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

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2015

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:

3. - Data and knowledge
3.1. - Data
3.1.4. - Uncertain data
3.3. - Data and knowledge analysis
3.3.2. - Data mining
3.3.3. - Big data analysis
3.4. - Machine learning and statistics
3.4.1. - Supervised learning
3.4.2. - Unsupervised learning
3.4.3. - Reinforcement learning
3.4.4. - Optimization and learning
3.4.6. - Neural networks
3.4.8. - Deep learning
3.5.2. - Recommendation systems
5.1. - Human-Computer Interaction
8. - Artificial intelligence
8.2. - Machine learning

Other Research Topics and Application Domains:

5.8. - Learning and training
6.1.1. - Software engineering
9.1.1. - E-learning, MOOC
9.4. - Sciences
9.4.5. - Data science

1. Members

Research Scientists

Alessandro Lazaric [Inria, Researcher]
Rémi Munos [Inria, Senior Researcher, HdR]
Daniil Ryabko [Inria, Researcher, HdR]
Michal Valko [Inria, Researcher]

Faculty Members

Philippe Preux [Team leader, Univ. Lille III, Professor, HdR]
Christos Dimitrakakis [Univ. Lille III, Associate Professor, since Oct 2015, HdR]
Romaric Gaudel [Univ. Lille III, Associate Professor]
Jérémie Mary [Univ. Lille III, Associate Professor]
Olivier Pietquin [Univ. Lille I, Professor, HdR]
Bilal Piot [Univ. Lille I, Associate Professor]

Engineer

Florian Strub [Inria, from Mar 2015]

PhD Students

Marc Abeille [Univ. Lille I]
Merwan Barlier [Orange Labs, granted by CIFRE]
Alexandre Berard [Univ. Lille I]
Daniele Calandriello [Inria]
Ronan Fruit [Inria, from Dec 2015]
Pratik Gajane [Orange Labs, granted by CIFRE]
Hadrien Glaude [Thales, until Jul 2015, granted by CIFRE]
Jean-Bastien Grill [Univ. Lille I, granted by ENS Paris]
Frédéric Guillou [Inria]
Adrien Hoarau [Inria, until Oct 2015, granted by DGA]
Tomáš Kocák [Inria]
Vincenzo Musco [Univ. Lille I and III]
Julien Perolat [Univ. Lille I]
Amir Sani [Inria, until Mar 2015]
Marta Soare [Inria, until Sep 2015, co-funded by Région Nord Pas de Calais]

Post-Doctoral Fellow

Gergely Neu [Inria, until Aug 2015, granted by OSEO Innovation]

Visiting Scientist

Cricia Zilda Felicio Paixao [, from Sep 2015]

Administrative Assistants

Natacha Oudoire [Inria]
Amelie Supervielle [Inria]

Others

Mastane Achab [École Polytechnique, granted by Univ. Lille I, Intern, from Mar 2015 until Jul 2015]
Antonin Carette [Univ. Lille I, Intern, from Jun 2015 until Aug 2015]
Akram Er-Raqabi [École Polytechnique, granted by Univ. Lille I, Intern, from May 2015 until Nov 2015]
Romain Philippon [Univ. Lille I, Intern, from Mar 2015 until Jul 2015]
Pierre Chainais [Ecole Centrale de Lille, Associate Professor, HdR]
Pierre-Victor Chaumier [Inria, until Jun 2015]
Emilie Kaufmann [CNRS, since Oct 2015]

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 [48]). 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 [46].

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 $\mathbb{P}\left(x_{t+1} = x'|x_t = x, a_t = a\right) = p(x'|x, a)$, 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:

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

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 [45]) 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) [43], 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 [44]. 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 $x \in \mathcal{X}$,

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

[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 $\arg\max_{a \in \mathcal{A}} f(a)$ 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 ([49]):

- 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 [47], 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.* [42] 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 piecewise 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 $x^1 = (x^1_1, \cdots, x^1_{n_1}), \cdots, x^N = (x^1_N, \cdots, x^N_{n_N})$, 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.

4. Application Domains

4.1. Recommendation systems in a broad sense

Recommendation systems have been a major field of applications of our research for a few years now. Recommendation systems should be understood in a broad sense, as systems that aim at providing personalized responses/items to users, based on their characteristics, and the environment in which the interaction happens.

