IN PARTNERSHIP WITH:
Université Charles de Gaulle (Lille 3)

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2017

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
A3.4.6. - Neural networks
A3.4.8. - Deep learning
A3.5.2. - Recommendation systems
A5.1. - Human-Computer Interaction
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:

B5.8. - Learning and training
B6.1. - Software industry
B7.2.1. - Smart vehicles
B9.1.1. - E-learning, MOOC
B9.4. - Sciences
B9.4.5. - Data science

1. Personnel

Research Scientists

Émilie Kaufmann [CNRS, Researcher]
Alessandro Lazaric [Inria, Researcher, on secondment at Facebook AI Research since June 2017]
Odalric Maillard [Inria, Researcher]
Daniil Ryabko [Inria, Researcher, HDR]
Michal Valko [Inria, Researcher, HDR]

Faculty Members

Philippe Preux [Team leader, Univ Charles de Gaulle, Professor, on Inria secondment since Sep 1st, 2016, HDR]
Romaric Gaudel [Univ Charles de Gaulle, Associate Professor, until Apr 2017]
Jérémie Mary [Univ Charles de Gaulle, Associate Professor, until May 2017, on secondment at Criteo Research since June 2017, HDR]
Bilal Piot [Univ des sciences et technologies de Lille, Associate Professor, until Jan 2017]

Post-Doctoral Fellows

Ralph Bourdoukan [Univ des sciences et technologies de Lille]
Édouard Oyallon [Inria, since Nov 2017]
Matteo Pirotta [Inria]
James Ridgway [Inria, until Aug 2017]

PhD Students

Marc Abeille [Univ des sciences et technologies de Lille, until Dec 2017]
Sheikh Waqas Akhtar [Inria, since Oct 2017]
Merwan Barlier [Orange Labs]
Alexandre Bérard [Univ des sciences et technologies de Lille]
Lilian Besson [CentraleSupélec Rennes]
Daniele Calandriello [Inria, until Dec 2017]
Nicolas Carrara [Orange Labs]
Ronan Fruit [Inria]
Pratik Gajane [Orange Labs, until Nov 2017]
Jean Bastien Grill [Inria/ENS Paris]
Édouard Leurent [Renault, since Oct 2017]
Alexis Martin [Inria, from Jan to Mar 2017]
Julien Pérolat [Univ des sciences et technologies de Lille, until Dec 2017]
Pierre Perrault [Inria, since Sep 2017]
Mathieu Seurin [Univ des sciences et technologies de Lille, since Sep 2017]
Julien Seznec [LeLivreScolaire.fr, since Mar 2017]
Xuedong Shang [Inria/ENS Rennes, since Oct 2017]
Florian Strub [Univ des sciences et technologies de Lille]
Kiewan Villatel [Critéo, since Oct 2017]
Romain Warlop [55]
Guillaume Gautier [Inria, since Feb 2017]
Harm de Vries [Université de Montréal, until Jun 2017]

Interns

Mahsa Asadi [Inria, from Sep 2017 to Dec 2017]
Subhojyoti Mukherjee [Inria, from Sep 2017 until Nov 2017]
Iuliia Olkhovskaia [Inria, from Feb 2017 until Jul 2017]
Georgios Papoudakis [Univ des sciences et technologies de Lille, from May 2017 until Sep 2017]
Xuedong Shang [Univ des sciences et technologies de Lille, from Feb 2017 until Jun 2017]

Administrative Assistant

Amelie Supervielle [Inria]

Visiting Scientists

Reda Alami [Orange Labs, since Oct 2017]
Aditya Gopalan [ITT Madras, Mar 2017]
Mohammad Sadegh Talebi Mazraeh Shahi [ANR, since Jun 2017 until Sep 2017]
Mohammadi Zaki [Indian institute of Sicence, Mar 2017]

External Collaborators

Rémi Bardenet [CNRS]

Pierre Chainais [Ecole centrale de Lille, HDR]
Jérémie Mary [Criteo, since Jun 2017, HDR]
Olivier Pietquin [Deepmind London, since May 2016, 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 [71]). 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 neither the best one to make a prediction, nor 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 models 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 [69].

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 $r : \mathcal{X} \times \mathcal{A} \times \mathcal{X} \to I\!R$ 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.

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:

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

[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.

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 [68]) and average reward settings. Note also that, here, we considered 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) [66], 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 [67]. 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 ([72]):

- 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 [70], 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.* [65] 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 $x_1, \cdots, x_n$ 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 behaviour (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 behaviour, or than a class of other behaviours.

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 yes then 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) behaviour 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 advertizing and recommandation 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 recommandation system). An other field of applications 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 works 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

- under the supervision of O. Pietquin and J. Mary, F. Strub and collaborators (among which University of Montreal) have introduced the **Guesswhat?!** game to study visually grounded dialogues interleaving vision and natural language. A dataset of 150k human-human dialogues has been collected and is freely available on the Internet. Supervised learning baselines and state-of-the-art reinforcement learning algoritms have been implemented and are available as open-source code. This work resulted in publications in prestigious conferences: as a spotlight at CVPR 2017, an oral at IJCAI 2017, and an other spotlight at NIPS 2017 [51], [29], [30]. Spotlight presentations concern less than 3.5% of submisisons to NIPS, and 5% of submissions to CVPR.

  See https://www.guesswhat.ai
- under the supervision of M. Valko and A. Lazaric, D. Calandriello and collaborators have provided the first breaking quadratic barrier for nonparametric learning. An open source implementation is available on the Internet. The work has been published in prestigious conferences: AI & STATS, ICML and NIPS [26], [28], [27].

6. New Software and Platforms

6.1. BAC

Bayesian Policy Gradient and Actor-Critic Algorithms
KEYWORDS: Machine learning - Incremental learning - Policy Learning
FUNCTIONAL DESCRIPTION: To address this issue, we proceed to supplement our Bayesian policy gradient framework with a new actor-critic learning model in which a Bayesian class of non-parametric critics, based on Gaussian process temporal difference learning, is used. Such critics model the action-value function as a Gaussian process, allowing Bayes' rule to be used in computing the posterior distribution over action-value functions, conditioned on the observed data. Appropriate choices of the policy parameterization and of the prior covariance (kernel) between action-values allow us to obtain closed-form expressions for the posterior distribution of the gradient of the expected return with respect to the policy parameters. We perform detailed experimental comparisons of the proposed Bayesian policy gradient and actor-critic algorithms with classic Monte-Carlo based policy gradient methods, as well as with each other, on a number of reinforcement learning problems.

- Contact: Michal Valko
- URL: https://team.inria.fr/sequel/Software/BAC/

6.2. GuessWhat?!

GuessWhat?! Visual object discovery through multi-modal dialogue
KEYWORDS: Deep learning - Dialogue System
FUNCTIONAL DESCRIPTION: This project train a AI to play the GuessWhat?! game. Thus, you can train an AI to ask questions, to answer questions about images. You can also perform basic visual reasoning. This project is a testbed for future interactive dialogue system.

- Partner: Universite de Montreal
- Contact: Florian Strub
- Publications: GuessWhat?! Visual object discovery through multi-modal dialogue - End-to-end optimization of goal-driven and visually grounded dialogue systems Harm de Vries

6.3. Squeak

Sequential sampling for kernel matrix approximation
KEYWORD: Machine learning

- Contact: Daniele Calandriello
- URL: http://researchers.lille.inria.fr/~valko/hp/serve.php?what=publications/squeak.py

6.4. OOR

Optimistic Optimization in R
KEYWORDS: Black-box optimization - Machine learning

- Contact: Mickael Binois
- URL: https://cran.r-project.org/web/packages/OOR/index.html

7. New Results

7.1. Decision-making Under Uncertainty

7.1.1. Reinforcement Learning

Thompson Sampling for Linear-Quadratic Control Problems, [22]

We consider the exploration-exploitation tradeoff in linear quadratic (LQ) control problems, where the state dynamics is linear and the cost function is quadratic in states and controls. We analyze the regret of Thompson sampling (TS) (a.k.a. posterior-sampling for reinforcement learning) in the frequentist setting, i.e., when the parameters characterizing the LQ dynamics are fixed. Despite the empirical and theoretical success in a wide range of problems from multi-armed bandit to linear bandit, we show that when studying the frequentist regret TS in control problems, we need to trade-off the frequency of sampling optimistic parameters and the frequency of switches in the control policy. This results in an overall regret of $O(T^{2/3})$, which is significantly worse than the regret $O(\sqrt{T})$ achieved by the optimism-in-face-of-uncertainty algorithm in LQ control problems.

Exploration–Exploitation in MDPs with Options, [33]

While a large body of empirical results show that temporally-extended actions and options may significantly affect the learning performance of an agent, the theoretical understanding of how and when options can be beneficial in online reinforcement learning is relatively limited. In this paper, we derive an upper and lower bound on the regret of a variant of UCRL using options. While we first analyze the algorithm in the general case of semi-Markov decision processes (SMDPs), we show how these results can be translated to the specific case of MDPs with options and we illustrate simple scenarios in which the regret of learning with options can be provably much smaller than the regret suffered when learning with primitive actions.

Regret Minimization in MDPs with Options without Prior Knowledge, [34]

The option framework integrates temporal abstraction into the reinforcement learning model through the introduction of macro-actions (i.e., options). Recent works leveraged the mapping of Markov decision processes (MDPs) with options to semi-MDPs (SMDPs) and introduced SMDP-versions of exploration-exploitation algorithms (e.g., RMAX-SMDP and UCRL-SMDP) to analyze the impact of options on the learning performance. Nonetheless, the PAC-SMDP sample complexity of RMAX-SMDP can hardly be translated into equivalent PAC-MDP theoretical guarantees, while the regret analysis of UCRL-SMDP requires prior knowledge of the distributions of the cumulative reward and duration of each option, which are hardly available in practice. In this paper, we remove this limitation by combining the SMDP view together with the inner Markov structure of options into a novel algorithm whose regret performance matches UCRL-SMDP's up to an additive regret term. We show scenarios where this term is negligible and the advantage of temporal abstraction is preserved. We also report preliminary empirical results supporting the theoretical findings.

Is the Bellman Residual a Bad Proxy?, [36]

This paper aims at theoretically and empirically comparing two standard optimization criteria for Reinforcement Learning: i) maximization of the mean value and ii) minimization of the Bellman residual. For that purpose, we place ourselves in the framework of policy search algorithms, that are usually designed to maximize the mean value, and derive a method that minimizes the residual T * v π – v π 1,ν over policies. A theoretical analysis shows how good this proxy is to policy optimization , and notably that it is better than its value-based counterpart. We also propose experiments on randomly generated generic Markov decision processes, specifically designed for studying the influence of the involved concentrability coefficient. They show that the Bellman residual is generally a bad proxy to policy optimization and that directly maximizing the mean value is much better, despite the current lack of deep theoretical analysis. This might seem obvious, as directly addressing the problem of interest is usually better, but given the prevalence of (projected) Bellman residual minimization in value-based reinforcement learning, we believe that this question is worth to be considered.

