuracy ϵ, whereas for non-regularized settings there are no known algorithms with guaranteed polynomial sample complexity in the worst case.

7.1.1.1. Deep reinforcement learning

"I'm sorry Dave, I'm afraid I can't do that" Deep Q-Learning From Forbidden Actions, [42]

The use of Reinforcement Learning (RL) is still restricted to simulation or to enhance human-operated systems through recommendations. Real-world environments (e.g. industrial robots or power grids) are generally designed with safety constraints in mind implemented in the shape of valid actions masks or contingency controllers. For example, the range of motion and the angles of the motors of a robot can be limited to physical boundaries. Violating constraints thus results in rejected actions or entering in a safe mode driven by an external controller, making RL agents incapable of learning from their mistakes. In this paper, we propose a simple modification of a state-of-the-art deep RL algorithm (DQN), enabling learning from forbidden actions. To do so, the standard Q-learning update is enhanced with an extra safety loss inspired by structured classification. We empirically show that it reduces the number of hit constraints during the learning phase and accelerates convergence to near-optimal policies compared to using standard DQN. Experiments are done on a Visual Grid World Environment and Text-World domain.

MERL: Multi-Head Reinforcement Learning, [39]

A common challenge in reinforcement learning is how to convert the agent's interactions with an environment into fast and robust learning. For instance, earlier work makes use of domain knowledge to improve existing reinforcement learning algorithms in complex tasks. While promising, previously acquired knowledge is often costly and challenging to scale up. Instead, we decide to consider problem knowledge with signals from quantities relevant to solve any task, e.g., self-performance assessment and accurate expectations. $\mathcal{V}^{ex}$ is such a quantity. It is the fraction of variance explained by the value function V and measures the discrepancy between V and the returns. Taking advantage of $\mathcal{V}^{ex}$, we propose MERL, a general framework for structuring reinforcement learning by injecting problem knowledge into policy gradient updates. As a result, the agent is not only optimized for a reward but learns using problem-focused quantities provided by MERL, applicable out-of-the-box to any task. In this paper: (a) We introduce and define MERL, the multi-head reinforcement learning framework we use throughout this work. (b) We conduct experiments across a variety of standard benchmark environments, including 9 continuous control tasks, where results show improved performance. (c) We demonstrate that MERL also improves transfer learning on a set of challenging pixel-based tasks. (d) We ponder how MERL tackles the problem of reward sparsity and better conditions the feature space of reinforcement learning agents.

Self-Educated Language Agent With Hindsight Experience Replay For Instruction Following, [45]

Language creates a compact representation of the world and allows the description of unlimited situations and objectives through compositionality. These properties make it a natural fit to guide the training of interactive agents as it could ease recurrent challenges in Reinforcement Learning such as sample complexity, generalization, or multi-tasking. Yet, it remains an open-problem to relate language and RL in even simple instruction following scenarios. Current methods rely on expert demonstrations, auxiliary losses, or inductive biases in neural architectures. In this paper, we propose an orthogonal approach called Textual Hindsight Experience Replay (THER) that extends the Hindsight Experience Replay approach to the language setting. Whenever the agent does not fulfill its instruction, THER learns to output a new directive that matches the agent trajectory, and it relabels the episode with a positive reward. To do so, THER learns to map a state into an instruction by using past successful trajectories, which removes the need to have external expert interventions to relabel episodes as in vanilla HER. We observe that this simple idea also initiates a learning synergy between language acquisition and policy learning on instruction following tasks in the BabyAI environment.

High-Dimensional Control Using Generalized Auxiliary Tasks, [47]

A long-standing challenge in reinforcement learning is the design of function approximations and efficient learning algorithms that provide agents with fast training, robust learning, and high performance in complex environments. To this end, the use of prior knowledge, while promising, is often costly and, in essence, challenging to scale up. In contrast, we consider problem knowledge signals, that are any relevant indicator useful to solve a task, e.g., metrics of uncertainty or proactive prediction of future states. Our framework consists of predicting such complementary quantities associated with self-performance assessment and accurate expectations. Therefore, policy and value functions are no longer only optimized for a reward but are learned using environment-agnostic quantities. We propose a generally applicable framework for structuring reinforcement learning by injecting problem knowledge in policy gradient updates. In this paper: (a) We introduce MERL, our multi-head reinforcement learning framework for generalized auxiliary tasks. (b) We conduct experiments across a variety of standard benchmark environments. Our results show that MERL improves performance for on- and off-policy methods. (c) We show that MERL also improves transfer learning on a set of challenging tasks. (d) We investigate how our approach addresses the problem of reward sparsity and pushes the function approximations into a better-constrained parameter configuration.

7.1.2. Multi-armed Bandit Theory

Asymptotically Optimal Algorithms for Budgeted Multiple Play Bandits, [15]

We study a generalization of the multi-armed bandit problem with multiple plays where there is a cost associated with pulling each arm and the agent has a budget at each time that dictates how much she can expect to spend. We derive an asymptotic regret lower bound for any uniformly efficient algorithm in our setting. We then study a variant of Thompson sampling for Bernoulli rewards and a variant of KL-UCB for

both single-parameter exponential families and bounded, finitely supported rewards. We show these algorithms are asymptotically optimal, both in rateand leading problem-dependent constants, including in the thick margin setting where multiple arms fall on the decision boundary.

Non-Asymptotic Pure Exploration by Solving Games, [46]

Pure exploration (aka active testing) is the fundamental task of sequentially gathering information to answer a query about a stochastic environment. Good algorithms make few mistakes and take few samples. Lower bounds (for multi-armed bandit models with arms in an exponential family) reveal that the sample complexity is determined by the solution to an optimisation problem. The existing state of the art algorithms achieve asymptotic optimality by solving a plug-in estimate of that optimisation problem at each step. We interpret the optimisation problem as an unknown game, and propose sampling rules based on iterative strategies to estimate and converge to its saddle point. We apply no-regret learners to obtain the first finite confidence guarantees that are adapted to the exponential family and which apply to any pure exploration query and bandit structure. Moreover, our algorithms only use a best response oracle instead of fully solving the optimisation problem.

Rotting bandits are not harder than stochastic ones, [32]

In bandits, arms' distributions are stationary. This is often violated in practice, where rewards change over time. In applications as recommendation systems, online advertising, and crowdsourcing, the changes may be triggered by the pulls, so that the arms' rewards change as a function of the number of pulls. In this paper, we consider the specific case of non-parametric rotting bandits, where the expected reward of an arm may decrease every time it is pulled. We introduce the filtering on expanding window average (FEWA) algorithm that at each round constructs moving averages of increasing windows to identify arms that are more likely to return high rewards when pulled once more. We prove that, without any knowledge on the decreasing behavior of the arms, FEWA achieves similar anytime problem-dependent, O(log(KT)), and problem-independent, O($sqrtKT$), regret bounds of near-optimal stochastic algorithms as UCB1 of Auer et al. (2002a). This result substantially improves the prior result of Levine et al. (2017) which needed knowledge of the horizon and decaying parameters to achieve problem-independent bound of only O(K1/3T2/3). Finally, we report simulations confirming the theoretical improvements of FEWA.

7.1.3. Black-box Optimization

General parallel optimization without a metric, [34]

Hierarchical bandits are an approach for global optimization of extremely irregular functions. This paper provides new elements regarding POO, an adaptive meta-algorithm that does not require the knowledge of local smoothness of the target function. We first highlight the fact that the subroutine algorithm used in POO should have a small regret under the assumption of local smoothness with respect to the chosen partitioning, which is unknown if it is satisfied by the standard subroutine HOO. In this work, we establish such regret guarantee for HCT, which is another hierarchical optimistic optimization algorithm that needs to know the smoothness. This confirms the validity of POO. We show that POO can be used with HCT as a subroutine with a regret upper bound that matches the one of best-known algorithms using the knowledge of smoothness up to a $\sqrt{\log n}$ factor. On top of that, we propose a general wrapper, called GPO, that can cope with algorithms that only have simple regret guarantees. Finally, we complement our findings with experiments on difficult functions.