In that broad sense, we have collaborated with companies on computational advertizing and recommendation systems. These collaborations have involved research studies on the following issues:

- cold-start problem,
- time varying environment,
- ability to deal with large amounts of users and items,
- ability to design algorithms to respond within a reasonnable amount of time, usually below 1 millisecond.

We have also competed in challenges, winning some of them [2], and we have also organized a challenge [3], on those topics.

A company has been awarded an innovation award in 2015, thanks to the research work done in collaboration with SEQUEL (*cf.* sec. 1).

In these works, we develop an original [4] point of view on such systems. While traditional (before say 2010) recommendation systems were seen as solving a supervised learning task, or a ranking task, we have developed the idea that recommender systems are really a problem of sequential decision making under uncertainty.

We also started a new work aiming to introduce deep learning in recommender systems. An engineer (Florian Strub) was recruited to work on this topic and presented some results at the NIPS'2015 workshop on "Machine Learning for (e-)Commerce". Moreover we released some code to handle sparse data with the Torch7 framework and GPUs https://github.com/fstrub95/nnsparse.

4.2. Spoken dialog systems

A Spoken Dialogue System (SDS) is a system enabling human people to interact with machines through speech. In contrast with command-and-control systems or question-answering systems that react to a single utterances, SDS build a real interaction over time and try to achieve complex tasks (like hotel booking, appointment scheduling etc.) by gathering pieces of information through several turns of dialogue. To do so, besides the required speech and language processing modules (*e.g.* speech recognition and synthesis, language understanding and generation), there is a need for a dialogue management module that decides what to say in any situation so as to achieve the goal in the most natural and efficient way, recovering from speech processing errors in a seamless manner.

The dialogue management module is thus taking sequences of decisions to achieve a long-term goal in an unknown, noisy and hard to model environment (since it includes human users). For this reason, we work on machine learning techniques such as reinforcement and imitation learning to optimize this specific sequential decision making under uncertainty problem.

In addition to bring novel and efficient solutions to this problem, we are interested in the new challenges brought to our research in machine learning by this type of application. Indeed, having the human in the learning loop typically requires dealing with non-stationarity, data-efficiency, safety as well as cooperation and imitation.

We collaborate with companies such as Orange Labs on this topic and several projects are ongoing (ANR MaRDi, CHIST-ERA IGLU). We will also be participating to a H2020 project on human robot-interaction starting in 2016 (BabyRobot). We organised a workshop at ICML this year: Machine Learning for Interactive Systems (MLIS). Olivier Pietquin was invited as a panelist at the NIPS Workshop on spoken language understanding and dialogue.

[2] SEQUEL ranked first and second at the "Pascal Exploration & Exploitation Challenge 2011"; SEQUEL ranked first at the "RecSys Challenge 2014: User Engagement as Evaluation".

[3] ICML 2012 new Challenges for Exploration & Exploitation 3.

[4] the originality fades away as years pass since this idea is exploited by other researchers.

4.3. Adaptive/learning systems more generally

Reinforcement learning leads to the design of systems that adapts their behavior to their environment, hence adaptive systems. We have worked on various applications of this idea, beyond the two main applications domaines mentioned above (recommendation systems, and spoken dialog systems). Let us briefly mention: educative tutoring systems; adaptive heating system in buildings; players that adapt their strength to that of their human opponent; bioreactor.

4.4. Prediction in general

Since the goal of our research is to design systems that learn to act in an optimal way in their environment, prediction is a major issue. Hence, we are doing some research activities on this particular task, without always being in direct connection with learning a policy.

We have done some research in the area of prediction web-server load in a non stationary environment. We also have activities in the prediction in bug in software code.

5. Highlights of the Year

5.1. Highlights of the Year

- organization of the 32nd International Conference on Machine Learning (ICML), in Lille, from Jul 6th to Jul 11th, 2015.

  ICML is the leading international conference in Machine Learning. This is the first time of its history that France hosts ICML. This edition has been the largest of all the times, with 1690 registrants (the previous record was 1400 in Beijing, in 2014).
- as an outcome of a contract with this start-up, Nuukik has been awarded “best data analysis” during the “connected commerce night” - http://www.retail-network.fr", 1500 participants, 80 projects in competition.

5.1.1. Awards

- V. Gabillon and B. Piot both received an AFIA award for their respective PhD, defended in 2014. They were both ranked second in this competition.
- Olivier Pietquin, Fellow of the “Institut Universitaire de France”.
- A. Lazaric and M. Valko received best reviewer awards at ICML 2015.