Faut-il minimiser le résidu de Bellman ou maximiser la valeur moyenne ?, [56]

Transfer Reinforcement Learning with Shared Dynamics, [38]

This article addresses a particular Transfer Reinforcement Learning (RL) problem: when dynamics do not change from one task to another, and only the reward function does. Our method relies on two ideas, the first one is that transition samples obtained from a task can be reused to learn on any other task: an immediate reward estimator is learnt in a supervised fashion and for each sample, the reward entry is changed by its reward estimate. The second idea consists in adopting the optimism in the face of uncertainty principle and to use upper bound reward estimates. Our method is tested on a navigation task, under four Transfer RL experimental settings: with a known reward function, with strong and weak expert knowledge on the reward function, and with a completely unknown reward function. It is also evaluated in a Multi-Task RL experiment and compared with the state-of-the-art algorithms. Results reveal that this method constitutes a major improvement for transfer/multi-task problems that share dynamics.

7.1.2. Multi-arm Bandit Theory

Trading Off Rewards and Errors in Multi-armed Bandits, [31]

In multi-armed bandits, the most common objective is the maximization of the cumulative reward. Alternative settings include active exploration, where a learner tries to gain accurate estimates of the rewards of all arms. While these objectives are contrasting, in many scenarios it is desirable to trade off rewards and errors. For instance, in educational games the designer wants to gather generalizable knowledge about the behavior of the students and teaching strategies (small estimation errors) but, at the same time, the system needs to avoid giving a bad experience to the players, who may leave the system permanently (large reward). In this paper, we formalize this tradeoff and introduce the ForcingBalance algorithm whose performance is provably close to the best possible tradeoff strategy. Finally, we demonstrate on real-world educational data that ForcingBalance returns useful information about the arms without compromising the overall reward.

Online Influence Maximization Under Independent Cascade Model with Semi-bandit Feedback, [54]

We study the online influence maximization problem in social networks under the independent cascade model. Specifically, we aim to learn the set of " best influencers " in a social network online while repeatedly interacting with it. We address the challenges of (i) combinatorial action space, since the number of feasible influencer sets grows exponentially with the maximum number of influencers, and (ii) limited feedback, since only the influenced portion of the network is observed. Under a stochastic semi-bandit feedback, we propose and analyze IMLinUCB, a computationally efficient UCB-based algorithm. Our bounds on the cumulative regret are polynomial in all quantities of interest, achieve near-optimal dependence on the number of interactions and reflect the topology of the network and the activation probabilities of its edges, thereby giving insights on the problem complexity. To the best of our knowledge, these are the first such results. Our experiments show that in several representative graph topologies, the regret of IMLinUCB scales as suggested by our upper bounds. IMLinUCB permits linear generalization and thus is both statistically and

computationally suitable for large-scale problems. Our experiments also show that IMLinUCB with linear generalization can lead to low regret in real-world online influence maximization.

Boundary Crossing for General Exponential Families, [39]

We consider parametric exponential families of dimension K on the real line. We study a variant of boundary crossing probabilities coming from the multi-armed bandit literature, in the case when the real-valued distributions form an exponential family of dimension K. Formally, our result is a concentration inequality that bounds the probability that B ψ (θ n , θ) f (t/n)/n, where θ is the parameter of an unknown target distribution, θ n is the empirical parameter estimate built from n observations, ψ is the log-partition function of the exponential family and B ψ is the corresponding Bregman divergence. From the perspective of stochastic multi-armed bandits, we pay special attention to the case when the boundary function f is logarithmic, as it enables to analyze the regret of the state-of-the-art KL-ucb and KL-ucb+ strategies, whose analysis was left open in such generality. Indeed, previous results only hold for the case when K = 1, while we provide results for arbitrary finite dimension K, thus considerably extending the existing results. Perhaps surprisingly, we highlight that the proof techniques to achieve these strong results already existed three decades ago in the work of T.L. Lai, and were apparently forgotten in the bandit community. We provide a modern rewriting of these beautiful techniques that we believe are useful beyond the application to stochastic multi-armed bandits.

The Non-stationary Stochastic Multi-armed Bandit Problem, Robin, Féraud, Maillard [64] [2]

Linear Thompson Sampling Revisited, [21]

We derive an alternative proof for the regret of Thompson sampling (TS) in the stochastic linear bandit setting. While we obtain a regret bound of order $\widetilde{O}(d^{3/2}\sqrt{T})$ as in previous results, the proof sheds new light on the functioning of the TS. We leverage on the structure of the problem to show how the regret is related to the sensitivity (i.e., the gradient) of the objective function and how selecting optimal arms associated to *optimistic* parameters does control it. Thus we show that TS can be seen as a generic randomized algorithm where the sampling distribution is designed to have a fixed probability of being optimistic, at the cost of an additional $\sqrt{d}$ regret factor compared to a UCB-like approach. Furthermore, we show that our proof can be readily applied to regularized linear optimization and generalized linear model problems.

Active Learning for Accurate Estimation of Linear Models, [47]

We explore the sequential decision-making problem where the goal is to estimate a number of linear models uniformly well, given a shared budget of random contexts independently sampled from a known distribution. For each incoming context, the decision-maker selects one of the linear models and receives an observation that is corrupted by the unknown noise level of that model. We present Trace-UCB, an adaptive allocation algorithm that learns the models' noise levels while balancing contexts accordingly across them, and prove bounds for its simple regret in both expectation and high-probability. We extend the algorithm and its bounds to the high dimensional setting , where the number of linear models times the dimension of the contexts is more than the total budget of samples. Simulations with real data suggest that Trace-UCB is remarkably robust , outperforming a number of baselines even when its assumptions are violated.

Learning the Distribution with Largest Mean: Two Bandit Frameworks, [18]

Over the past few years, the multi-armed bandit model has become increasingly popular in the machine learning community, partly because of applications including online content optimization. This paper reviews two different sequential learning tasks that have been considered in the bandit literature ; they can be formulated as (sequentially) learning which distribution has the highest mean among a set of distributions, with some constraints on the learning process. For both of them (regret minimization and best arm identification) we present recent, asymptotically optimal algorithms. We compare the behaviors of the sampling rule of each algorithm as well as the complexity terms associated to each problem.

On Bayesian Index Policies for Sequential Resource Allocation, [19]

[2] This work has been done while OA. Maillard was at Inria Saclay, in the TAO team.

This paper is about index policies for minimizing (frequentist) regret in a stochastic multi-armed bandit model, inspired by a Bayesian view on the problem. Our main contribution is to prove that the Bayes-UCB algorithm, which relies on quantiles of posterior distributions, is asymptotically optimal when the reward distributions belong to a one-dimensional exponential family, for a large class of prior distributions. We also show that the Bayesian literature gives new insight on what kind of exploration rates could be used in frequentist, UCB-type algorithms. Indeed, approximations of the Bayesian optimal solution or the Finite Horizon Gittins indices provide a justification for the kl-UCB+ and kl-UCB-H+ algorithms, whose asymptotic optimality is also established.

Multi-Player Bandits Models Revisited, [59]

Multi-player Multi-Armed Bandits (MAB) have been extensively studied in the literature, motivated by applications to Cognitive Radio systems. Driven by such applications as well, we motivate the introduction of several levels of feedback for multi-player MAB algorithms. Most existing work assume that sensing information is available to the algorithm. Under this assumption, we improve the state-of-the-art lower bound for the regret of any decentralized algorithms and introduce two algorithms, RandTopM and MCTopM, that are shown to empirically outperform existing algorithms. Moreover, we provide strong theoretical guarantees for these algorithms, including a notion of asymptotic optimality in terms of the number of selections of bad arms. We then introduce a promising heuristic, called Selfish, that can operate without sensing information, which is crucial for emerging applications to Internet of Things networks. We investigate the empirical performance of this algorithm and provide some first theoretical elements for the understanding of its behavior.

Multi-Armed Bandit Learning in IoT Networks: Learning helps even in non-stationary settings, [57]

Setting up the future Internet of Things (IoT) networks will require to support more and more communicating devices. We prove that intelligent devices in unlicensed bands can use Multi-Armed Bandit (MAB) learning algorithms to improve resource exploitation. We evaluate the performance of two classical MAB learning algorithms, UCB1 and Thompson Sampling, to handle the decentralized decision-making of Spectrum Access, applied to IoT networks; as well as learning performance with a growing number of intelligent end-devices. We show that using learning algorithms does help to fit more devices in such networks, even when all end-devices are intelligent and are dynamically changing channel. In the studied scenario, stochastic MAB learning provides a up to 16% gain in term of successful transmission probabilities, and has near optimal performance even in non-stationary and non-i.i.d. settings with a majority of intelligent devices.

7.1.3. Nonparametric Statistics of Time Series

Efficient Tracking of a Growing Number of Experts, [41]

We consider a variation on the problem of prediction with expert advice, where new forecasters that were unknown until then may appear at each round. As often in prediction with expert advice, designing an algorithm that achieves near-optimal regret guarantees is straightforward, using aggregation of experts. However, when the comparison class is sufficiently rich, for instance when the best expert and the set of experts itself changes over time, such strategies naively require to maintain a prohibitive number of weights (typically exponential with the time horizon). By contrast, designing strategies that both achieve a near-optimal regret and maintain a reasonable number of weights is highly non-trivial. We consider three increasingly challenging objectives (simple regret, shifting regret and sparse shifting regret) that extend existing notions defined for a fixed expert ensemble; in each case, we design strategies that achieve tight regret bounds, adaptive to the parameters of the comparison class, while being computationally inexpensive. Moreover, our algorithms are anytime , agnostic to the number of incoming experts and completely parameter-free. Such remarkable results are made possible thanks to two simple but highly effective recipes: first the " abstention trick " that comes from the specialist framework and enables to handle the least challenging notions of regret, but is limited when addressing more sophisticated objectives. Second, the " muting trick " that we introduce to give more flexibility. We show how to combine these two tricks in order to handle the most challenging class of comparison strategies.

7.1.4. Stochastic Games

Monte-Carlo Tree Search by Best Arm Identification, [37]

Recent advances in bandit tools and techniques for sequential learning are steadily enabling new applications and are promising the resolution of a range of challenging related problems. We study the game tree search problem, where the goal is to quickly identify the optimal move in a given game tree by sequentially sampling its stochastic payoffs. We develop new algorithms for trees of arbitrary depth, that operate by summarizing all deeper levels of the tree into confidence intervals at depth one, and applying a best arm identification procedure at the root. We prove new sample complexity guarantees with a refined dependence on the problem instance. We show experimentally that our algorithms outperform existing elimination-based algorithms and match previous special-purpose methods for depth-two trees.

Learning Nash Equilibrium for General-Sum Markov Games from Batch Data, [46]

This paper addresses the problem of learning a Nash equilibrium in γ-discounted mul-tiplayer general-sum Markov Games (MGs) in a batch setting. As the number of players increases in MG, the agents may either collaborate or team apart to increase their final rewards. One solution to address this problem is to look for a Nash equilibrium. Although , several techniques were found for the subcase of two-player zero-sum MGs, those techniques fail to find a Nash equilibrium in general-sum Markov Games. In this paper, we introduce a new definition of-Nash equilibrium in MGs which grasps the strategy's quality for multiplayer games. We prove that minimizing the norm of two Bellman-like residuals implies to learn such an-Nash equilibrium. Then, we show that minimizing an empirical estimate of the L p norm of these Bellman-like residuals allows learning for general-sum games within the batch setting. Finally, we introduce a neural network architecture that successfully learns a Nash equilibrium in generic multiplayer general-sum turn-based MGs.

7.1.5. Automata Learning

Spectral Learning from a Single Trajectory under Finite-State Policies, [23]

We present spectral methods of moments for learning sequential models from a single trajectory, in stark contrast with the classical literature that assumes the availability of multiple i.i.d. trajectories. Our approach leverages an efficient SVD-based learning algorithm for weighted automata and provides the first rigorous analysis for learning many important models using dependent data. We state and analyze the algorithm under three increasingly difficult scenarios: probabilistic automata, stochastic weighted automata, and reactive predictive state representations controlled by a finite-state policy. Our proofs include novel tools for studying mixing properties of stochastic weighted automata.