A simple dynamic bandit algorithm for hyper-parameter tuning, [33]

Hyper-parameter tuning is a major part of modern machine learning systems. The tuning itself can be seen as a sequential resource allocation problem. As such, methods for multi-armed bandits have been already applied. In this paper, we view hyper-parameter optimization as an instance of best-arm identification in infinitely many-armed bandits. We propose D-TTTS, a new adaptive algorithm inspired by Thompson sampling, which dynamically balances between refining the estimate of the quality of hyper-parameter configurations previously explored and adding new hyper-parameter configurations to the pool of candidates. The algorithm is easy to implement and shows competitive performance compared to state-of-the-art algorithms for hyper-parameter tuning.

7.1.4. Statistics for Machine Learning

Non-asymptotic analysis of a sequential rupture detection test and its application to non-stationary bandits, [36]

We study a strategy for online change-point detection based on generalized likelihood ratios (GLR) and that can be expressed with the binary relative entropy. This test is used to detect a change in the mean of a bounded distribution, and we propose a non-asymptotic control of its false alarm probability and detection delay. We then explain how it can be useful for sequential decision making by proposing the GLR-klUCB bandit strategy, which is efficient in piece-wise stationary multi-armed bandit models.

Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds, [27]

We consider change-point detection in a fully sequential setup, when observations are received one by one and one must raise an alarm as early as possible after any change. We assume that both the change points and the distributions before and after the change are unknown. We consider the class of piecewise-constant mean processes with sub-Gaussian noise, and we target a detection strategy that is uniformly good on this class (this constrains the false alarm rate and detection delay). We introduce a novel tuning of the GLR test that takes here a simple form involving scan statistics, based on a novel sharp concentration inequality using an extension of the Laplace method for scan-statistics that holds doubly-uniformly in time. This also considerably simplifies the implementation of the test and analysis. We provide (perhaps surprisingly) the first fully non-asymptotic analysis of the detection delay of this test that matches the known existing asymptotic orders, with fully explicit numerical constants. Then, we extend this analysis to allow some changes that are not-detectable by any uniformly-good strategy (the number of observations before and after the change are too small for it to be detected by any such algorithm), and provide the first robust, finite-time analysis of the detection delay.

Learning Multiple Markov Chains via Adaptive Allocation, [35]

We study the problem of learning the transition matrices of a set of Markov chains from a single stream of observations on each chain. We assume that the Markov chains are ergodic but otherwise unknown. The learner can sample Markov chains sequentially to observe their states. The goal of the learner is to sequentially select various chains to learn transition matrices uniformly well with respect to some loss function. We introduce a notion of loss that naturally extends the squared loss for learning distributions to the case of Markov chains, and further characterize the notion of being uniformly good in all problem instances. We present a novel learning algorithm that efficiently balances exploration and exploitation intrinsic to this problem, without any prior knowledge of the chains. We provide finite-sample PAC-type guarantees on the performance of the algorithm. Further, we show that our algorithm asymptotically attains an optimal loss.

7.1.5. DPP

On two ways to use determinantal point processes for Monte Carlo integration, [40]

This paper focuses on Monte Carlo integration with determinantal point processes (DPPs) which enforce negative dependence between quadrature nodes. We survey the properties of two unbiased Monte Carlo estimators of the integral of interest: a direct one proposed by Bardenet & Hardy (2016) and a less obvious 60-year-old estimator by Ermakov & Zolotukhin (1960) that actually also relies on DPPs. We provide an efficient implementation to sample exactly a particular multidimen-sional DPP called multivariate Jacobi ensemble. This let us investigate the behavior of both estima-tors on toy problems in yet unexplored regimes.

7.2. Applications

7.2.1. Autonomous car

Practical Open-Loop Optimistic Planning, [25]

We consider the problem of online planning in a Markov Decision Process when given only access to a generative model, restricted to open-loop policies-i.e. sequences of actions-and under budget constraint. In this setting, the Open-Loop Optimistic Planning (OLOP) algorithm enjoys good theoretical guarantees but is overly conservative in practice, as we show in numerical experiments. We propose a modified version of the algorithm with tighter upper-confidence bounds, KL-OLOP, that leads to better practical performances while retaining the sample complexity bound. Finally, we propose an efficient implementation that significantly improves the time complexity of both algorithms.

Budgeted Reinforcement Learning in Continuous State Space, [20]

A Budgeted Markov Decision Process (BMDP) is an extension of a Markov Decision Process to critical applications requiring safety constraints. It relies on a notion of risk implemented in the shape of a cost signal constrained to lie below an-adjustable-threshold. So far, BMDPs could only be solved in the case of finite state spaces with known dynamics. This work extends the state-of-the-art to continuous spaces environments and unknown dynamics. We show that the solution to a BMDP is a fixed point of a novel Budgeted Bellman Optimality operator. This observation allows us to introduce natural extensions of Deep Reinforcement Learning algorithms to address large-scale BMDPs. We validate our approach on two simulated applications: spoken dialogue and autonomous driving.

7.2.2. Cognitive radio

Decentralized Spectrum Learning for IoT Wireless Networks Collision Mitigation, [28]

This paper describes the principles and implementation results of reinforcement learning algorithms on IoT devices for radio collision mitigation in ISM unlicensed bands. Learning is here used to improve both the IoT network capability to support a larger number of objects as well as the autonomy of IoT devices. We first illustrate the efficiency of the proposed approach in a proof-of-concept based on USRP software radio platforms operating on real radio signals. It shows how collisions with other RF signals present in the ISM band are diminished for a given IoT device. Then we describe the first implementation of learning algorithms on LoRa devices operating in a real LoRaWAN network, that we named IoTligent. The proposed solution adds neither processing overhead so that it can be ran in the IoT devices, nor network overhead so that no change is required to LoRaWAN. Real life experiments have been done in a realistic LoRa network and they show that IoTligent device battery life can be extended by a factor 2 in the scenarios we faced during our experiment.

GNU Radio Implementation of MALIN: "Multi-Armed bandits Learning for Internet-of-things Networks", [37]

We implement an IoT network in the following way: one gateway, one or several intelligent (i.e., learning) objects, embedding the proposed solution, and a traffic generator that emulates radio interferences from many other objects. Intelligent objects communicate with the gateway with a wireless ALOHA-based protocol, which does not require any specific overhead for the learning. We model the network access as a discrete sequential decision making problem, and using the framework and algorithms from Multi-Armed Bandit (MAB) learning, we show that intelligent objects can improve their access to the network by using low complexity and decentralized algorithms, such as UCB1 and Thompson Sampling. This solution could be added in a straightforward and costless manner in LoRaWAN networks, just by adding this feature in some or all the devices, without any modification on the network side.

7.2.3. Other

Accurate reconstruction of EBSD datasets by a multimodal data approach using an evolutionary algorithm, [14]

A new method has been developed for the correction of the distortions and/or enhanced phase differentiation in Electron Backscatter Diffraction (EBSD) data. Using a multi-modal data approach, the method uses segmented images of the phase of interest (laths, precipitates, voids, inclusions) on images gathered by backscattered or secondary electrons of the same area as the EBSD map. The proposed approach then search for the best transformation to correct their relative distortions and recombines the data in a new EBSD file. Speckles of the features of interest are first segmented in both the EBSD and image data modes. The speckle extracted

from the EBSD data is then meshed, and the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) is implemented to distort the mesh until the speckles superimpose. The quality of the matching is quantified via a score that is linked to the number of overlapping pixels in the speckles. The locations of the points of the distorted mesh are compared to those of the initial positions to create pairs of matching points that are used to calculate the polynomial function that describes the distortion the best. This function is then applied to un-distort the EBSD data, and the phase information is inferred using the data of the segmented speckle. Fast and versatile, this method does not require any human annotation and can be applied to large datasets and wide areas. Besides, this method requires very few assumptions concerning the shape of the distortion function. It can be used for the single compensation of the distortions or combined with the phase differentiation. The accuracy of this method is of the order of the pixel size. Some application examples in multiphase materials with feature sizes down to 1 μm are presented, including Ti-6Al-4V Titanium alloy, Rene 65 and additive manufactured Inconel 718 Nickel-base superalloys.