6. New Software and Platforms

6.1. Function optimization

Participants: Jean-Bastien Grill, Michal Valko, Rémi Munos.

6.1.1. POO

This is a black-box function optimization toolkit that finds the global optimum of a function given a finite budget of noisy evaluations. The algorithm does not require the knowledge of the function’s smoothness. It works for a larger class of functions than what was previously considered, especially for functions that are difficult to optimize, in a precise sense.

7. New Results

7.1. Decision-making Under Uncertainty

7.1.1. Reinforcement Learning

Nonparametric multiple change point estimation in highly dependent time series [7]

Given a heterogeneous time-series sample, the objective is to find points in time, called change points, where the probability distribution generating the data has changed. The data are assumed to have been generated by arbitrary unknown stationary ergodic distributions. No modelling, independence or mixing assumptions are made. A novel, computationally efficient, nonparametric method is proposed, and is shown to be asymptotically consistent in this general framework. The theoretical results are complemented with experimental evaluations.

Explore no more: Improved high-probability regret bounds for non-stochastic bandits [26]

This work addresses the problem of regret minimization in non-stochastic multi-armed bandit problems, focusing on performance guarantees that hold with high probability. Such results are rather scarce in the literature since proving them requires a large deal of technical effort and significant modifications to the standard, more intuitive algorithms that come only with guarantees that hold on expectation. One of these modifications is forcing the learner to sample arms from the uniform distribution at least $\Omega(\sqrt{T})$ times over T rounds, which can adversely affect performance if many of the arms are suboptimal. While it is widely conjectured that this property is essential for proving high-probability regret bounds, we show in this paper that it is possible to achieve such strong results without this undesirable exploration component. Our result relies on a simple and intuitive loss-estimation strategy called Implicit eXploration (IX) that allows a remarkably clean analysis. To demonstrate the flexibility of our technique, we derive several improved high-probability bounds for various extensions of the standard multi-armed bandit framework. Finally, we conduct a simple experiment that illustrates the robustness of our implicit exploration technique.

First-order regret bounds for combinatorial semi-bandits [27]

We consider the problem of online combinatorial optimization under semi-bandit feedback, where a learner has to repeatedly pick actions from a combinatorial decision set in order to minimize the total losses associated with its decisions. After making each decision, the learner observes the losses associated with its action, but not other losses. For this problem, there are several learning algorithms that guarantee that the learner's expected regret grows as $O(\sqrt{T})$ with the number of rounds T. In this paper, we propose an algorithm that improves this scaling to $O(\sqrt{L * T})$, where L * T is the total loss of the best action. Our algorithm is among the first to achieve such guarantees in a partial-feedback scheme, and the first one to do so in a combinatorial setting.

Random-Walk Perturbations for Online Combinatorial Optimization [4]

We study online combinatorial optimization problems where a learner is interested in minimizing its cumulative regret in the presence of switching costs. To solve such problems, we propose a version of the follow-the-perturbed-leader algorithm in which the cumulative losses are perturbed by independent symmetric random walks. In the general setting, our forecaster is shown to enjoy near-optimal guarantees on both quantities of interest, making it the best known efficient algorithm for the studied problem. In the special case of prediction with expert advice, we show that the forecaster achieves an expected regret of the optimal order $O(\sqrt{n \log N})$ where n is the time horizon and N is the number of experts, while guaranteeing that the predictions are switched at most $O(\sqrt{n \log N})$ times, in expectation.

Qualitative Multi-Armed Bandits: A Quantile-Based Approach [32]

We formalize and study the multi-armed bandit (MAB) problem in a generalized stochastic setting, in which rewards are not assumed to be numerical. Instead, rewards are measured on a qualitative scale that allows for comparison but invalidates arithmetic operations such as averaging. Correspondingly, instead of characterizing an arm in terms of the mean of the underlying distribution, we opt for using a quantile of that distribution as a representative value. We address the problem of quantile-based online learning both for the case of a finite (pure exploration) and infinite time horizon (cumulative regret minimization). For both cases, we propose suitable algorithms and analyze their properties. These properties are also illustrated by means of first experimental studies.