7.1.6. Online Kernel and Graph-Based Methods

Distributed Adaptive Sampling for Kernel Matrix Approximation, [26]

Most kernel-based methods, such as kernel regression, kernel PCA, ICA, or k-means clustering, do not scale to large datasets, because constructing and storing the kernel matrix $\mathbf{K}_n$ requires at least $O(n^2)$ time and space for n samples. Recent works (Alaoui 2014, Musco 2016) show that sampling points with replacement according to their ridge leverage scores (RLS) generates small dictionaries of relevant points with strong spectral approximation guarantees for $\mathbf{K}_n$. The drawback of RLS-based methods is that computing exact RLS requires constructing and storing the whole kernel matrix. In this paper, we introduce SQUEAK, a new algorithm for kernel approximation based on RLS sampling that sequentially processes the dataset, storing a dictionary which creates accurate kernel matrix approximations with a number of points that only depends on the effective dimension $d_{\text{eff}}(\gamma)$ of the dataset. Moreover since all the RLS estimations are efficiently performed using only the small dictionary, SQUEAK never constructs the whole matrix $\mathbf{K}_n$ runs in linear time $\widetilde{O}(nd_{\text{eff}}(\gamma)^3)$ w.r.t. n, and requires only a single pass over the dataset. We also propose a parallel and distributed version of SQUEAK achieving similar accuracy in as little as $\widetilde{O}(log(n)d_{\text{eff}}(\gamma)^3)$ time.

Second-Order Kernel Online Convex Optimization with Adaptive Sketching, [28]

Kernel online convex optimization (KOCO) is a framework combining the expressiveness of non-parametric kernel models with the regret guarantees of online learning. First-order KOCO methods such as functional gradient descent require only $O(t)$ time and space per iteration, and, when the only information on the losses is their convexity, achieve a minimax optimal $O(\sqrt{T})$ regret. Nonetheless, many common losses in kernel problems, such as squared loss, logistic loss, and squared hinge loss posses stronger curvature that can be exploited. In this case, second-order KOCO methods achieve $O(\log(Det(K)))$ regret, which we show scales as $O(d_{eff}\log T)$, where d_{eff} is the effective dimension of the problem and is usually much smaller than $O(\sqrt{T})$. The main drawback of second-order methods is their much higher $O(t^2)$ space and time complexity. In this paper, we introduce kernel online Newton step (KONS), a new second-order KOCO method that also achieves $O(d_{eff}\log T)$ regret. To address the computational complexity of second-order methods, we introduce a new matrix sketching algorithm for the kernel matrix K, and show that for a chosen parameter $\gamma \leq 1$ our Sketched-KONS reduces the space and time complexity by a factor of γ^2 to $O(t^2\gamma^2)$ space and time per iteration, while incurring only $1/\gamma$ times more regret.

Efficient Second-order Online Kernel Learning with Adaptive Embedding, [27]

Online kernel learning (OKL) is a flexible framework to approach prediction problems, since the large approximation space provided by reproducing kernel Hilbert spaces can contain an accurate function for the problem. Nonetheless, optimizing over this space is computationally expensive. Not only first order methods accumulate $O(\sqrt{T})$ more loss than the optimal function, but the curse of kernelization results in a $O(t)$ per step complexity. Second-order methods get closer to the optimum much faster, suffering only $O(\log(T))$ regret, but second-order updates are even more expensive, with a $O(t^2)$ per-step cost. Existing approximate OKL methods try to reduce this complexity either by limiting the Support Vectors (SV) introduced in the predictor, or by avoiding the kernelization process altogether using embedding. Nonetheless, as long as the size of the approximation space or the number of SV does not grow over time, an adversary can always exploit the approximation process. In this paper, we propose PROS-N-KONS, a method that combines Nystrom sketching to project the input point in a small, accurate embedded space, and performs efficient second-order updates in this space. The embedded space is continuously updated to guarantee that the embedding remains accurate, and we show that the per-step cost only grows with the effective dimension of the problem and not with T. Moreover, the second-order updated allows us to achieve the logarithmic regret. We empirically compare our algorithm on recent large-scales benchmarks and show it performs favorably.

Zonotope Hit-and-run for Efficient Sampling from Projection DPPs, [35]

Determinantal point processes (DPPs) are distributions over sets of items that model diversity using kernels. Their applications in machine learning include summary extraction and recommendation systems. Yet, the cost of sampling from a DPP is prohibitive in large-scale applications, which has triggered an effort towards efficient approximate samplers. We build a novel MCMC sampler that combines ideas from combinatorial geometry, linear programming, and Monte Carlo methods to sample from DPPs with a fixed sample cardinality, also called projection DPPs. Our sampler leverages the ability of the hit-and-run MCMC kernel to efficiently move across convex bodies. Previous theoretical results yield a fast mixing time of our chain when targeting a distribution that is close to a projection DPP, but not a DPP in general. Our empirical results demonstrate that this extends to sampling projection DPPs, i.e., our sampler is more sample-efficient than previous approaches which in turn translates to faster convergence when dealing with costly-to-evaluate functions, such as summary extraction in our experiments.

7.2. Statistical Learning and Bayesian Analysis

Universality of Bayesian mixture predictors, [50]

The problem is that of sequential probability forecasting for finite-valued time series. The data is generated by an unknown probability distribution over the space of all one-way infinite sequences. It is known that this measure belongs to a given set C, but the latter is completely arbitrary (uncountably infinite, without any structure given). The performance is measured with asymptotic average log loss. In this work it is shown that the minimax asymptotic performance is always attainable, and it is attained by a convex combination of a

countably many measures from the set C (a Bayesian mixture). This was previously only known for the case when the best achievable asymptotic error is 0. This also contrasts previous results that show that in the non-realizable case all Bayesian mixtures may be suboptimal, while there is a predictor that achieves the optimal performance.

Hypotheses Testing on Infinite Random Graphs, [48]

Drawing on some recent results that provide the formalism necessary to definite stationarity for infinite random graphs, this paper initiates the study of statistical and learning questions pertaining to these objects. Specifically, a criterion for the existence of a consistent test for complex hypotheses is presented, generalizing the corresponding results on time series. As an application, it is shown how one can test that a tree has the Markov property, or,more generally, to estimate its memory.

Independence Clustering (Without a Matrix), [49]

The independence clustering problem is considered in the following formulation: given a set S of random variables, it is required to find the finest partitioning $\{U\_1, \cdots, U\_k\}$ of S into clusters such that the clusters $U\_1, \cdots, U\_k$ are mutually independent. Since mutual independence is the target, pairwise similarity measurements are of no use, and thus traditional clustering algorithms are inapplicable. The distribution of the random variables in S is, in general, unknown, but a sample is available. Thus, the problem is cast in terms of time series. Two forms of sampling are considered: i.i.d. and stationary time series, with the main emphasis being on the latter, more general, case. A consistent, computationally tractable algorithm for each of the settings is proposed, and a number of open directions for further research are outlined.

7.3. Applications

7.3.1. Dialogue Systems and Natural Language

End-to-end Optimization of Goal-driven and Visually Grounded Dialogue Systems, [51]

End-to-end design of dialogue systems has recently become a popular research topic thanks to powerful tools such as encoder-decoder architectures for sequence-to-sequence learning. Yet, most current approaches cast human-machine dialogue management as a supervised learning problem, aiming at predicting the next utterance of a participant given the full history of the dialogue. This vision is too simplistic to render the intrinsic planning problem inherent to dialogue as well as its grounded nature , making the context of a dialogue larger than the sole history. This is why only chitchat and question answering tasks have been addressed so far using end-to-end architectures. In this paper, we introduce a Deep Reinforcement Learning method to optimize visually grounded task-oriented dialogues , based on the policy gradient algorithm. This approach is tested on a dataset of 120k dialogues collected through Mechanical Turk and provides encouraging results at solving both the problem of generating natural dialogues and the task of discovering a specific object in a complex picture.

Online Learning and Transfer for User Adaptation in Dialogue Systems, [58]

We address the problem of user adaptation in Spoken Dialogue Systems. The goal is to quickly adapt online to a new user given a large amount of dialogues collected with other users. Previous works using Transfer for Reinforcement Learning tackled this problem when the number of source users remains limited. In this paper, we overcome this constraint by clustering the source users: each user cluster, represented by its centroid, is used as a potential source in the state-of-the-art Transfer Reinforcement Learning algorithm. Our benchmark compares several clustering approaches , including one based on a novel metric. All experiments are led on a negotiation dialogue task, and their results show significant improvements over baselines.

GuessWhat?! Visual Object Discovery Through Multi-modal Dialogue, [29]

We introduce GuessWhat?!, a two-player guessing game as a testbed for research on the interplay of computer vision and dialogue systems. The goal of the game is to locate an unknown object in a rich image scene by asking a sequence of questions. Higher-level image understanding, like spatial reasoning and language grounding, is required to solve the proposed task. Our key contribution is the collection of a large-scale dataset

consisting of 150K human-played games with a total of 800K visual question-answer pairs on 66K images. We explain our design decisions in collecting the dataset and introduce the oracle and questioner tasks that are associated with the two players of the game. We prototyped deep learning models to establish initial base-lines of the introduced tasks.

LIG-CRIStAL System for the WMT17 Automatic Post-Editing Task, [25]

This paper presents the LIG-CRIStAL submission to the shared Automatic Post-Editing task of WMT 2017. We propose two neural post-editing models: a mono-source model with a task-specific attention mechanism, which performs particularly well in a low-resource scenario; and a chained architecture which makes use of the source sentence to provide extra context. This latter architecture manages to slightly improve our results when more training data is available. We present and discuss our results on two datasets (en-de and de-en) that are made available for the task.

7.3.2. Recommendation systems

A Multi-Armed Bandit Model Selection for Cold-Start User Recommendation, [32]

How can we effectively recommend items to a user about whom we have no information? This is the problem we focus on, known as the cold-start problem. In this paper, we focus on the cold user problem.In most existing works, the cold-start problem is handled through the use of many kinds of information available about the user. However, what happens if we do not have any information?Recommender systems usually keep a substantial amount of prediction models that are available for analysis. Moreover, recommendations to new users yield uncertain returns. Assuming a number of alternative prediction models is available to select items to recommend to a cold user, this paper introduces a multi-armed bandit based model selection, named PdMS.In comparison with two baselines, PdMS improves the performance as measured by the nDCG.These improvements are demonstrated on real, public datasets.

7.3.3. Software development

A Large-scale Study of Call Graph-based Impact Prediction using Mutation Testing, [20]

In software engineering, impact analysis consists in predicting the software elements (e.g. modules, classes, methods) potentially impacted by a change in the source code. Impact analysis is required to optimize the testing effort. In this paper, we propose a framework to predict error propagation. Based on 10 open-source Java projects and 5 classical mutation operators, we create 17000 mutants and study how the error they introduce propagates. This framework enables us to analyze impact prediction based on four types of call graph. Our results show that the sophistication indeed increases completeness of impact prediction. However, and surprisingly to us, the most basic call graph gives the highest trade-off between precision and recall for impact prediction.