Energy Management for Microgrids: a Reinforcement Learning Approach, [41]

This paper presents a framework based on reinforcement learning for energy management and economic dispatch of an islanded microgrid without any forecasting module. The architecture of the algorithm is divided in two parts: a learning phase trained by a reinforcement learning (RL) algorithm on a small dataset and the testing phase based on a decision tree induced from the trained RL. An advantage of this approach is to create an autonomous agent, able to react in real-time, considering only the past. This framework was tested on real data acquired at Ecole Polytechnique in France over a long period of time, with a large diversity in the type of days considered. It showed near optimal, efficient and stable results in each situation.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Lelivrescolaire.fr

- Contract with http://Lelivrescolaire.fr; PI: Michal Valko

  Title: Sequential Machine Learning for Adaptive Educational Systems

  Duration: 3 years (Mar 2018 – Feb 2021)

  Abstract: This contract comes along the CIFRE grant on the same topic. Adaptive educational content are technologies which adapt to the difficulties encountered by students. With the rise of digital content in schools, the mass of data coming from education enables but also ask for machine learning methods. Since 2010, Lelivrescolaire.fr has been developing some learning materials for teachers and students through collaborative creation process. For instance, during the school year 2015/2016, students has achieved more than 8 000 000 exercises on its homework platform Afterclasse.fr. Our approach would be based on sequential machine learning: the algorithm learns to recommend some exercises which adapt to students gradually as they answer.
  Participants: Julien Seznec, Michal Valko.

8.1.2. Renault

- Contract with Renault; PI: Philippe Preux

  Title: Control of an autonomous vehicle

  Duration: 3 years (Dec 2017 – Nov 2020)

  Abstract: This contract comes along the CIFRE grant on the same topic. This work is done in collaboration with the NON-A team-project.
  Participants: Édouard Leurent, Odalric-Ambrym Maillard, Philippe Preux.

8.1.3. Critéo

- Contract with “Criteo”; PI: Philippe Preux

  Title: Computational advertizing

  Duration: 3 years (Dec 2017 – Jun 2019)

  Abstract: This contract comes along the CIFRE grant on the same topic. The goal is to investigate reinforcmeent learning and deep learning on the problem of ad selection on the Internet.

  Note: this contract came to its end because the PhD candidate quitted Critéo, hence aborting his PhD studies.
  Participants: Philippe Preux, Kiewan Villatel.

8.1.4. Share My Space

- Contract with “Share My Space”.

  Duration: 6 months
  Participant: Philippe Preux.

9. Partnerships and Cooperations

9.1. Regional Initiatives

9.1.1. With U. INSERM 1190, CHU Lille

Participants: Odalric-Ambrym Maillard, Philippe Preux, Philippe Preux.

Title: Bandits for Health (B4H)

Type: I-SITE Lille

Coordinator: Philippe Preux

Duration: 2019–2023

Abstract: B4H is a fundamental research project on a certain type of bandit algorithms, tailored to be applied to post-surgical patient follow-up. Bandit in a non-stationary environment will be studied. This work is performed in collaboration with Pr. F. Pattou and his group.

Title: No title

Type: Informal

Coordinator: Philippe Preux

Duration: 2019–2020

Abstract: This is mostly a data analysis work in order to study whether a certain disease may be predicted based on a certain dataset collected by U. INSERM 1190. Estelle Chatelain, a BiLille engineer, is involved in this project. This work is performed in collaboration with Pr. F. Pattou and his group.

9.1.2. With Service de Radiologie et Imagerie Musculosquelettique, CHU Lille

Participants: Philippe Preux, Franck Valentini.

Title: Radiology AI Demonstrator (RAID)

Type: CPER, Région Hauts-de-France

Coordinator: Philippe Preux

Duration: 2019–2020

Abstract: The goal of the RAID project is to assess the potential of deep learning for radio analysis and patient triage. Various applications are investigated.

9.2. National Initiatives

9.2.1. ANR BOLD

Participants: Émilie Kaufmann, Michal Valko, Pierre Ménard, Xuedong Shang, Omar Darwiche Domingues.

Title: Beyond Online Learning for better Decision making

Type: National Research Agency

Coordinator: Vianney Perchet (ENS Paris-Saclay / ENSAE)

Duration: 2019–2023

Abstract: Reactive machine learning algorithms adapt to data generating processes, typically do not require large computational power and, moreover, can be translated into offline (as opposed to online) algorithms if needed. Introduced in the 30s in the context of clinical trials, online ML algorithms have been gaining a lot of theoretical interest for the last 15 years because of their applications to the optimization of recommender systems, click through rates, planning in congested networks, to name just a few. However, in practice, such algorithms are not used as much as they should, because the traditional low-level modelling assumptions they are based upon are not appropriate, as it appears.

Instead of trying to complicate and generalise arbitrarily a framework unfit for potential applications, we will tackle this problem from another perspective. We will seek a better understanding of the simple original problem and extend it in the appropriate directions. There are currently three main barriers to a broader development of online learning, that this project aim at overcoming. 1) The classical "one step, one decision, one reward" paradigm is unfit. 2) Optimality is defined with respect to worst-case generic lower bounds and mechanics behind online learning are not fully understood. 3) Algorithms were designed in a non strategic or interactive environment.

The project gathers four parnters: ENS Paris-Saclay, University of Toulouse, Inria Lille and Université Paris Descartes.

9.2.2. ANR BoB

Participant: Michal Valko.

Title: Bayesian statistics for expensive models and tall data

Type: National Research Agency

Coordinator: CNRS (Rémi Bardenet)

Duration: 2016–2020

Abstract: Bayesian methods are a popular class of statistical algorithms for updating scientific beliefs. They turn data into decisions and models, taking into account uncertainty about models and their parameters. This makes Bayesian methods popular among applied scientists such as biologists, physicists, or engineers. However, at the heart of Bayesian analysis lie 1) repeated sweeps over the full dataset considered, and 2) repeated evaluations of the model that describes the observed physical process. The current trends to large-scale data collection and complex models thus raises two main issues. Experiments, observations, and numerical simulations in many areas of science nowadays generate terabytes of data, as does the LHC in particle physics for instance. Simultaneously, knowledge creation is becoming more and more data-driven, which requires new paradigms addressing how data are captured, processed, discovered, exchanged, distributed, and analyzed. For statistical algorithms to scale up, reaching a given performance must require as few iterations and as little access to data as possible. It is not only experimental measurements that are growing at a rapid pace. Cell biologists tend to have scarce data but large-scale models of tens of nonlinear differential equations to describe complex dynamics. In such settings, evaluating the model once requires numerically solving a large system of differential equations, which may take minutes for some tens of differential equations on today's hardware. Iterative statistical processing that requires a million sequential runs of the model is thus out of the question. In this project, we tackle the fundamental cost-accuracy

trade-off for Bayesian methods, in order to produce generic inference algorithms that scale favorably with the number of measurements in an experiment and the number of runs of a statistical model. We propose a collection of objectives with different risk-reward trade-offs to tackle these two goals. In particular, for experiments with large numbers of measurements, we further develop existing subsampling-based Monte Carlo methods, while developing a novel decision theory framework that includes data constraints. For expensive models, we build an ambitious programme around Monte Carlo methods that leverage determinantal processes, a rich class of probabilistic tools that lead to accurate inference with limited model evaluations. In short, using innovative techniques such as subsampling-based Monte Carlo and determinantal point processes, we propose in this project to push the boundaries of the applicability of Bayesian inference.

9.2.3. ANR Badass

Participants: Odalric-Ambrym Maillard, Émilie Kaufmann.