Predicting the outcomes of every process for which an asymptotically accurate stationary predictor exists is impossible [30]

The problem of prediction consists in forecasting the conditional distribution of the next outcome given the past. Assume that the source generating the data is such that there is a stationary predictor whose error converges to zero (in a certainsense). The question is whether there is a universal predictor for all such sources, that is, a predictor whose error goes to zero if any of the sources that have this property is chosen to generate the data. This question is answered in the negative, contrasting a number of previously established positive results concerning related but smaller sets of processes.

Improved Regret Bounds for Undiscounted Continuous Reinforcement Learning [22]

We consider the problem of undiscounted reinforcement learning in continuous state space. Regret bounds in this setting usually hold under various assumptions on the structure of the reward andtransition function. Under the assumption that the rewards andtransition probabilities are Lipschitz, for 1-dimensional state space a regret bound of $\widetilde{O}(T^{\frac{3}{4}})$ after any T steps has been given by.Here we improve upon this result by using non-parametric kernel density estimation for estimating the transition probability distributions,and obtain regret bounds that depend on the smoothness of the transition probability distributions.In particular, under the assumption that the transition probability functions are smoothly differentiable, the regret bound is shown to be $\widetilde{O}(T^{\frac{2}{3}})$ asymptotically for reinforcement learning in 1-dimensional state space. Finally, we also derive improved regret bounds for higher dimensional state space.

Maximum Entropy Semi-Supervised Inverse Reinforcement Learning [9]

A popular approach to apprenticeship learning (AL) is to formulate itas an inverse reinforcement learning (IRL) problem. The MaxEnt-IRL algorithm successfully integrates the maximum entropy principleinto IRL and unlike its predecessors, it resolves theambiguity arising from the fact that a possibly large number of policies couldmatch the expert's behavior. In this paper, we study an AL setting in which inaddition to the expert's trajectories,a number of unsupervised trajectories is available. We introduce MESSI,a novel algorithm that combines MaxEnt-IRLwith principles coming from semi-supervised learning. In particular, MESSIintegrates the unsupervised data intothe MaxEnt-IRL framework using a pairwise penalty on trajectories. Empiricalresults in a highway driving and grid-world problems indicate that MESSI is able to take advantage of the unsupervised trajectories and improve the performance ofMaxEnt-IRL.

Direct Policy Iteration with Demonstrations [12]

We consider the problem of learning the optimal policy of an unknown Markov decision process (MDP) when expert demonstrations are available along with interaction samples. We build on classification-based policy iteration to perform a seamless integration of interaction and expert data, thus obtaining an algorithm which can benefit from both sources of information at the same time. Furthermore , we provide a full theoretical analysis of the performance across iterations providing insights on how the algorithm works. Finally, we report an empirical evaluation of the algorithm and a comparison with the state-of-the-art algorithms.

Approximate Modified Policy Iteration and its Application to the Game of Tetris [8]

Modified policy iteration (MPI) is a dynamic programming (DP) algorithm that contains the two celebrated policy and value iteration methods. Despite its generality, MPI has not been thoroughly studied, especially its approximation form which is used when the state and/or action spaces are large or infinite. In this paper, we propose three implementations of approximate MPI (AMPI) that are extensions of the well-known approximate DP algorithms: fitted-value iteration, fitted-Q iteration, and classification-based policy iteration. We provide error propagation analysis that unify those for approximate policy and value iteration. We develop the finite-sample analysis of these algorithms, which highlights the influence of their parameters. In the classification-based version of the algorithm (CBMPI), the analysis shows that MPI's main parameter controls the balance between the estimation error of the classifier and the overall value function approximation. We illustrate and evaluate the behavior of these new algorithms in the Mountain Car and Tetris problems. Remarkably, in Tetris, CBMPI outperforms the existing DP approaches by a large margin, and competes with the current state-of-the-art methods while using fewer samples.

7.1.2. Multi-arm Bandit Theory

Simple regret for infinitely many armed bandits [11]

We consider a stochastic bandit problem with infinitely many arms. In this setting, the learner has no chance of trying all the arms even once and has to dedicate its limited number of samples only to a certain number of arms. All previous algorithms for this setting were designed for minimizing the cumulative regret of the learner. In this paper, we propose an algorithm aiming at minimizing the simple regret. As in the cumulative regret setting of infinitely many armed bandits , the rate of the simple regret will depend on a parameter β characterizing the distribution of the near-optimal arms. We prove that depending on β, our algorithm is minimax optimal either up to a multiplicative constant or up to a log(n) factor. We also provide extensions to several important cases: when β is unknown, in a natural setting where the near-optimal arms have a small variance , and in the case of unknown time horizon.