Correctness Attraction: A Study of Stability of Software Behavior under Runtime Perturbation, [15]

Can the execution of a software be perturbed without breaking the correctness of the output? In this paper, we devise a novel protocol to answer this rarely investigated question. In an experimental study, we observe that many perturbations do not break the correctness in ten subject programs. We call this phenomenon “correctness attraction”. The uniqueness of this protocol is that it considers a systematic exploration of the perturbation space as well as perfect oracles to determine the correctness of the output. To this extent, our findings on the stability of software under execution perturbations have a level of validity that has never been reported before in the scarce related work. A qualitative manual analysis enables us to set up the first taxonomy ever of the reasons behind correctness attraction.

7.3.4. Graph theory

A generative model for sparse, evolving digraphs, [43]

Generating graphs that are similar to real ones is an open problem, while the similarity notion is quite elusive and hard to formalize. In this paper, we focus on sparse digraphs and propose SDG, an algorithm that aims at generating graphs similar to real ones. Since real graphs are evolving and this evolution is important to study in order to understand the underlying dynamical system, we tackle the problem of generating series of graphs. We propose SEDGE, an algorithm meant to generate series of graphs similar to a real series. SEDGE is an extension of SDG. We consider graphs that are representations of software programs and show experimentally that our approach outperforms other existing approaches. Experiments show the performance of both algorithms.

A Spectral Algorithm with Additive Clustering for the Recovery of Overlapping Communities in Networks, [17]

This paper presents a novel spectral algorithm with additive clustering designed to identify overlapping communities in networks. The algorithm is based on geometric properties of the spectrum of the expected adjacency matrix in a random graph model that we call stochastic blockmodel with overlap (SBMO). An adaptive version of the algorithm, that does not require the knowledge of the number of hidden communities, is proved to be consistent under the SBMO when the degrees in the graph are (slightly more than) logarithmic. The algorithm is shown to perform well on simulated data and on real-world graphs with known overlapping communities.

7.3.5. Deep Learning

Modulating early visual processing by language, [30]

It is commonly assumed that language refers to high-level visual concepts while leaving low-level visual processing unaffected. This view dominates the current literature in computational models for language-vision tasks, where visual and linguistic inputs are mostly processed independently before being fused into a single representation. In this paper, we deviate from this classic pipeline and propose to modulate the entire visual processing by a linguistic input. Specifically, we introduce Conditional Batch Normalization (CBN) as an efficient mechanism to modulate convolutional feature maps by a linguistic embedding. We apply CBN to a pre-trained Residual Network (ResNet), leading to the MODulatEd ResNet (MODERN) architecture, and show that this significantly improves strong baselines on two visual question answering tasks. Our ablation study confirms that modulating from the early stages of the visual processing is beneficial.

FiLM: Visual Reasoning with a General Conditioning Layer, [45]

We introduce a general-purpose conditioning method for neural networks called FiLM: Feature-wise Linear Modulation. FiLM layers influence neural network computation via a simple , feature-wise affine transformation based on conditioning information. We show that FiLM layers are highly effective for visual reasoning - answering image-related questions which require a multi-step, high-level process - a task which has proven difficult for standard deep learning methods that do not explicitly model reasoning. Specifically, we show on visual reasoning tasks that FiLM layers 1) halve state-of-the-art error for the CLEVR benchmark, 2) modulate features in a coherent manner, 3) are robust to ablations and architectural modifications, and 4) generalize well to challenging, new data from few examples or even zero-shot.

Learning Visual Reasoning Without Strong Priors, [44]

Achieving artificial visual reasoning - the ability to answer image-related questions which require a multi-step, high-level process - is an important step towards artificial general intelligence. This multi-modal task requires learning a question-dependent, structured reasoning process over images from language. Standard deep learning approaches tend to exploit biases in the data rather than learn this underlying structure, while leading methods learn to visually reason successfully but are hand-crafted for reasoning. We show that a general-purpose, Conditional Batch Normalization approach achieves state-of-the-art results on the CLEVR Visual Reasoning benchmark with a 2.4% error rate. We outperform the next best end-to-end method (4.5%) and even methods that use extra supervision (3.1%). We probe our model to shed light on how it reasons, showing it has learned a question-dependent, multi-step process. Previous work has operated under the assumption that visual reasoning calls for a specialized architecture, but we show that a general architecture

with proper conditioning can learn to visually reason effectively. Index Terms: Deep Learning, Language and Vision Note: A full paper extending this study is available at http: //arxiv.org/abs/1709.07871, with additional references , experiments, and analysis.

HoME: a Household Multimodal Environment, [24]

We introduce HoME: a Household Multimodal Environment for artificial agents to learn from vision, audio, semantics, physics, and interaction with objects and other agents, all within a realistic context. HoME integrates over 45,000 diverse 3D house layouts based on the SUNCG dataset, a scale which may facilitate learning, generalization, and transfer. HoME is an open-source, OpenAI Gym-compatible platform extensible to tasks in reinforcement learning, language grounding, sound-based navigation, robotics, multi-agent learning, and more. We hope HoME better enables artificial agents to learn as humans do: in an interactive, multimodal, and richly contextualized setting.

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: Mar. 2018 – Feb. 2021

  Abstract: 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, Alessandro Lazaric, Michal Valko.

8.1.2. OtherLang

- contract with “OtherLang”; PI: Romaric Gaudel

  Title: Tool to support foreign language practice

  Duration: 2 months

  Abstract: OtherLang develops an application to learn a foreign language by reading documents and interacting wit other people. During the time-line of the contract, SequeL brought his knowledge about Recommender Systems which may be used either to recommend documents to users or to recommend users to users.
  Participants: Romaric Gaudel, Philippe Preux.

8.1.3. Sidexa

- contract with “Sidexa”; PI: Jérémie Mary and then Philippe Preux

  Title: vision applied to the segmentation and recognition of car body parts parts

  Duration: 3 months

  Abstract: We investigate deep learning to perform car body segmentation. The result being very good, a second contract will follow up this one in 2018.
  Participants: Jérémie Mary, Philippe Preux.

8.1.4. Renault

- contract with “Renault”; PI: Philippe Preux

  Title: State of the art in reinforcement learning regarding autonomous car control and path planning.

  Duration: 3 months (Jan–Mar 2017)

  Abstract: This work has consisted in surveying the litterature related to autonomous car control, and reinforcement learning.
  Participants: Alexis Martin, Odalric Maillard, Philippe Preux.

- contract with Renault; PI: Philippe Preux

  Title: Control of an autonomous vehicle

  Duration: 3 years (12/2017–11/2020)

  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 Maillard, Philippe Preux.

8.1.5. Critéo

- contract with “Criteo”; PI: Philippe Preux

  Title: Computational advertizing

  Duration: 3 years (12/2017–11/2020)

  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.
  Participants: Philippe Preux, Kiewan Villatel.

8.1.6. Orange Labs

- contract with “Orange Labs”; PI: Philippe Preux

  Title: Sequential Learning and Decision Making under Partial Monitoring

  Duration: Oct. 2014 – Sep. 2017

  Abstract: This contract comes along the CIFRE grant on the same topic. In applications such as recommendation systems, or computational advertising, the return collected from the user is partial: (s)he clicks on one item, or no item at all. We study this setting in which only a “partial” information is gathered in particular how to learn to behave optimaly in such a setting.
  Participants: Pratik Gajane, Philippe Preux.

8.1.7. Orange Labs

- contract with “Orange Labs”; PI: Olivier Pietquin

  Title: Inter User Transfer in dialogue systems

  Duration: 3 years

  Abstract: This contract comes along the CIFRE grant on the same topic. The research aims at developing new algorithms to learn fast adaptation strategies for dialogue systems when a new user starts using them while we collected data from previous interactions with other users. Especially, it addresses the cold-start problem encountered when a new user faces the system, before samples can be collected to optimize the interaction strategy.
  Participants: Merwan Barlier, Nicolas Carrara, Olivier Pietquin.

8.1.8. 55

- contract with “55”; PI: Jérémie Mary

  Title: Novel Learning and Exploration-Exploitation Methods for Effective Recommender Systems

  Duration: Oct. 2015 – Sep. 2018

  Abstract: This contract comes along the CIFRE grant on the same topic. In this Ph.D. thesis we intend to deal with this problem by developing novel and more sophisticated recommendation strategies in which the collection of data and the improvement of the performance are considered as a unique process, where the trade-off between the quality of the data and the performance of the recommendation strategy is optimized over time. This work also consider tensor methods (one layer of the tensor can be the time) with the goal to scale them at RS level.

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. ANR BoB

Participants: Rémi Bardenet, 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 favourably 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.

9.1.2. ANR Badass

Participants: Odalric 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.

9.1.3. ANR ExTra-Learn

Participants: Alessandro Lazaric, Jérémie Mary, Michal Valko.

- *Title*: Extraction and Transfer of Knowledge in Reinforcement Learning
- *Type*: National Research Agency (ANR-9011)
- *Coordinator*: Inria Lille (A. Lazaric)
- *Duration*: 2014-2018
- *Abstract*: ExTra-Learn is directly motivated by the evidence that one of the key features that allows humans to accomplish complicated tasks is their ability of building knowledge from past experience and transfer it while learning new tasks. We believe that integrating transfer of learning in machine learning algorithms will dramatically improve their learning performance and enable them to solve complex tasks. We identify in the reinforcement learning (RL) framework the most suitable candidate for this integration. RL formalizes the problem of learning an optimal control policy from the experience directly collected from an unknown environment. Nonetheless, practical limitations of current algorithms encouraged research to focus on how to integrate prior knowledge

into the learning process. Although this improves the performance of RL algorithms, it dramatically reduces their autonomy. In this project we pursue a paradigm shift from designing RL algorithms incorporating prior knowledge, to methods able to incrementally discover, construct, and transfer "prior" knowledge in a fully automatic way. More in detail, three main elements of RL algorithms would significantly benefit from transfer of knowledge. *(i)* For every new task, RL algorithms need exploring the environment for a long time, and this corresponds to slow learning processes for large environments. Transfer learning would enable RL algorithms to dramatically reduce the exploration of each new task by exploiting its resemblance with tasks solved in the past. *(ii)* RL algorithms evaluate the quality of a policy by computing its state-value function. Whenever the number of states is too large, approximation is needed. Since approximation may cause instability, designing suitable approximation schemes is particularly critical. While this is currently done by a domain expert, we propose to perform this step automatically by constructing features that incrementally adapt to the tasks encountered over time. This would significantly reduce human supervision and increase the accuracy and stability of RL algorithms across different tasks. *(iii)* In order to deal with complex environments, hierarchical RL solutions have been proposed, where state representations and policies are organized over a hierarchy of subtasks. This requires a careful definition of the hierarchy, which, if not properly constructed, may lead to very poor learning performance. The ambitious goal of transfer learning is to automatically construct a hierarchy of skills, which can be effectively reused over a wide range of similar tasks.

- *Activity Report*: Research in ExTra-Learn continued in investigating how knowledge can be transferred into reinforcement learning algorithms to improve their performance. Pierre-Victor Chaumier did a 4 months internship in SequeL studying how to perform transfer neural networks across different games in the Atari platform. Unfortunately, the preliminary results we obtained were not very positive. We investigated different transfer models, from basic transfer of a fully trained network, to co-train over multiple games and retrain with initialization from a previous network. In most of the cases, the improvement from transfer was rather limited and in some cases even negative transfer effects appeared. This seems to be intrinsic in the neural network architecture which tends to overfit on one single task and it poorly generlizes over alternative tasks. Another activity was related to the study of macro-actions in RL. We proved for the first time under which conditions macro-actions can actually improve the learning speed of an RL exploration-exploitation algorithm. This is the first step towards the automatic identification and construction of useful macro-actions across multiple tasks.