Title: BAnDits for non-Stationarity and Structure

Type: National Research Agency

Coordinator: Inria Lille (O. Maillard)

Duration: 2016–2020

Abstract: Motivated by the fact that a number of modern applications of sequential decision making require developing strategies that are especially robust to change in the stationarity of the signal, and in order to anticipate and impact the next generation of applications of the field, the BADASS project intends to push theory and application of MAB to the next level by incorporating non-stationary observations while retaining near optimality against the best not necessarily constant decision strategy. Since a non-stationary process typically decomposes into chunks associated with some possibly hidden variables (states), each corresponding to a stationary process, handling non-stationarity crucially requires exploiting the (possibly hidden) structure of the decision problem. For the same reason, a MAB for which arms can be arbitrary non-stationary processes is powerful enough to capture MDPs and even partially observable MDPs as special cases, and it is thus important to jointly address the issue of non-stationarity together with that of structure. In order to advance these two nested challenges from a solid theoretical standpoint, we intend to focus on the following objectives: *(i)* To broaden the range of optimal strategies for stationary MABs: current strategies are only known to be provably optimal in a limited range of scenarios for which the class of distribution (structure) is perfectly known; also, recent heuristics possibly adaptive to the class need to be further analyzed. *(ii)* To strengthen the literature on pure sequential prediction (focusing on a single arm) for non-stationary signals via the construction of adaptive confidence sets and a novel measure of complexity: traditional approaches consider a worst-case scenario and are thus overly conservative and non-adaptive to simpler signals. *(iii)* To embed the low-rank matrix completion and spectral methods in the context of reinforcement learning, and further study models of structured environments: promising heuristics in the context of e.g. contextual MABs or Predictive State Representations require stronger theoretical guarantees.

This project will result in the development of a novel generation of strategies to handle non-stationarity and structure that will be evaluated in a number of test beds and validated by a rigorous theoretical analysis. Beyond the significant advancement of the state of the art in MAB and RL theory and the mathematical value of the program, this JCJC BADASS is expected to strategically impact societal and industrial applications, ranging from personalized health-care and e-learning to computational sustainability or rain-adaptive river-bank management to cite a few.

9.2.4. Grant of Fondation Mathématique Jacques Hadamard

Participants: Michal Valko, Ronan Fruit.

Title: Theoretically grounded efficient algorithms for high-dimensional and continuous reinforcement learning

Type: PGMO-IRMO, funded by Criteo

PI: Michal Valko

Criteo contact: Marc Abeille

Duration: 2018–2020

Abstract: While learning how to behave optimally in an unknown environment, a reinforcement learning (RL) agent must trade off the exploration needed to collect new information about the dynamics and reward of the environment, and the exploitation of the experience gathered so far to gain as much reward as possible. A good measure of the agent's performance is the regret, which measures the difference between the performance of optimal policy and the actual rewards accumulated by the agent. Two common approaches to the exploration-exploitation dilemma with provably good regret guarantees are the optimism in the face of uncertainty principle and Thompson Sampling. While these approaches have been successfully applied to small environments with a finite number of states and action (tabular scenario), existing approach for large or continuous environments either rely on heuristics and come with no regret guarantees, or can be proved to achieve small regret but cannot be implemented efficiently. In this project, we propose to make a significant contribution in the understanding of large and/or continuous RL problems by developing and analyzing new algorithms that perform well both in theory and practice.

This research line can have a practical impact in all the applications requiring continuous interaction with an unknown environment. Recommendation systems belong to this category and, by definition, they can be modeled has a sequence of repeated interaction between a learning agent and a large (possibly continuous) environment.

9.2.5. With CIRAD and CGIAR

Participants: Philippe Preux, Odalric-Ambrym Maillard, Romain Gautron.

Title: Crop management

Duration: 2019–2022

Abstract: We study how reinforcement learning may be used to provide recommendations of practices to small farm holders in under-developped countries. In such countries, agriculture remains mostly a non mechanized activity, dealing with fields of very small surface.

This is a very challenging application for RL: data is scarce, recommendations made to farmers should be of quality: we can not just learn by making millions of bad recommendations to people who use them to live and feed their family. Modeling the problem as an RL is yet an other challenge.

We feel that it is very interesting to challenge RL with such complex tasks. Solving games with RL is nice and fun, but we should assess RL abilities to solve real risky tasks.

This pioneering work is done within Romain Gautron's PhD, in collaboration with CIRAD, the CGIAR, and in relation with the Africa Rising program.

9.2.6. Project CNRS-INSERM REPOS

Participants: Émilie Kaufmann, Clémence Réda [INSERM].

Title: Repositionnement de médicaments basé sur leurs effets transcriptionnels par des approches de réseaux géniques

Type: Appel à projet Santé Numérique

PI: Pr. Andrée Delahaye-Duriez (INSERM, UMR1141)

Duration: 2019

Abstract: Drug repurposing consists in studying molecules already commercialized and find other therapies in which they may be efficient. The quality of therapeutic components is often assessed by their affinity to a given protein, but it can also be assessed in terms of their impact at the transciptomic level. The aim of this project is to develop a method for selecting which drugs could be used for a given disease based on their ability to inverse the transcriptomic signature of a pathological phenotype. We will propose a new method based on algorithms for sequential decision making (bandit algorithms) to adaptively select which drug should be explored, where exploring a drug means performing simulations to propagate the perturbation (using for example gene regulatory networks) and estimate the transcriptomic impact of the perturbation induced by the drug. These simulations will hinge on existing gene expression data that are already available for many drugs, but also on new transcriptomic data generated for a mouse model of a rare disease called the Ondine syndrom.

9.2.7. National Partners

- ENS Paris-Saclay
  - M. Valko collaborated with V. Perchet on structured bandit problem. They co-supervise a PhD student (P. Perrault) together
  - O-A. Maillard collaborates with V. Perchet on automated feature learning. They co-supervise a PhD student (R. Ouhamma) together
  - E. Kaufmann collaborated with V. Perchet and E. Boursier on Multi-Player bandits
- Institut de Mathématiques de Toulouse, then Ecole Normale Supérieure de Lyon
  - E. Kaufmann collaborated with Aurélien Garivier on sequential testing and structured bandit problems
- Centrale-Supélec Rennes:
  - E. Kaufmann co-advises Lilian Besson, who works at CentraleSupélec with Christophe Moy on MAB for cognitive radio and Internet-of-Things communications
- Participation to the Inria Project Lab (IPL) "HPC – Big Data": Started in 2018, this IPL gathers a dozen Inria team-projects, mixing researchers in HPC with researchers in machine learning and data science. SEQUEL contribution in this project is about how we can take advantage of HPC for our computational needs regarding deep learning and deep reinforcement learning, and also how such learning algorithms might be redesigned or re-implemented in order to take advantage of HPC architectures.
- Participation to the Inria Project Lab (IPL) "HYAIAI": Started in 2019, this IPL gathers Magnet and SEQUEL in Lille, Tau in Saclay, Lacodam in Rennes, Orpailleur and Multispeech in Nancy. The goal of this IPL is to study machine learning combining symbolic and numeric approaches, to obtain interpretable AI systems.
- PCIM (École Polytechnique)
  - Ph. Preux collaborates with Tanguy Levent (PhD student) on the control of smartgrids with reinforcement learning
- Defrost (Inria Lille)
  - Ph. Preux collaborates with Pierre Schegg (PhD student) on the control of soft robots with reinforcement learning

9.3. European Initiatives

9.3.1. Collaborations in European Programs, Except FP7 & H2020

9.3.1.1. DELTA

Participants: Michal Valko, Émilie Kaufmann, Omar Darwiche Domingues, Pierre Ménard.

Program: CHIST-ERA

Project acronym: DELTA

Project title: Dynamically Evolving Long-Term Autonomy

Duration: October 2017 - December 2021

Coordinator: Anders Jonsson (PI)

Inria Coordinator: Michal Valko

Other partners: UPF Spain, MUL Austria, ULG Belgium

Abstract: Many complex autonomous systems (e.g., electrical distribution networks) repeatedly select actions with the aim of achieving a given objective. Reinforcement learning (RL) offers a powerful framework for acquiring adaptive behaviour in this setting, associating a scalar reward with each action and learning from experience which action to select to maximise long-term reward. Although RL has produced impressive results recently (e.g., achieving human-level play in Atari games and beating the human world champion in the board game Go), most existing solutions only work under strong assumptions: the environment model is stationary, the objective is fixed, and trials end once the objective is met. The aim of this project is to advance the state of the art of fundamental research in lifelong RL by developing several novel RL algorithms that relax the above assumptions. The new algorithms should be robust to environmental changes, both in terms of the observations that the system can make and the actions that the system can perform. Moreover, the algorithms should be able to operate over long periods of time while achieving different objectives. The proposed algorithms will address three key problems related to lifelong RL: planning, exploration, and task decomposition. Planning is the problem of computing an action selection strategy given a (possibly partial) model of the task at hand. Exploration is the problem of selecting actions with the aim of mapping out the environment rather than achieving a particular objective. Task decomposition is the problem of defining different objectives and assigning a separate action selection strategy to each. The algorithms will be evaluated in two realistic scenarios: active network management for electrical distribution networks, and microgrid management. A test protocol will be developed to evaluate each individual algorithm, as well as their combinations.

9.4. International Initiatives

9.4.1. Inria International Partners

- É. Kaufmann visited CWI, Amsterdam for one week in February, working with Wouter Koolen, Rémy Degenne and Rianne De Heide. Pierre Ménard also collaborated with them.

9.5. International Research Visitors

9.5.1. Visits of International Scientists

- Anders Jonsson, Pompeu Fabra University, Spain ,sabbatical year Sep 2019 – Jul 2020
- Kaige Yang, University College London, UK, Oct 9 & Jan 9 2020
- Rianne de Heide, CWI, The Netherlands, April 23 – August 3, 2019
- Chuan-Zheng Lee, Stanford University, USA, June – October 2019
- Arun Verma, IIT Bombay, June 1 – November 30, 2019

9.5.1.1. Internships

- Alessio Della Libera, from Jul 2019 until Sep 2019

  TD-Gammon, and his github with the gym-backgammon code

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events: Organisation

- the 1st Reinforcement Learning Summer Scool, July 1-12, 2019, Villeneuve d'Ascq
- the 3rd Vigil workshop at NeurIPS 2019

10.1.1.1. Member of the Organizing Committees

- F. Strub, co-organizer of the workshop "Visually Grounded Interaction and Language (ViGIL)" at NeurIPS 2019
- The whole SEQUEL team has organized RLSS

10.1.2. Scientific Events: Selection

10.1.2.1. Member of the Conference Program Committees

- Émilie Kaufmann: ALT
- Odalric-Ambrym Maillard: ICML, ECAI, SIF
- Philippe Preux: ECML, EGC, SFC

10.1.2.2. Reviewer

In 2019, we have reviewed submissions for: AI&Stats, NeurIPS, ALT, ICML, COLT, IJCAI, AAAI, CDC, ECAI

10.1.3. Journal

10.1.3.1. Reviewer - Reviewing Activities

- Journal of Machine Learning Research
- Journal of Artificial Intelligence Research
- The Annals of Statistics
- Bernoulli
- IEEE Transactions on Knowledge and Data Engineering
- Machine Learning
- Information and Inference: A Journal of the IMA

10.1.4. Invited Talks

- E. Kaufmann
  - "Beyond Classical Bandit Tools for Monte-Carlo Tree Search", AAAI workshop on Reinforcement Learning for Games, Honolulu, Jan 2019
  - "New tools for Adaptive Testing and Applications to Bandit Problems", Machine Learning and Optimization Working Group, Ecole des Ponts, Feb 2019
  - "Generalized Likelihood Ratios Tests applied to Sequential Decision Making", Statistics Seminar, Agro ParisTech, Paris, May 2019
  - "Generalized Likelihood Ratios Tests applied to Sequential Decision Making", Machine Learning Seminar, University of Leiden, The Netherlands, May 2019
  - "Quelques outils statistiques pour la prise de décision séquentielle", Conférence plénière du GRETSI, Lille, Aug 2019
  - "Practical algorithm for multi-player bandits", MAPLE workshop, Milan, Italy Sep 2019
  - "Practical algorithm for multi-player bandits", Invited session of the Allerton Conference, Urbana-Champaign, USA Sep 2019
- Odalric-Ambrym Maillard:
  - "La prise de décision séquentielle au service de la société de demain", Euratechnologie, Lille, Feb 2019

  - "Change of mean detection, non-asymptotic delay and aggregation", 3rd non-stationary day, Institut Henry Poincaré, Paris, Mar 2019
  - "A tour of time-uniform concentration inequalities: Laplace, Peeling, Kernel", Workshop on empirical Processes and Applications to Statistics, Besançon, May 2019
  - "A tour of time-uniform concentration inequalities: Laplace, Peeling, Kernel", CWI, Amsterdam, The Netherlands, Jun 2019
  - "Reinforcement Learning: successes and promises", Ecole Polytechnique, Palaiseau, Nov 2019
- Philippe Preux:
  - A brief introduction to supervised learning and reinforcement learning, 1st humAIn seminar, Villeneuve d'Ascq, Feb 2019
  - "Sous le contrôle des bandits", AFCE, June 2019
  - Explainability in machine learning, 3rd humAIn seminar, Lille, June 2019
  - "Learning to act", ENS-Paris-Saclay, Conférence de rentrée, Sep 2019
  - "Apprentissage par renforcement : mythe et réalité", FOOR, Tourcoing, Nov 2019
- Jill-Jênn Vie:
  - "IA, éducation et formation", Hermès, Paris, Oct 2019
  - "JJ Vie's Factorization IV", LaBRI, Bordeaux, Nov 2019
  - "Deep Learning for Anime & Manga", Paris Open Source Summit, Dec 2019
  - "Deep Learning for Recommender Systems", Université Cergy-Pontoise, Dec 2019
- R. Gautron, O-A. Maillard, Ph. Preux, "Reinforcement learning for crop-management: a sequential decision-making under uncertainty approach", CGIAR convention, Hydebarad, India, Oct 2019
- Ph. Preux, M. Seurin, "L'IA, les données, ... et l'Homme dans tout ça ?", congress "Les données et leurs usages dans les technologies du numérique", Douai, Oct 2019

10.1.5. Scientific Expertise

- Émilie Kaufmann:
  - member of the hiring committee for an assistant professor in probability/statistics at Université Paris-Sud
- Odalric-Ambrym Maillard:
  - member of the hiring committee for CRCN at Inria Lille
- Philippe Preux:
  - member of the hiring committee for CRCN at Inria Rennes
  - member of the hiring committee for CRCN at Inria (national)
  - evaluation of submissions to ANRT (he also declined many such invitations due to lack of time, *e.g.* with ANR)

10.1.6. Research Administration

- Odalric-Ambrym Maillard is:
  - member of the CER at Inria Lille
- Philippe Preux is:
  - "délégué scientifique adjoint" of the Inria center in Lille
  - member of the Inria evaluation committee (CE)
  - member of the Inria internal scientific committee (COSI)
  - member of the scientific committee of CRIStAL until Jan 2019

– the head of the "Data Intelligence" thematic group at CRIStAL until Jan 2019

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Doctorat: Émilie Kaufmann and Odalric-Ambrym-Maillard, "Bandit algorithms I", RLSS Summer School, Lille, 9h, July 2019

Master: Émilie Kaufmann, "Data Mining", 36h, M1, Université de Lille, Jan-Apr 2019

Master: Émilie Kaufmann, "Reinforcement Learning", 24h, M2, Ecole Centrale de Lille, Nov 2019-Jan 2020

Master: Odalric-Ambrym Maillard, "Reinforcement Learning", 38h equivalent TD, M2, Ecole Polytechnique, Palaiseau, Jan-Mar 2019

Doctorat: Odalric-Ambrym Maillard, "Bandit algorithms II", RLSS Summer School, Lille, 9h, July 2019

Doctorat: Philippe Preux, "Reinforcement Learning", Fall School on AI (IA2) of the GDR IA (CNRS), Lyon, 3h, Oct 2019

Doctorat: Philippe Preux, "AI learns to act", MOMI, Sophia-Antipolis, 1h30, Feb 2019

10.2.2. Supervision

HdR: Odalric-Ambrym Maillard, Mathematics of Sequential Decision Making, Université de Lille, Feb 11, 2019

PhD: Lilian Besson, Multi-players Bandit Algorithms for Internet of Things Networks, Centrale-Supélec Rennes, Nov 20, 2019, supervisors: Christophe Moy (Université de Rennes) et Émilie Kaufmann

PhD: Ronan Fruit, Exploration–exploitation dilemma in Reinforcement Learning under various form of prior knowledge, Université de Lille, Nov 6, 2019, supervisor: Alessandro Lazaric

PhD: Nicolas Carrara, "Apprentissage par renforcement pour optimisation de systèmes de dialogue via l'adaptation à chaque utilisateur", Université de Lille, Dec 18, 2019, supervisor: Oliier Pietquin

PhD in progress: Dorian Baudry, "Efficient Exploration for Structured Bandits and Reinforcement Learning", since Nov 2019, supervisors: É. Kaufmann, O-A. Maillard

PhD in progress: Omar Darwiche Domingues, "Sequential Learning in Dynamic Environments", since Oct 2018, supervisors: É. Kaufmann, M. Valko

PhD in progress: Johan Ferret, "Explainable Reinforcement Learning via Deep Neural Networks", since Fall 2019, supervisor: Ph. Preux, O. Pietquin

PhD in progress: Yannis Flet-Berliac, "Deep reinforcement learning in stochastic and non stationary environments", since Oct 2018, supervisor: Ph. Preux

PhD in progress: Guillaume Gautier, DPPs in ML, started Oct 2016, defense scheduled in March 2020. Supervisors: R. Bardenet, M. Valko.