9.1.4. ANR KEHATH

Participants: Olivier Pietquin, Alexandre Bérard.

- *Acronym*: KEHATH
- *Title*: Advanced Quality Methods for Post-Edition of Machine Translation
- *Type*: ANR
- *Coordinator*: Lingua & Machina
- *Duration*: 2014-2017
- *Other partners*: Univ. Lille 1, Laboratoire d'Informatique de Grenoble (LIG)
- *Abstract*: The translation community has seen a major change over the last five years. Thanks to progress in the training of statistical machine translation engines on corpora of existing translations, machine translation has become good enough so that it has become advantageous for translators to post-edit machine outputs rather than translate from scratch. However, current enhancement of machine translation (MT) systems from human post-edition (PE) are rather basic: the post-edited output is added to the training corpus and the translation model and language model are re-trained, with no clear view of how much has been improved and how much is left to be improved. Moreover, the final PE result is the only feedback used: available technologies do not take advantages of logged sequences of post-edition actions, which inform on the cognitive processes of the post-editor. The

KEHATH project intends to address these issues in two ways. Firstly, we will optimise advanced machine learning techniques in the MT+PE loop. Our goal is to boost the impact of PE, that is, reach the same performance with less PE or better performance with the same amount of PE. In other words, we want to improve machine translation learning curves. For this purpose, active learning and reinforcement learning techniques will be proposed and evaluated. Along with this, we will have to face challenges such as MT systems heterogeneity (statistical and/or rule-based), and ML scalability so as to improve domain-specific MT. Secondly, since quality prediction (QP) on MT outputs is crucial for translation project managers, we will implement and evaluate in real-world conditions several confidence estimation and error detection techniques previously developed at a laboratory scale. A shared concern will be to work on continuous domain-specific data flows to improve both MT and the performance of indicators for quality prediction. The overall goal of the KEHATH project is straightforward: gain additional machine translation performance as fast as possible in each and every new industrial translation project, so that post-edition time and cost is drastically reduced. Basic research is the best way to reach this goal, for an industrial impact that is powerful and immediate.

9.1.5. PEPS Project BIO

Participants: Émilie Kaufmann, Lilian Besson.

- *Title*: Bandits pour l'Internet des Objets
- *Type*: CNRS PEPS project
- *Coordinator*: CNRS (E. Kaufmann)
- *Duration*: april-december 2017
- *Abstract*: (in French) Dans le but d'améliorer le qualité et de minimiser les coûts énergétiques des communications entre les objets communicants et leurs stations de base, nous cherchons dans ce projet à adapter les avancées récentes du domaine de la radio intelligente à la spécificité des communications de type Internet des Objets. Vu l'engorgement du spectre fréquentiel, il est nécessaire pour ces objets d'apprendre à détecter de manière adaptative quand et sur quelle fréquence communiquer. Nous proposons pour cette tâche l'utilisation d'algorithmes dits de bandit à plusieurs bras, déjà connus dans le contexte de la radio intelligente, mais pas toujours adaptés à la spécificité des communications pour l'Internet des Objets. Nous introduirons de nouveaux algorithmes de bandit multi-joueurs, traduisant la coordination nécessaire entre les multiples objets en plus de l'apprentissage de la qualité des canaux fréquentiel. Ensuite nous envisagerons une nouvelle modélisation, de type bandit adversarial, pour décrire les communications dans des standards comme LoRa où les objets reçoivent des messages de confirmation des stations de bases, conduisant à des algorithmes minimisant la latence de ces communications.

9.1.6. National Partners

- ENS Paris-Saclay
  - M. Valko collaborated with V. Perchet on structured bandit problem. They co-supervise a PhD student (P. Perrault) together.
- Institut de Mathématiques de Toulouse
  - E. Kaufmann collaborated with Aurélien Garivier on sequential testing and structured bandit problems.
- CentraleSupélec Rennes
  - E. Kaufmann co-advises Lilian Besson, who works at CentraleSupélec with Christophe Moy. Christophe, Lilian and Émilie worked together on a PEPS project about bandits for Internet Of Things. One paper was published to the CROWNCOM conference, and another has been submitted to the ALT conference.

9.2. European Initiatives

9.2.1. FP7 & H2020 Projects

9.2.1.1. H2020 BabyRobot

Program: H2020

Project acronym: BabyRobot

Project title: Child-Robot Communication and Collaboration

Duration: 01/2016 - 12/2018

Coordinator: Alexandros Potamianos (Athena Research and Innovation Center in Information Communication and Knowledge Technologies, Greece)

Other partners: Institute of Communication and Computer Systems (Greece), The University of Hertfordshire Higher Education Corporation (UK), Universitaet Bielefeld (Germany), Kunlgliga Tekniska Hoegskolan (Sweden), Blue Ocean Robotics ApS (Denmark), Univ. Lille (France), Furhat Robotics AB (Sweden)

Abstract: The crowning achievement of human communication is our unique ability to share intentionality, create and execute on joint plans. Using this paradigm we model human-robot communication as a three step process: sharing attention, establishing common ground and forming shared goals. Prerequisites for successful communication are being able to decode the cognitive state of people around us (mindreading) and building trust. Our main goal is to create robots that analyze and track human behavior over time in the context of their surroundings (situational) using audio-visual monitoring in order to establish common ground and mind-reading capabilities. On BabyRobot we focus on the typically developing and autistic spectrum children user population. Children have unique communication skills, are quick and adaptive learners, eager to embrace new robotic technologies. This is especially relevant for special eduation where the development of social skills is delayed or never fully develops without intervention or therapy. Thus our second goal is to define, implement and evaluate child-robot interaction application scenarios for developing specific socio-affective, communication and collaboration skills in typically developing and autistic spectrum children. We will support not supplant the therapist or educator, working hand-inhand to create a low risk environment for learning and cognitive development. Breakthroughs in core robotic technologies are needed to support this research mainly in the areas of motion planning and control in constrained spaces, gestural kinematics, sensorimotor learning and adaptation. Our third goal is to push beyond the state-of-the-art in core robotic technologies to support natural human-robot interaction and collaboration for edutainment and healthcare applications. Creating robots that can establish communication protocols and form collaboration plans on the fly will have impact beyond the application scenarios investigated here.

9.2.1.2. CHIST-ERA DELTA

Participants: Michal Valko, Émilie Kaufmann.

Program: CHIST-ERA

Project acronym: DELTA

Project title: Dynamically Evolving Long-Term Autonomy

Duration: October 2017 - December 2021

Coordinator: Anders Jonsson (PI)

Inria coPI: 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.

9.2.1.3. CHIST-ERA IGLU

Program: CHIST-ERA

Project acronym: IGLU

Project title: Interactively Grounded Language Understanding

Duration: 11/2015 - 10/2018

Coordinator: Jean Rouat (Université de Sherbrooke, Canada)

Other partners: UMONS (Belgique), Inria (France), Univ-Lille (France), KTH (sweden), Universidad de Zaragoza (Spain)

Abstract: Language is an ability that develops in young children through joint interaction with their caretakers and their physical environment. At this level, human language understanding could be referred as interpreting and expressing semantic concepts (e.g. objects, actions and relations) through what can be perceived (or inferred) from current context in the environment. Previous work in the field of artificial intelligence has failed to address the acquisition of such perceptually-grounded knowledge in virtual agents (avatars), mainly because of the lack of physical embodiment (ability to interact physically) and dialogue, communication skills (ability to interact verbally). We believe that robotic agents are more appropriate for this task, and that interaction is a so important aspect of human language learning and understanding that pragmatic knowledge (identifying or conveying intention) must be present to complement semantic knowledge. Through a developmental approach where knowledge grows in complexity while driven by multimodal experience and language interaction with a human, we propose an agent that will incorporate models of dialogues, human emotions and intentions as part of its decision-making process. This will lead anticipation and reaction not only based on its internal state (own goal and intention, perception of the environment), but also on the perceived state and intention of the human interactant. This will be possible through the development of advanced machine learning methods (combining developmental, deep and reinforcement learning) to handle large-scale multimodal inputs, besides leveraging state-of-the-art technological components involved in a language-based dialog system available within the consortium. Evaluations of learned skills and knowledge will be performed using an integrated architecture in a culinary use-case, and novel databases enabling research in grounded human language understanding will be released.

9.3. International Initiatives

9.3.1. With CWI

Title: Non-parametric sequential prediction project

Centrum Wiskunde & Informatica (CWI), Amsterdam (NL) - Peter Grünwald

Duration: 2016 - 2018

Start year: 2016

Abstract: The aim is to develop the theory of learning for sequential decision making under uncertainty problems.

In 2017, this collaboration involved D. Ryabko, É. Kaufmann, J. Ridgway, M. Valko, O. Maillard. A post-doc funded by Inria has been recruited in Fall 2016.

https://project.inria.fr/inriacwi/projects/non-parametric-sequential-prediction-project/

9.3.2. EduBand

Title: Educational Bandits

International Partner (Institution - Laboratory - Researcher):

Carnegie Mellon University (United States) - Department of Computer Science, Theory of computation lab - Emma Brunskill

Start year: 2015

See also: https://project.inria.fr/eduband/

Education can transform an individual's capacity and the opportunities available to him. The proposed collaboration will build on and develop novel machine learning approaches towards enhancing (human) learning. Massive open online classes (MOOCs) are enabling many more people to access education, but mostly operate using status quo teaching methods. Even more important than access is the opportunity for online software to radically improve the efficiency, engagement and effectiveness of education. Existing intelligent tutoring systems (ITSs) have had some promising successes, but mostly rely on learning sciences research to construct hand-built strategies for automated teaching. Online systems make it possible to actively collect substantial amount of data about how people learn, and offer a huge opportunity to substantially accelerate progress in improving education. An essential aspect of teaching is providing the right learning experience for the student, but it is often unknown a priori exactly how this should be achieved. This challenge can often be cast as an instance of decision-making under uncertainty. In particular, prior work by Brunskill and colleagues demonstrated that reinforcement learning (RL) and multi-arm bandit (MAB) can be very effective approaches to solve the problem of automated teaching. The proposed collaboration is thus intended to explore the potential interactions of the fields of online education and RL and MAB. On the one hand, we will define novel RL and MAB settings and problems in online education. On the other hand, we will investigate how solutions developed in RL and MAB could be integrated in ITS and MOOCs and improve their effectiveness.

9.3.3. Allocate

Participants: Pierre Perrault, Julien Seznec, Michal Valko, Émilie Kaufmann, Odalric Maillard.

Title: Adaptive allocation of resources for recommender systems

Inria contact: Michal Valko

International Partner (Institution - Laboratory - Researcher):

Univertät Potsdam, Germany A. Carpentier

Start year: 2017

We plan to improve a practical scenario of *resource allocation in market surveys*, such as product appraisals and music recommendation. In practice, the market is typically divided into segments: geographic regions, age groups, ...These groups are then queried for preference with some fixed rule of a number of queries per group. This testing is *costly and non-adaptive*. The reason is some groups are easier to estimate than others, but this is impossible to know a priori. Our challenge is **adaptively allocate the optimal number of samples** to each group and improve the efficient of market studies, by providing *sample-efficient* solutions.

9.3.4. Informal International Partners

Adobe Research

Branislav Kveton *Collaborator*

Zheng Wen *Collaborator*

Sharan Vaswani *Collaborator*

M. Valko collaborated with Adobe Research on online influence maximization in social networks. This led to a publication in NIPS 2017.

Massachusetts Institute of Technology

Victor-Emmanuel Brunel *Collaborator*

M. Valko collaborated with V.-E. Brunel on the estimation of low rank determinantal point processes useful for diverse recommender systems.

Univertät Potsdam

Alexandra Carpentier *Collaborator*

M. Valko collaborated with A. Carpentier on adaptive estimation of the block-diagonal matrices with application to market segmentations. This collaboration formalized in September 2017 by creating a north-european associate team.

University of California, Berkeley

Victor Gabillon *Collaborator*

M. Valko collaborated with V. Gabillon on the sample complexities in unknown type of environments.

University of Southern California

Haipeng Luo *Collaborator*

M. Valko collaborated with H. Luo on online submodular minimization.

Adobe Research

Mohammad Ghavamzadeh *Collaborator*

A. Lazaric collaborated with Adobe Research on active learning for accurate estimation of linear models. This led to a publication in ICML 2017.

Stanford University

Carlos Riquelme *Collaborator*

A. Lazaric collaborated with Carlos Riquelme on active learning for accurate estimation of linear models. This led to a publication in ICML 2017.

Stanford University

Emma Brunskill *Collaborator*

A. Lazaric collaborated with Emma Brunskill on exploration-exploitation with options in reinforcement learning. This led to a publication in NIPS 2017.

University of California, Irvine

Anima Anandkumar *Collaborator*

Kamyar Azzizade *Collaborator*

A. Lazaric collaborated with A. Anandkumar and K. Azzizade on exploration-exploitation with in reinforcement learning with state clustering. This led to a submission to AI&Stats 2018.

University of Leoben

Ronald Ortner *Collaborator*

A. Lazaric collaborated with R. Ortner on exploration-exploitation in reinforcement learning with regularized optimization. This will lead to a submission to ICML 2018.

Politecnico di Milano

Marcello Restelli *Collaborator*

Matteo Pirotta collaborate with M. Restelli on several topics in reinforcement learning. This will lead to publications to ICML 2017 and NIPS 2017.

Lancaster University

B. Balle *Collaborator*

O. Maillard collaborated on spectral learning of Hankel matrices. This led to a publication at ICML.

Mila, Université de Montréal

A. Courville *Collaborator*

F. Strub and O. Pietquin collaborate on deep reinforcement learning for language acquisition. This led to several papers at IJCAI, CVPR, and NIPS, as well as the guesswhat?! dataset and protocol, and the HOME dataset.

Uberlandia University, Brasil

C. Felicio *Collaborator*

Ph. Preux supervises this PhD on recommendation systems. This led to the defense of C. Felicio and a paper at UMAP.

9.3.5. International Initiatives

SequeL

Title: The multi-armed bandit problem

International Partner (Institution - Laboratory - Researcher):

University of Leoben (Austria) Peter Auer

Duration: 2014 - 2018

Start year: 2014

In a nutshell, the collaboration is focusing on nonparametric algorithms for active learning problems, mainly involving theoretical analysis of reinforcement learning and bandits problems beyond the traditional settings of finite-state MDPs (for RL) or i.i.d. rewards (for bandits). Peter Auer from University of Leoben is a worldwide leader in the field, having introduced the UCB approach around 2000, along with its finite-time analysis. Today, SequeL is likely to be the largest research group working in this field in the world, enjoying worldwide recognition. SequeL and P. Auer's group have been collaborating for a couple of years now; they have co-authored papers, visited each other (sabbatical stay, post-doc), coorganized workshops; the STREP Complacs partially funds this very active collaboration.

9.3.6. International Initiatives

Contextual multi-armed bandits with hidden structure

Title: Contextual multi-armed bandits with hidden structure

International Partner (Institution - Laboratory - Researcher):

IISc Bangalore (India) – Aditya Gopalan

Duration: 2015 - 2017

Recent advances in Multi-Armed Bandit (MAB) theory have yielded key insights into, and driven the design of applications in, sequential decision making in stochastic dynamical systems. Notable among these are recommender systems, which have benefited greatly from the study of contextual MABs incorporating user-specific information (the context) into the decision problem from a rigorous theoretical standpoint. In the proposed initiative, the key features of (a) sequential interaction between a learner and the users, and (b) a relatively small number of interactions per user with the system, motivate the goal of efficiently exploiting the underlying collective structure of users. The state-of-the-art lacks a wellgrounded strategy with provably near-optimal guarantees for general, low-rank user structure. Combining expertise in the foundations of MAB theory together with recent advances in spectral methods and low-rank matrix completion, we target the first provably near-optimal sequential low-rank MAB

9.4. International Research Visitors

9.4.1. Visits of International Scientists

9.4.1.1. Internships

- Harm de Vries, PhD stduent, University of Montreal, Canada, Jan-Jun 2017
- Mohammad Sadegh Talebi Mazraeh Shahi, PhD student, KTH Royal Institute of Technology, Sweden, Jun-Sep 2017
- Xuedong Shang, master student, ENS Rennes, Feb–Jun 2017
- Iuliia Olkhovskaia, master student, Moscow Institute of Physics and Technology, Russia, Feb–Jul 2017
- Georgios Papoudakis, master student, Aristotle University of Thessalnoniki, Greece, May–Sep 2017
- Subhojyoti Mukherjee, master student, Indian Institute of technology, Sep-Nov 2017
- Mahsa Asadi, Shiraz University, Iran, Sep-Dec 2017

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

- *Visually grounded interaction and language*, workshop at NIPS 2017, organized by Florian Strub, Harm de Vries, Abhishek Das, Satwik Kottur, Stefan Lee, Mateusz Malinowski, Olivier Pietquin, Devi Parikh, Dhruv Batra, Aaron C Courville, Jérémie Mary. URL: https://nips.cc/Conferences/2017/Schedule?showEvent=8766
- O. Maillard: Workshop of the working group *Sequential Structured Statistical Learning*, May 17 2017 at Institut des Hautes Etudes Scientifiques (Bures-sur-Yvette). URL: https://sites.google.com/site/groupedetravailsssl

10.1.1.1. Member of the Conference Program Committees

Members of SEQUEL have been involved in the following program committees in 2017:

- Senior PC for International Joint Conference on Artificial Intelligence (IJCAI 2017)
- Senior PC for ACM KDD 2017
- International Conference on Artificial Intelligence and Statistics (AI & STATS 2017)
- PC member for the international Conference On Learning Theory (COLT 2017)
- European Conference on Machine Learning (ECML 2017)
- 1st Workshop on Transfer in Reinforcement Learning (TiRL) 2017
- The Third International Conference on Machine Learning, Optimization and Big Data (MOD 2017)
- French conferences:
  - Extraction et Gestion de Conaissances (EGC),
  - Journées Francophones de Planification, Décision, Apprentissage (JFPDA)
  - Journées de la Société Francophone de Classification (SFC)
  - Conférence francophone sur l'Apprentissage Automatique (CAp)

10.1.1.2. Reviewer

Édouard Oyallon receives a "best NIPS reviewer award".

Members of SEQUEL have reviewed papers for the following conferences:

- AI&Stats, COLT, ECML, ICML, IJCAI, NIPS, ALT.

10.1.2. Journal

10.1.2.1. Reviewer - Reviewing Activities

- Automatica
- IEEE Transactions on Pattern Analysis and Machine Intelligence - Journal Reviewer
- IEEE transaction on Software Engineering
- International Federation of Automatic Control
- Bernoulli Journal
- Journal of Machine Learning Research
- IEEE Transaction on Signal Processing

10.1.3. Invited Talks

- R. Gaudel, *Recommendation as a Sequential Process*, Presented on Februaray 1st, 2017, at Séminaire CMLA, Paris, France (*CMLA 2017*)
- R. Gaudel, *Recommendation as a Sequential Process*, Presented on January 10th, 2017, at Séminaire ENSAI, Rennes (Bruz), France (*ENSAI 2017*)
- A. Lazaric, *Spectral Methods for Reinforcement Learning*, Presented on April 10, 2017, at Amazon, Berlin, Germany
- M. Valko, *SequeL, graphs in ML, and online recommender systems*, Presented on November 9th, 2017 at Plateau Inria Euratechnologies in Lille, France (*Euratechnologies 2017*)
- M. Valko, *Sequential sampling for kernel matrix approximation and online learning* Presented on September 19th, DeepMind, London, UK (*DeepMind 2017*)
- M. Valko, *Active learning on networks and online influence maximization*, Presented on September 18th, 2017, Decision Theory and Network Science: Methods and Applications, Lancaster, UK (*STOR-i 2017*)
- M. Valko, *Side observation in graph bandits*, Presented on July 11th, 2017, ICML 2017 workshop on Picky Learners, Sydney, Australia (*ICML 2017*)
- M. Valko, *Distributed sequential sampling for kernel matrix approximation*, Presented on June 28th, 2017, L'Institut de Mathématiques de Toulouse, France (*IMT 2017*)
- M. Valko, *Online sequential solutions for recommender systems*, Presented on June 14th, 2017 at Journées Scientifiques Inria 2017 in Nice, France (*JS 2017*)
- M. Valko, *Where is Justin Bieber?*, Presented on March 30th, 2017 at Dating day in Lille, France (*Dating 2017*)
- M. Valko, *Distributed sequential sampling for kernel matrix approximation*, Presented on March 22nd, 2017, for Universität Potsdam at Amazon (*Berlin 2017*)

10.1.4. Scientific Expertise

- É. Kaufmann was a member of the committee of Experts for Hiring junior faculty in the maths departement of Université of Lille 1
- J.Mary was a member of the industrial transfer commission of Inria Lille
- Alessandro Lazaric was reviewer for NSFC-ISF Research Grant
- Philippe Preux is a member of the evaluation committee and participates in the hiring, promotion, and evaluation juries of Inria:
  - Inria CR1 hiring committee
  - Inria Lille CR2 hiring committee
  - Inria committee for researcher promotion
  - Inria committee for PEDR

- Philippe Preux was a member of the hiring committees for 1 professor and 2 associate professors at the Université de Lille 3
- Philippe Preux was a member of the committee for PhD grant of the "Pôle Métropolitain de la Côte d'Opale"
- Philippe Preux reviewed a proposal for ANRT (and declined invitation from ANR)
- M. Valko is an elected member of the evaluation committee and participates in the hiring, promotion, and evaluation juries of Inria, notably
  - Hiring committee for junior researchers at Inria Saclay (2017)
  - Inria work group for deontological ethics (2017)
  - Selection committee for Inria award for scientific excellence of junior and confirmed researchers (2017)
- M. Valko was a member national Inria acceptance committee for hiring junior researchers
- M. Valko was a member of the committee of Experts for Hiring junior faculty at CMLA, ENS Paris-Saclay

10.1.5. Research Administration

- *M. Gaudel* was member of the Board of CRIStAL.
- 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
  - the head of the "Data Intelligence" thematic group at CRIStAL

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Master: É. Kaufmann, 2017/2018 Fall: Machine Learning, 18h eq TD, M2 Maths/Finances, Université de Lille 1

Master: É. Kaufmann, 2016/2017 Spring: Data Mining, 36h eq TD, M1 Maths/Finances, Université de Lille 1

Master: A. Lazaric, 2017/2018 Fall: Reinforcement Learning, 36h eqTD, M2, ENS Cachan

Master: M. Valko, 2017/2018 Fall: Graphs in Machine Learning, 36h eqTD, M2, ENS Cachan