PhD in progress: Jean-Bastien Grill, "Création et analyse d'algorithmes efficaces pour la prise de décision dans un environnement inconnu et incertain", started Oct 2014, defended on Dec 19, 2019. Supervisors: R. Munos, M. Valko

PhD in progress: Nathan Grinsztajn, "Apprentissage par renforcement pour la résolution séquentielle de problèmes d'optimisation combinatoire incertains et partiellement définis", since Fall 2019, supervisor: Ph. Preux

PhD in progress: Léonard Hussenot, "Adversarial reinforcement learning: attacks and robustness", since Fall 2019, supervisor: Ph. Preux, O. Pietquin

PhD in progress: Édouard Leurent, “Autonomous vehicle control: application of machine learning to contextualized path planning”, since Oct 2017, supervisors: O-A. Maillard, D. Effimov (Valse), W. Perruquetti (CRIStAL)

PhD in progress: Reda Ouhamma, “Automated feature representation”, since Fall 2019, O-A. Maillard

PhD in progress: Pierre Perrault, “Online Learning on Streaming Graphs”, since Sep 2017, supervisors: M. Valko, V. Perchet

PhD in progress: Sarah Perrin, “Reinforcement Learning in Mean Field Games”, since Fall 2019, supervisors: O. Pietquin, R. Elie

PhD in progress: Hassan Saber, “Structured multi-armed bandits”, since Oct 2018, Structured Multi-armed bandits, supervisor: O-A. Maillard.

PhD in progress: Mathieu Seurin, “Multi-scale rewards in reinforcement learning”, since Oct 2017, supervisors: O. Pietquin, Ph. Preux

PhD in progress: Julien Seznec, “Sequential Learning for Educational Systems”, since Mar 2017, supervisors: M. Valko, A. Lazaric, J. Banon

PhD in progress: Xuedong Shang, “Adaptive methods for optimization in stochastic environments”, started Oct 2017, supervisors: É. Kaufmann, M. Valko

PhD in progress: Florian Strub, “Reinforcement Learning for visually grounded interaction”, since Jan 2016, defense scheduled for Jan 2020, supervisors: O. Pietquin and J. Mary

PhD in progress: Kiewan Villatel, “Deep Learning for Conversion Rate Prediction in Online Advertising”, started Oct 2017, aborted June 2019, supervisor: Ph. Preux

10.2.3. Juries

- Émilie Kaufmann:
  - Aristide Tossou, member of the jury, Chalmers University, Sweden, Nov 18, 2019
  - Rémi Degenne, member of the jury, Université Paris-Diderot, Dec 18, 2019
  - member of the Mathematics jury for the admission competition of ENS, section B/L
- Odalric-Ambrym Maillard:
  - Léonard Torossian, reviewer, Université Toulouse III, Dec 17, 2019.
- Philippe Preux:
  - Quentin Waymel (medical doctorate), member of the jury, Université de Lille, Jun 2019
  - Adrien Legrand, reviewer, Université de Picardie, Amiens, Nov 29, 2019
  - Erinc Merdivan, reviewer, Centrale-Supélec Metz, Dec 17, 2019
  - Nicolas Carrara, member of the jury, Université de Lille, Dec 18, 2019
- Michal Valko:
  - Aristide Tossou, opponent, Chalmers University, Sweden, Nov 18, 2019

10.3. Popularization

10.3.1. Articles and contents

- Philippe Preux:
  - interviewed by *Le Monde* pulished in Sep 2019
  - interview on I-SITE project B4H, Inria

10.3.2. Education

- Odalric-Ambrym Maillard:

– "Reinforcement Learning: successes and promises", Executive Master, Ecole Polytechnique, Palaiseau, Nov 2019

10.3.3. Interventions

- Philippe Preux:
  - panel on "Promises and perils of AI", CGIAR, Hyderabad, India, Oct 2019
  - panel on "AI and man", Euratechnologies, Lille, Sep 2019
- Yannis Flet-Berliac and Philippe Preux: panel on "Who's the pilot: man of software?", FOOR, Le Fresnoy, Tourcoing, Nov 2019
- Yannis Flet-Berliac: "Princess of parallelograms" installation (with Thomas Depas), Le Fresnoy, Tourcoing, "Damien & The Love Guru" gallery in Brussels, Belgium, Sep–Dec 2019

Figure 1.

Princess of parallelograms is a collaborative project between Yannis Flet-Berliac and a student from Le Fresnoy National Studio of Contemporary Arts. They created an interactive sculpture made of a variety of computer vision attributes: a support for anthropomorphic projections, a set of generated virtual masks, or a new form of photographic trap. When visitors stand in front of the device's webcam, a Deep Convolutional Conditional-GAN Auto-Encoder model applies a filter on their face with virtual flesh, hair, and facial expressions in real-time. In the meantime, an emotion detection model trained on the FER-2013 dataset is running in the background. The system allows the users to actively interact with the installation. So far, the project has been exposed at Le Fresnoy and in Brussels.

11. Bibliography

Major publications by the team in recent years

[1] O. CAPPÉ, A. GARIVIER, O.-A. MAILLARD, R. MUNOS, G. STOLTZ. *Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation*, in "Annals of Statistics", 2013, vol. 41, n^{o} 3, pp. 1516-1541, https://hal.archives-ouvertes.fr/hal-00738209

[2] A. CARPENTIER, M. VALKO. *Revealing Graph Bandits for Maximizing Local Influence*, in "Proceedings of the 19th International Conference on Artificial Intelligence and Statistics", Cadiz, Spain, A. GRETTON, C. C. ROBERT (editors), Proceedings of Machine Learning Research, PMLR, May 2016, vol. 51, pp. 10-18, http://proceedings.mlr.press/v51/carpentier16a.html

[3] H. DE VRIES, F. STRUB, J. MARY, H. LAROCHELLE, O. PIETQUIN, A. COURVILLE. *Modulating early visual processing by language*, in "Conference on Neural Information Processing Systems", Long Beach, United States, December 2017, pp. 6594-6604, https://hal.inria.fr/hal-01648683

[4] N. GATTI, A. LAZARIC, M. ROCCO, F. TROVÒ. *Truthful Learning Mechanisms for Multi–Slot Sponsored Search Auctions with Externalities*, in "Artificial Intelligence", October 2015, vol. 227, pp. 93-139, https://hal.inria.fr/hal-01237670

[5] M. GHAVAMZADEH, Y. ENGEL, M. VALKO. *Bayesian Policy Gradient and Actor-Critic Algorithms*, in "Journal of Machine Learning Research", January 2016, vol. 17, n^{o} 66, pp. 1-53, https://hal.inria.fr/hal-00776608

[6] H. KADRI, E. DUFLOS, P. PREUX, S. CANU, A. RAKOTOMAMONJY, J. AUDIFFREN. *Operator-valued Kernels for Learning from Functional Response Data*, in "Journal of Machine Learning Research (JMLR)", April 2016, vol. 17, n^{o} 20, pp. 1-54, https://hal.archives-ouvertes.fr/hal-01221329

[7] E. KAUFMANN, O. CAPPÉ, A. GARIVIER. *On the Complexity of Best Arm Identification in Multi-Armed Bandit Models*, in "Journal of Machine Learning Research", January 2016, vol. 17, pp. 1-42, https://hal.archives-ouvertes.fr/hal-01024894

[8] A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Analysis of Classification-based Policy Iteration Algorithms*, in "Journal of Machine Learning Research", 2016, vol. 17, pp. 1-30, https://hal.inria.fr/hal-01401513

[9] R. MUNOS. *From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning*, in "Foundations and Trends in Machine Learning", 2014, vol. 7, n^{o} 1, pp. 1-129, http://dx.doi.org/10.1561/2200000038

[10] R. ORTNER, D. RYABKO, P. AUER, R. MUNOS. *Regret bounds for restless Markov bandits*, in "Journal of Theoretical Computer Science (TCS)", 2014, vol. 558, pp. 62-76 [*DOI :* 10.1016/J.TCS.2014.09.026], https://hal.inria.fr/hal-01074077

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] N. CARRARA. *Reinforcement learning for Dialogue Systems optimization with user adaptation*, Ecole Doctoral Science pour l'Ingénieur Université Lille Nord-de-France, December 2019, https://tel.archives-ouvertes.fr/tel-02422691

[12] R. FRUIT. *Exploration-exploitation dilemma in Reinforcement Learning under various form of prior knowledge*, Université de Lille 1, Sciences et Technologies; CRIStAL UMR 9189, November 2019, https://tel.archives-ouvertes.fr/tel-02388395

[13] O.-A. MAILLARD. *Mathematics of Statistiscal Sequential Decision Making*, Université de Lille Nord de France, February 2019, Habilitation à diriger des recherches, https://hal.archives-ouvertes.fr/tel-02077035

Articles in International Peer-Reviewed Journals

[14] M.-A. CHARPAGNE, F. STRUB, T. M. POLLOCK. *Accurate reconstruction of EBSD datasets by a multimodal data approach using an evolutionary algorithm*, in "Materials Characterization", April 2019, vol. 150, pp.

184-198, https://arxiv.org/abs/1903.02988 - A short version of this paper exists towards people working in Machine Learning, namely arxiv:1903.02982 [*DOI :* 10.1016/J.MATCHAR.2019.01.033], https://hal.archives-ouvertes.fr/hal-02062098

[15] A. R. LUEDTKE, E. KAUFMANN, A. CHAMBAZ. *Asymptotically Optimal Algorithms for Budgeted Multiple Play Bandits*, in "Machine Learning Journal", September 2019, vol. 108, n° 11, pp. 1919-1949, https://arxiv.org/abs/1606.09388 , https://hal.archives-ouvertes.fr/hal-01338733

International Conferences with Proceedings

[16] *Best Paper*
M. ASADI, M. S. TALEBI, H. BOUREL, O.-A. MAILLARD. *Model-Based Reinforcement Learning Exploiting State-Action Equivalence*, in "ACML 2019, Proceedings of Machine Learning Research", Nagoya, Japan, 2019, vol. 101, pp. 204 - 219, https://hal.archives-ouvertes.fr/hal-02378887.

[17] P. BARTLETT, V. GABILLON, J. HEALEY, M. VALKO. *Scale-free adaptive planning for deterministic dynamics & discounted rewards*, in "International Conference on Machine Learning", Long Beach, United States, 2019, https://hal.inria.fr/hal-02387484

[18] P. BARTLETT, V. GABILLON, M. VALKO. *A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption*, in "Algorithmic Learning Theory", Chicago, United States, 2019, https://hal.inria.fr/hal-01885368

[19] D. CALANDRIELLO, L. CARRATINO, A. LAZARIC, M. VALKO, L. ROSASCO. *Gaussian process optimization with adaptive sketching: Scalable and no regret*, in "Conference on Learning Theory", Phoenix, United States, 2019, https://hal.inria.fr/hal-02144311

[20] N. CARRARA, E. LEURENT, R. LAROCHE, T. URVOY, O.-A. MAILLARD, O. PIETQUIN. *Budgeted Reinforcement Learning in Continuous State Space*, in "Conference on Neural Information Processing Systems", Vancouver, Canada, Advances in Neural Information Processing Systems, December 2019, vol. 32, https://arxiv.org/abs/1903.01004 , https://hal.archives-ouvertes.fr/hal-02375727

[21] M. DEREZIŃSKI, D. CALANDRIELLO, M. VALKO. *Exact sampling of determinantal point processes with sublinear time preprocessing*, in "Neural Information Processing Systems", Vancouver, Canada, 2019, https://hal.inria.fr/hal-02387524

[22] C. DIMITRAKAKIS, Y. LIU, D. PARKES, G. RADANOVIC. *Bayesian Fairness*, in "AAAI 2019 - Thirty-Third AAAI Conference on Artificial Intelligence", Honolulu, United States, January 2019, https://hal.inria.fr/hal-01953311

[23] G. GAUTIER, R. BARDENET, M. VALKO. *On two ways to use determinantal point processes for Monte Carlo integration – Long version*, in "NeurIPS 2019 - Thirty-third Conference on Neural Information Processing Systems", Vancouver, Canada, Advances in Neural Information Processing Systems, 2019, https://hal.archives-ouvertes.fr/hal-02277739

[24] J.-B. GRILL, O. D. DOMINGUES, P. MÉNARD, R. MUNOS, M. VALKO. *Planning in entropy-regularized Markov decision processes and games*, in "Neural Information Processing Systems", Vancouver, Canada, 2019, https://hal.inria.fr/hal-02387515

[25] E. LEURENT, O.-A. MAILLARD. *Practical Open-Loop Optimistic Planning*, in "European Conference on Machine Learning", Würzburg, Germany, European Conference on Machine Learning, September 2019, https://arxiv.org/abs/1904.04700 , https://hal.archives-ouvertes.fr/hal-02375697

[26] A. LOCATELLI, A. CARPENTIER, M. VALKO. *Active multiple matrix completion with adaptive confidence sets*, in "International Conference on Artificial Intelligence and Statistics", Okinawa, Japan, 2019, https://hal.inria.fr/hal-02387468

[27] O.-A. MAILLARD. *Sequential change-point detection: Laplace concentration of scan statistics and non-asymptotic delay bounds*, in "Algorithmic Learning Theory", Chicago, United States, 2019, vol. 98, pp. 1 - 23, https://hal.archives-ouvertes.fr/hal-02351665

[28] C. MOY, L. BESSON. *Decentralized Spectrum Learning for IoT Wireless Networks Collision Mitigation*, in "ISIoT 2019 - 1st International Workshop on Intelligent Systems for the Internet of Things", Santorin, Greece, May 2019, https://arxiv.org/abs/1906.00614 , https://hal.inria.fr/hal-02144465

[29] R. ORTNER, M. PIROTTA, R. FRUIT, A. LAZARIC, O.-A. MAILLARD. *Regret Bounds for Learning State Representations in Reinforcement Learning*, in "Conference on Neural Information Processing Systems", Vancouver, Canada, Conference on Neural Information Processing Systems, December 2019, https://hal.archives-ouvertes.fr/hal-02375715

[30] P. PERRAULT, V. PERCHET, M. VALKO. *Exploiting structure of uncertainty for efficient matroid semi-bandits*, in "International Conference on Machine Learning", Long Beach, United States, 2019, https://hal.inria.fr/hal-02387478

[31] P. PERRAULT, V. PERCHET, M. VALKO. *Finding the bandit in a graph: Sequential search-and-stop*, in "International Conference on Artificial Intelligence and Statistics", Okinawa, Japan, 2019, https://hal.inria.fr/hal-02387465

[32] J. SEZNEC, A. LOCATELLI, A. CARPENTIER, A. LAZARIC, M. VALKO. *Rotting bandits are not harder than stochastic ones*, in "International Conference on Artificial Intelligence and Statistics", Naha, Japan, 2019, https://hal.inria.fr/hal-01936894

[33] X. SHANG, E. KAUFMANN, M. VALKO. *A simple dynamic bandit algorithm for hyper-parameter tuning*, in "Workshop on Automated Machine Learning at International Conference on Machine Learning", Long Beach, United States, AutoML@ICML 2019 - 6th ICML Workshop on Automated Machine Learning, June 2019, https://hal.inria.fr/hal-02145200