10.2.2. Supervision

PhD in progress: Marc Abeille, Exploration-exploitation in reinforcement learning, started Sept. 2014, advisor: Remi Munos, Alessandro Lazaric

PhD in progress: Merwan Barlier, Human-in-the loop reinforrcement learning for dialogue systems, started Oct. 2014, advisor: Olivier Pietquin

PhD in progress: Alexandre Bérard, Deep learning for post-editing and automatic translation, started Oct. 2014, advisor: Olivier Pietquin

PhD in progress: Lilian Besson, Bandit approach to improve Internet Of Things Communications, started Oct. 2016, advisor: Émilie Kaufmann, Christophe Moy (CentraleSupélec Rennes)

PhD in progress: Daniele Calandriello, Efficient Sequential Learning in Structured and Constrained Environment, Inria, started Oct. 2014, advisor: Michal Valko, Alessandro Lazaric

PhD in progress: Ronan Fruit, Exploration-exploitation in hierarchical reinforcement learning, Inria, started Dec. 2015, advisor: Daniil Ryabko, Alessandro Lazaric

PhD in progress: Pratik Gajane, Multi-armed bandits with unconventional feedback, started Oct. 2014, defended Nov. 14th 2017, advisor: Philippe Preux

PhD in progress: Guillaume Gautier, DPPs in ML, started Oct. 2016, advisor: Michal Valko; Rémi Bardenet

PhD in progress: Jean-Bastien Grill, Création et analyse d'algorithmes efficaces pour la prise de décision dans un environnement inconnu et incertain, Inria/ENS Paris/Lille 1, started Oct. 2014, advisor: Rémi Munos, Michal Valko

PhD in progress: Édouard Leurent, Autonomous vehicle control: application of machine learning to contextualized path planning, started Oct. 2017, advisor: Odalric Maillard, Philippe Preux, Denis Effimov (NON-A), Wilfrid Perruquetti (NON-A)

PhD in progress: Sheikh Waqas Akhtar, Bandits for non-stationarity and structure, started Oct. 2017, advisor: Odalric Maillard, Daniil Ryabko.

PhD in progress: Julien Perolat, Reinforcement learning: the multi-player case, started Oct. 2014, advisor: Olivier Pietquin

PhD in progress: Pierre Perrault, Online Learning on Streaming Graphs, started Sep. 2017, advisor: Michal Valko; Vianney Perchet

PhD in progress: Mathieu Seurin, Multi-scale rewards in reinforcement learning, started Oct. 2017, advisor: Olivier Pietquin, Philippe Preux

PhD in progress: Julien Seznec, Sequential Learning for Educational Systems, started Mar. 2017, advisor: Michal Valko; Alessandro Lazaric, Jonathan Banon

PhD in progress: Xuedong Shang, Adaptive methods for optimization in stochastic environments, started Oct. 2017, advisor: Émilie Kaufmann, Michal Valko

PhD in progress: Florian Strub, Reinforcement Learning for visually grounded interaction, started Jan. 2016, advisors: Olivier Pietquin and Jeremie Mary

PhD in progress: Kiewan Villatel, Deep Learning for Conversion Rate Prediction in Online Advertising, started Oct. 2017, advisor: Philippe Preux

10.2.3. Juries

PhD and HDR juries:

- É. Kaufmann, *Navikumar Modi*, CentraleSupélec Rennes, May 2017
- A. Lazaric:
  - *Stefano Paladino*, Politecnico di Milano, Dec 2017
  - *Micheal Castronovo*, Université de Liege, March 2017
  - *Raffaello Camoriano*, Universitá di Genova, April 2017
  - *Claire Vernade*, TelecomParis Tech, October 2017
- Ph. Preux:
  - Cricia Zilda Felicio Paixao, Uniervity Uberlandia, Brasil
  - Thibault Gisselbrecht, LIP 6, UPMC, Paris
  - Pratik Gajane, CRIStAL, Lille
- M. Valko: *Clément Bouttier*, Université Toulouse 3 Paul Sabatier, June 2017

PhD mid-term evaluation:

- M. Valko: *Thibault Liétard*, Université Lille, September 2017

10.3. Popularization

- CNRS publishes an article about zonotope sampling presented at ICML (see http://www.cnrs.fr/ins2i/spip.php?article2633).
- Julien Seznec publishes an article in *Les Echos* that discusses ML for education (November 2017).
- Émilie Kaufmann gave a popularization talk about bandit algorithms aimed at high school/prepa students at the MathPark seminar, organized at IHP in Paris (April 2017).
- *Avec GuessWhat?! quand l'humain joue, l'ordinateur s'initie au langage*, https://www.inria.fr/centre/lille/actualites/avec-guesswhat-!-quand-l-humain-joue-l-ordinateur-s-initie-au-langage
- Florian Strub and Mathieu Seurin demonstrated guesswhat?! during the celebrations of Inria 50th anniversary (November 2017).
- Philippe Preux:
  - interviewed for an article on *L'intelligence artificielle, est-ce vraiment de l'intelligence ?* in *BioTech.info*, Jan. 2017.
  - participates to a debate about Artificial Intelligence, as part of the franceIA tour (Eurate-chnologies, Lille).
  - interview by AFP in relation to alphaGo.
  - interviewed for an article on AI and games, published in *Le figaro*.
  - an interview that led to a publication in ATOS Connexion, the ATOS internal journal.
  - a video has been made with him being interviewed on Artificial Intelligence by NordEka (to be available on youtube).
  - has been selected to be protrayed at the "Soirée partenaires de l'université de Lille", Nov.
  - was a member of the organization comittee of the celebrations of the 50th Inria anniversary in Lille.
  - co-organizes a meet-up on big data and machine learning at Inria.
- M. Valko, *Comment maximiser la détection des influenceurs sur les réseaux sociaux ?*, popularization talk, Presented on May 30th, 2017 at 13 France (*Inria 13:45 2017*)

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° 3, pp. 1516-1541, Accepted, to appear in Annals of Statistics, https://hal.archives-ouvertes.fr/hal-00738209

[2] A. CARPENTIER, M. VALKO. *Revealing graph bandits for maximizing local influence*, in "International Conference on Artificial Intelligence and Statistics", Seville, Spain, May 2016, https://hal.inria.fr/hal-01304020

[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, 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)", 2016, 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] M. ABEILLE. *Exploration-Exploitation with Thompson Sampling in Linear Systems*, Université de Lille, December 2017

[12] D. CALANDRIELLO. *Efficient Sequential Learning in Structured and Constrained Environments*, Université de Lille, December 2017

[13] P. GAJANE. *Multi-armed bandits with unconventional feedback*, Université de Lille, November 2017

[14] J. PÉROLAT. *Reinforcement learning: the multiplayer case*, Université de Lille, December 2017

Articles in International Peer-Reviewed Journals

[15] B. DANGLOT, P. PREUX, B. BAUDRY, M. MONPERRUS. *Correctness Attraction: A Study of Stability of Software Behavior Under Runtime Perturbation*, in "Empirical Software Engineering", 2017, https://arxiv.org/abs/1611.09187 [*DOI :* 10.1007/s10664-017-9571-8], https://hal.archives-ouvertes.fr/hal-01378523

[16] C. DIMITRAKAKIS, B. NELSON, Z. ZHANG, A. MITROKOTSA, B. I. P. RUBINSTEIN. *Differential Privacy for Bayesian Inference through Posterior Sampling*, in "Journal of Machine Learning Research", April 2017, vol. 18, n^{o} 11, 1-39 p. , https://hal.inria.fr/hal-01500302

[17] E. KAUFMANN, T. BONALD, M. LELARGE. *A Spectral Algorithm with Additive Clustering for the Recovery of Overlapping Communities in Networks*, in "Journal of Theoretical Computer Science (TCS)", 2017, https://arxiv.org/abs/1506.04158 , forthcoming, https://hal.archives-ouvertes.fr/hal-01163147

[18] E. KAUFMANN, A. GARIVIER. *Learning the distribution with largest mean: two bandit frameworks*, in "ESAIM: Proceedings and Surveys", 2017, vol. 2017, pp. 1 - 10, https://arxiv.org/abs/1702.00001 , forthcoming, https://hal.archives-ouvertes.fr/hal-01449822

[19] E. KAUFMANN. *On Bayesian index policies for sequential resource allocation*, in "Annals of Statistics", 2017, https://arxiv.org/abs/1601.01190 , forthcoming, https://hal.archives-ouvertes.fr/hal-01251606

[20] V. MUSCO, M. MONPERRUS, P. PREUX. *A Large-scale Study of Call Graph-based Impact Prediction using Mutation Testing*, in "Software Quality Journal", September 2017, vol. 25, n° 3, pp. 921–950 [*DOI :* 10.1007/s11219-016-9332-8], https://hal.inria.fr/hal-01346046

International Conferences with Proceedings

[21] M. ABEILLE, A. LAZARIC. *Linear Thompson Sampling Revisited*, in "AISTATS 2017 - 20th International Conference on Artificial Intelligence and Statistics", Fort Lauderdale, United States, April 2017, https://hal.inria.fr/hal-01493561

[22] M. ABEILLE, A. LAZARIC. *Thompson Sampling for Linear-Quadratic Control Problems*, in "AISTATS 2017 - 20th International Conference on Artificial Intelligence and Statistics", Fort Lauderdale, United States, April 2017, https://hal.inria.fr/hal-01493564

[23] B. BALLE, O.-A. MAILLARD. *Spectral Learning from a Single Trajectory under Finite-State Policies*, in "International conference on Machine Learning", Sidney, France, Proceedings of the International conference on Machine Learning, July 2017, https://hal.archives-ouvertes.fr/hal-01590940

[24] S. BRODEUR, E. PEREZ, A. ANAND, F. GOLEMO, L. CELOTTI, F. STRUB, J. ROUAT, H. LAROCHELLE, A. COURVILLE. *HoME: a Household Multimodal Environment*, in "NIPS 2017's Visually-Grounded Interaction and Language Workshop", Long Beach, United States, December 2017, https://arxiv.org/abs/1711.11017 , https://hal.inria.fr/hal-01653037

[25] A. BÉRARD, O. PIETQUIN, L. BESACIER. *LIG-CRIStAL System for the WMT17 Automatic Post-Editing Task*, in "Second conference on machine translation (WMT17) during EMNLP 2017", Copenhague, Denmark, September 2017, https://hal.archives-ouvertes.fr/hal-01580881

[26] D. CALANDRIELLO, A. LAZARIC, M. VALKO. *Distributed adaptive sampling for kernel matrix approximation*, in "International Conference on Artificial Intelligence and Statistics", Fort Lauderdale, United States, 2017, https://hal.inria.fr/hal-01482760

[27] D. CALANDRIELLO, A. LAZARIC, M. VALKO. *Efficient second-order online kernel learning with adaptive embedding*, in "NIPS 2017 : The Thirty-first Annual Conference on Neural Information Processing Systems", Long Beach, United States, December 2017, pp. 1-17, https://hal.inria.fr/hal-01643961

[28] D. CALANDRIELLO, A. LAZARIC, M. VALKO. *Second-Order Kernel Online Convex Optimization with Adaptive Sketching*, in "International Conference on Machine Learning", Sydney, Australia, 2017, https://hal.inria.fr/hal-01537799