[34] X. SHANG, E. KAUFMANN, M. VALKO. *General parallel optimization without a metric*, in "Algorithmic Learning Theory", Chicago, United States, 2019, vol. 98, https://hal.inria.fr/hal-02047225

[35] M. S. TALEBI, O.-A. MAILLARD. *Learning Multiple Markov Chains via Adaptive Allocation*, in "Advances in Neural Information Processing Systems 32 (NIPS 2019)", Vancouver, Canada, December 2019, https://hal.archives-ouvertes.fr/hal-02387345

National Conferences with Proceedings

[36] L. BESSON, E. KAUFMANN. *Non-asymptotic analysis of a sequential rupture detection test and its application to non-stationary bandits*, in "GRETSI 2019 - XXVIIème Colloque francophone de traitement du signal et des images", Lille, France, August 2019, https://hal.inria.fr/hal-02152243

Conferences without Proceedings

[37] L. BESSON, R. BONNEFOI, C. MOY. *GNU Radio Implementation of MALIN: "Multi-Armed bandits Learning for Internet-of-things Networks"*, in "IEEE WCNC 2019 - IEEE Wireless Communications and Networking Conference", Marrakech, Morocco, April 2019, https://arxiv.org/abs/1902.01734 , https://hal.inria.fr/hal-02006825

[38] R. BONNEFOI, L. BESSON, J. MANCO-VASQUEZ, C. MOY. *Upper-Confidence Bound for Channel Selection in LPWA Networks with Retransmissions*, in "The 1st International Workshop on Mathematical Tools and technologies for IoT and mMTC Networks Modeling", Marrakech, Morocco, Philippe Mary, Samir Perlaza, Petar Popovski, April 2019, https://arxiv.org/abs/1902.10615 - The source code (MATLAB or Octave) used for the simula-tions and the figures is open-sourced under the MIT License, atBitbucket.org/scee_ietr/ucb_smart_retrans, https://hal.inria.fr/hal-02049824

[39] Y. FLET-BERLIAC, P. PREUX. *MERL: Multi-Head Reinforcement Learning*, in "NeurIPS 2019 Deep Reinforcement Learning Workshop", Vancouver, Canada, December 2019, https://arxiv.org/abs/1909.11939 , https://hal.inria.fr/hal-02305105

[40] G. GAUTIER, R. BARDENET, M. VALKO. *On two ways to use determinantal point processes for Monte Carlo integration*, in "NEGDEPML 2019 - ICML Workshop on Negative Dependence in ML", Long Beach, CA, United States, June 2019, https://hal.archives-ouvertes.fr/hal-02160382

[41] T. LEVENT, P. PREUX, E. LE PENNEC, J. BADOSA, G. HENRI, Y. BONNASSIEUX. *Energy Management for Microgrids: a Reinforcement Learning Approach*, in "ISGT-Europe 2019 - IEEE PES Innovative Smart Grid Technologies Europe", Bucharest, France, IEEE, September 2019, pp. 1-5 [*DOI :* 10.1109/ISGTEUROPE.2019.8905538], https://hal.archives-ouvertes.fr/hal-02382232

[42] M. SEURIN, P. PREUX, O. PIETQUIN. *"I'm sorry Dave, I'm afraid I can't do that" Deep Q-Learning From Forbidden Actions*, in "Workshop on Safety and Robustness in Decision Making (NeurIPS 2019)", Vancouver, Canada, December 2019, https://hal.inria.fr/hal-02387419

Other Publications

[43] L. BESSON, E. KAUFMANN. *The Generalized Likelihood Ratio Test meets klUCB: an Improved Algorithm for Piece-Wise Non-Stationary Bandits*, February 2019, https://arxiv.org/abs/1902.01575 - working paper or preprint, https://hal.inria.fr/hal-02006471

[44] E. BOURSIER, E. KAUFMANN, A. MEHRABIAN, V. PERCHET. *A Practical Algorithm for Multiplayer Bandits when Arm Means Vary Among Players*, May 2019, https://arxiv.org/abs/1902.01239 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02006069

[45] G. CIDERON, M. SEURIN, F. STRUB, O. PIETQUIN. *Self-Educated Language Agent With Hindsight Experience Replay For Instruction Following*, November 2019, https://arxiv.org/abs/1910.09451 - working paper or preprint [*DOI :* 10.09451], https://hal.archives-ouvertes.fr/hal-02386585

[46] R. DEGENNE, W. M. KOOLEN, P. MÉNARD. *Non-Asymptotic Pure Exploration by Solving Games*, December 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02402665

[47] Y. FLET-BERLIAC, P. PREUX. *High-Dimensional Control Using Generalized Auxiliary Tasks*, November 2019, working paper or preprint, https://hal.inria.fr/hal-02295705

[48] Y. FLET-BERLIAC, P. PREUX. *Samples Are Useful? Not Always: denoising policy gradient updates using variance explained*, September 2019, https://arxiv.org/abs/1904.04025 - working paper or preprint, https://hal.inria.fr/hal-02091547

[49] A. GARIVIER, H. HADIJI, P. MÉNARD, G. STOLTZ. *KL-UCB-switch: optimal regret bounds for stochastic bandits from both a distribution-dependent and a distribution-free viewpoints*, November 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01785705

[50] A. GARIVIER, E. KAUFMANN. *Non-Asymptotic Sequential Tests for Overlapping Hypotheses and application to near optimal arm identification in bandit models*, May 2019, https://arxiv.org/abs/1905.03495 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02123833

[51] E. LEURENT, Y. BLANCO, D. EFIMOV, O.-A. MAILLARD. *Approximate Robust Control of Uncertain Dynamical Systems*, February 2019, https://arxiv.org/abs/1903.00220 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01931744

[52] E. LEURENT, J. MERCAT. *Social Attention for Autonomous Decision-Making in Dense Traffic*, November 2019, https://arxiv.org/abs/1911.12250 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02383940

[53] O.-A. MAILLARD, T. A. MANN, R. ORTNER, S. MANNOR. *Active Roll-outs in MDP with Irreversible Dynamics*, July 2019, working paper or preprint, https://hal.archives-ouvertes.fr/hal-02177808

[54] X. SHANG, R. DE HEIDE, E. KAUFMANN, P. MÉNARD, M. VALKO. *Fixed-confidence guarantees for Bayesian best-arm identification*, October 2019, https://arxiv.org/abs/1910.10945 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02330187

[55] F. STRUB, M.-A. CHARPAGNE, T. M. POLLOCK. *Accurate reconstruction of EBSD datasets by a multimodal data approach using an evolutionary algorithm*, March 2019, https://arxiv.org/abs/1903.02988 - A short version of this paper exists towards people working in Machine Learning, namely arxiv:1903.02982 [*DOI :* 10.1016/J.MATCHAR.2019.01.033], https://hal.archives-ouvertes.fr/hal-02062104

[56] C. TRINH, E. KAUFMANN, C. VERNADE, R. COMBES. *Solving Bernoulli Rank-One Bandits with Unimodal Thompson Sampling*, December 2019, https://arxiv.org/abs/1912.03074 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-02396943

References in notes

[57] P. AUER, N. CESA-BIANCHI, P. FISCHER. *Finite-time analysis of the multi-armed bandit problem*, in "Machine Learning", 2002, vol. 47, n^{O} 2/3, pp. 235–256

[58] R. BELLMAN. *Dynamic Programming*, Princeton University Press, 1957

[59] D. BERTSEKAS, S. SHREVE. *Stochastic Optimal Control (The Discrete Time Case)*, Academic Press, New York, 1978

[60] D. BERTSEKAS, J. TSITSIKLIS. *Neuro-Dynamic Programming*, Athena Scientific, 1996

[61] M. PUTERMAN. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*, John Wiley and Sons, 1994

[62] H. ROBBINS. *Some aspects of the sequential design of experiments*, in "Bull. Amer. Math. Soc.", 1952, vol. 55, pp. 527–535

[63] R. SUTTON, A. BARTO. *Reinforcement learning: an introduction*, MIT Press, 1998

[64] P. WERBOS. *ADP: Goals, Opportunities and Principles*, IEEE Press, 2004, pp. 3–44, Handbook of learning and approximate dynamic programming