[29] H. DE VRIES, F. STRUB, S. CHANDAR, O. PIETQUIN, H. LAROCHELLE, A. COURVILLE. *GuessWhat?! Visual object discovery through multi-modal dialogue*, in "Conference on Computer Vision and Pattern Recognition", Honolulu, United States, July 2017, https://arxiv.org/abs/1611.08481 , https://hal.inria.fr/hal-01549641

[30] H. DE VRIES, F. STRUB, J. MARY, H. LAROCHELLE, O. PIETQUIN, A. COURVILLE. *Modulating early visual processing by language*, in "NIPS 2017 - Conference on Neural Information Processing Systems", Long Beach, United States, December 2017, pp. 1-14, https://arxiv.org/abs/1707.00683 , https://hal.inria.fr/hal-01648683

[31] A. ERRAQABI, A. LAZARIC, M. VALKO, E. BRUNSKILL, Y.-E. LIU. *Trading off rewards and errors in multi-armed bandits*, in "International Conference on Artificial Intelligence and Statistics", Fort Lauderdale, United States, 2017, https://hal.inria.fr/hal-01482765

[32] C. Z. FELÍCIO, K. V. R. PAIXÃO, C. A. Z. BARCELOS, P. PREUX. *A Multi-Armed Bandit Model Selection for Cold-Start User Recommendation*, in "25th ACM Conference on User Modelling, Adaptation and Personalization (UMAP)", Bratislava, Slovakia, July 2017, https://hal.inria.fr/hal-01517967

[33] R. FRUIT, A. LAZARIC. *Exploration–Exploitation in MDPs with Options*, in "AISTATS 2017 - 20th International Conference on Artificial Intelligence and Statistics", Fort Lauderdale, United States, April 2017, https://hal.inria.fr/hal-01493567

[34] R. FRUIT, M. PIROTTA, A. LAZARIC, E. BRUNSKILL. *Regret Minimization in MDPs with Options without Prior Knowledge*, in "NIPS 2017 - Neural Information Processing Systems", Long Beach, United States, December 2017, pp. 1-36, https://hal.inria.fr/hal-01649082

[35] G. GAUTIER, R. BARDENET, M. VALKO. *Zonotope hit-and-run for efficient sampling from projection DPPs*, in "International Conference on Machine Learning", Sydney, Australia, 2017, https://hal.inria.fr/hal-01526577

[36] M. GEIST, B. PIOT, O. PIETQUIN. *Is the Bellman residual a bad proxy?*, in "NIPS 2017 - Advances in Neural Information Processing Systems", Long Beach, United States, December 2017, pp. 1-13, https://hal.archives-ouvertes.fr/hal-01629739

[37] E. KAUFMANN, W. M. KOOLEN. *Monte-Carlo Tree Search by Best Arm Identification*, in "NIPS 2017 - 31st Annual Conference on Neural Information Processing Systems", Long Beach, United States, Advances in Neural Information Processing Systems, December 2017, pp. 1-23, https://arxiv.org/abs/1706.02986 , https://hal.archives-ouvertes.fr/hal-01535907

[38] R. LAROCHE, M. BARLIER. *Transfer Reinforcement Learning with Shared Dynamics*, in "AAAI-17 - Thirty-First AAAI Conference on Artificial Intelligence", San Francisco, United States, February 2017, 7 p. , https://hal.archives-ouvertes.fr/hal-01548649

[39] O.-A. MAILLARD. *Boundary Crossing for General Exponential Families*, in "Algorithmic Learning Theory", Kyoto, Japan, Proceedings of Algorithmic Learning Theory, October 2017, vol. 1, pp. 1 - 34, https://hal.archives-ouvertes.fr/hal-01615427

[40] A. M. METELLI, M. PIROTTA, M. RESTELLI. *Compatible Reward Inverse Reinforcement Learning*, in "The Thirty-first Annual Conference on Neural Information Processing Systems - NIPS 2017", Long Beach, United States, December 2017, https://hal.inria.fr/hal-01653328

[41] J. MOURTADA, O.-A. MAILLARD. *Efficient tracking of a growing number of experts*, in "Algorithmic Learning Theory", Tokyo, Japan, Proceedings of Algorithmic Learning Theory, October 2017, vol. 76, pp. 1 - 23, https://hal.archives-ouvertes.fr/hal-01615424

[42] M. PAPINI, M. PIROTTA, M. RESTELLI. *Adaptive Batch Size for Safe Policy Gradients*, in "The Thirty-first Annual Conference on Neural Information Processing Systems (NIPS)", Long Beach, United States, December 2017, https://hal.inria.fr/hal-01653330

[43] G. PAPOUDAKIS, P. PREUX, M. MONPERRUS. *A generative model for sparse, evolving digraphs*, in "6th International Conference on Complex Networks and their applications", Lyon, France, November 2017, https://arxiv.org/abs/1710.06298 [*DOI :* 10.1007/978-3-319-72150-7_43], https://hal.inria.fr/hal-01617851

[44] E. PEREZ, H. DE VRIES, F. STRUB, V. DUMOULIN, A. COURVILLE. *Learning Visual Reasoning Without Strong Priors*, in "ICML 2017's Machine Learning in Speech and Language Processing Workshop", Sidney, France, August 2017, https://arxiv.org/abs/1709.07871 , https://hal.inria.fr/hal-01648684

[45] E. PEREZ, F. STRUB, H. DE VRIES, V. DUMOULIN, A. COURVILLE. *FiLM: Visual Reasoning with a General Conditioning Layer*, in "AAAI Conference on Artificial Intelligence", New Orleans, United States, February 2018, https://arxiv.org/abs/1707.03017 , https://hal.inria.fr/hal-01648685

[46] J. PÉROLAT, F. STRUB, B. PIOT, O. PIETQUIN. *Learning Nash Equilibrium for General-Sum Markov Games from Batch Data*, in "AISTATS 2017 - The 20th International Conference on Artificial Intelligence and Statistics", Fort Lauderdale, United States, April 2017, pp. 1-14, https://hal.inria.fr/hal-01648489

[47] C. RIQUELME, M. GHAVAMZADEH, A. LAZARIC. *Active Learning for Accurate Estimation of Linear Models*, in "ICML 2017 - 34th International Conference on Machine Learning", Sydney, Australia, August 2017, 36 p. , https://hal.inria.fr/hal-01538762

[48] D. RYABKO. *Hypotheses testing on infinite random graphs*, in "ALT 2017 - 28th International Conference on Algorithmic Learning Theory", kyoto, Japan, October 2017, pp. 1-12, https://arxiv.org/abs/1708.03131 , https://hal.inria.fr/hal-01627330

[49] D. RYABKO. *Independence clustering (without a matrix)*, in "NIPS 2017 - Thirty-first Annual Conference on Neural Information Processing Systems", Long Beach, United States, December 2017, pp. 1-14, https://arxiv.org/abs/1703.06700 , https://hal.inria.fr/hal-01627333

[50] D. RYABKO. *Universality of Bayesian mixture predictors*, in "ALT 2017 - 28th International Conference on Algorithmic Learning Theory", Kyoto, Japan, October 2017, pp. 1-13, https://arxiv.org/abs/1610.08249 , https://hal.inria.fr/hal-01627332

[51] F. STRUB, H. DE VRIES, J. MARY, B. PIOT, A. COURVILLE, O. PIETQUIN. *End-to-end optimization of goal-driven and visually grounded dialogue systems Harm de Vries*, in "International Joint Conference on Artificial Intelligence", Melbourne, Australia, August 2017, https://arxiv.org/abs/1703.05423 , https://hal.inria.fr/hal-01549642

[52] S. TOSATTO, M. PIROTTA, C. D'ERAMO, M. RESTELLI. *Boosted Fitted Q-Iteration*, in "34th International Conference on Machine Learning (ICML)", Sydney, Australia, August 2017, https://hal.inria.fr/hal-01653332

[53] N. TZIORTZIOTIS, C. DIMITRAKAKIS. *Bayesian Inference for Least Squares Temporal Difference Regularization*, in "ECML 2017 - European Conference on Machine Learning", Skopje, Macedonia, 2017-09-22, September 2017, https://hal.inria.fr/hal-01593212

[54] Z. WEN, B. KVETON, M. VALKO, S. VASWANI. *Online influence maximization under independent cascade model with semi-bandit feedback*, in "NIPS 2017 - Neural Information Processing Systems", Long Beach, United States, December 2017, pp. 1-24, https://hal.inria.fr/hal-01643976

[55] M. ZANON BOITO, A. BÉRARD, A. VILLAVICENCIO, L. BESACIER. *Unwritten Languages Demand Attention Too! Word Discovery with Encoder-Decoder Models*, in "IEEE Automatic Speech Recognition and Understanding (ASRU)", Okinawa, Japan, December 2017, https://hal.archives-ouvertes.fr/hal-01592091

National Conferences with Proceedings

[56] M. GEIST, B. PIOT, O. PIETQUIN. *Faut-il minimiser le résidu de Bellman ou maximiser la valeur moyenne ?*, in "Journées Francophones sur la Planification, la Décision et l'Apprentissage pour la conduite de systèmes (JFPDA 2017)", Caen, France, Actes des Journées Francophones sur la Planification, la Décision et l'Apprentissage pour la conduite de systèmes (JFPDA 2017), July 2017, https://hal.archives-ouvertes.fr/hal-01576347

Conferences without Proceedings

[57] R. BONNEFOI, L. BESSON, C. MOY, E. KAUFMANN, J. PALICOT. *Multi-Armed Bandit Learning in IoT Networks: Learning helps even in non-stationary settings*, in "CROWNCOM 2017 - 12th EAI International Conference on Cognitive Radio Oriented Wireless Networks", Lisbon, Portugal, September 2017, https://hal.archives-ouvertes.fr/hal-01575419

[58] N. CARRARA, R. LAROCHE, O. PIETQUIN. *Online learning and transfer for user adaptation in dialogue systems*, in "SIGDIAL/SEMDIAL joint special session on negotiation dialog 2017", Saarbrücken, Germany, August 2017, https://hal.archives-ouvertes.fr/hal-01557775

Other Publications

[59] L. BESSON, E. KAUFMANN. *Multi-Player Bandits Models Revisited*, October 2017, https://arxiv.org/abs/1711.02317 - working paper or preprint, https://hal.inria.fr/hal-01629733

[60] C. DIMITRAKAKIS, F. JARBOUI, D. PARKES, L. SEEMAN. *Multi-view Sequential Games: The Helper-Agent Problem*, February 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01408294

[61] C. DIMITRAKAKIS, Y. LIU, D. PARKES, G. RADANOVIC. *Subjective Fairness: Fairness is in the eye of the beholder*, July 2017, https://arxiv.org/abs/1706.00119 - working paper or preprint, https://hal.inria.fr/hal-01531849

[62] A. R. LUEDTKE, E. KAUFMANN, A. CHAMBAZ. *Asymptotically Optimal Algorithms for Budgeted Multiple Play Bandits*, October 2017, https://arxiv.org/abs/1606.09388 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01338733

[63] O.-A. MAILLARD. *Basic Concentration Properties of Real-Valued Distributions*, September 2017, Lecture, https://hal.archives-ouvertes.fr/cel-01632228

References in notes

[64] R. ALLESIARDO, R. FÉRAUD, O.-A. MAILLARD. *The Non-stationary Stochastic Multi-armed Bandit Problem*, in "International Journal of Data Science and Analytics", 2017, vol. 3, n^{o} 4, pp. 267–283 [*DOI :* 10.1007/s41060-017-0050-5], https://hal.archives-ouvertes.fr/hal-01575000

[65] 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

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

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

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

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

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

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

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