Black-box optimization of noisy functions with unknown smoothness [20]

We study the problem of black-box optimization of a function f of any dimension, given function evaluations perturbed by noise. The function is assumed to be locally smooth around one of its global optima, but this smoothness is unknown. Our contribution is an adaptive optimization algorithm, POO or parallel optimistic optimization, that is able to deal with this setting. POO performs almost as well as the best known algorithms requiring the knowledge of the smoothness. Furthermore, POO works for a larger class of functions than what was previously considered, especially for functions that are difficult to optimize, in a very precise sense. We provide a finite-time analysis of POO's performance, which shows that its error after n evaluations is at most a factor of $\sqrt{\ln n}$ away from the error of the best known optimization algorithms using the knowledge of the smoothness.

Cheap Bandits [21]

We consider stochastic sequential learning problems where the learner can observe the average reward of several actions. Such a setting is interesting in many applications involving monitoring and surveillance, where the set of the actions to observe represent some (geographical) area. The importance of this setting is that in these applications , it is actually cheaper to observe average reward of a group of actions rather than the reward of a single action. We show that when the reward is smooth over a given graph representing the neighboring actions, we can maximize the cumulative reward of learning while minimizing the sensing cost. In this paper we propose CheapUCB, an algorithm that matches the regret guarantees of the known algorithms for this setting and at the same time guarantees a linear cost again over them. As a by-product of our analysis , we establish a (p dT) lower bound on the cumulative regret of spectral bandits for a class of graphs with effective dimension d.

Truthful Learning Mechanisms for Multi–Slot Sponsored Search Auctions with Externalities [5]

Sponsored Search Auctions (SSAs) constitute one of the most successful applications of microeconomic mechanisms. In mechanism design, auctions are usually designed to incentivize advertisers to bid their truthful valuations and, at the same time, to guarantee both the advertisers and the auctioneer a non–negative utility. Nonetheless, in sponsored search auctions, the Click–Through–Rates (CTRs) of the advertisers are often unknown to the auctioneer and thus standard truthful mechanisms cannot be directly applied and must be paired with an effective learning algorithm for the estimation of the CTRs. This introduces the critical problem of designing a learning mechanism able to estimate the CTRs at the same time as implementing a truthful mechanism with a revenue loss as small as possible compared to the mechanism that can exploit the true CTRs. Previous work showed that, when dominant–strategy truthfulness is adopted, in single–slot auctions the problem can be solved using suitable exploration–exploitation mechanisms able to achieve a cumulative regret (on the auctioneer's revenue) of order $O(T^{(2/3)})$, where T is the number of times the auction is repeated. It is also known that, when truthfulness in expectation is adopted, a cumulative regret (over the social welfare) of order $O(T^{(1/2)})$ can be obtained. In this paper, we extend the results available in the literature to the more realistic case of multi–slot auctions. In this case, a model of the user is needed to characterize how the CTR of an ad changes as its position in the allocation changes. In particular, we adopt the cascade model, one of the most popular models for sponsored search auctions, and we prove a number of novel upper bounds and lower bounds on both auctioneer's revenue loss and social welfare w.r.t. to the Vickrey–Clarke–Groves (VCG) auction. Furthermore, we report numerical simulations investigating the accuracy of the bounds in predicting the dependency of the regret on the auction parameters.

A Relative Exponential Weighing Algorithm for Adversarial Utility-based Dueling Bandits *[37]*

We study the K-armed dueling bandit problem which is a variation of the classical Multi-Armed Bandit (MAB) problem in which the learner receives only relative feedback about the selected pairs of arms. We propose a new algorithm called Relative Exponential-weight algorithm for Exploration and Exploitation (REX3) to handle the adversarial utility-based formulation of this problem. This algorithm is a non-trivial extension of the Exponential-weight algorithm for Exploration and Exploitation (EXP3) algorithm. We prove a finite time expected regret upper bound of order O(sqrt(K ln(K)T)) for this algorithm and a general lower bound of order omega(sqrt(KT)). At the end, we provide experimental results using real data from information retrieval applications.

Simultaneous Optimistic Optimization on the Noiseless BBOB Testbed [15]

We experiment the SOO (Simultaneous Optimistic Optimization) global optimizer on the BBOB testbed. We report results for both the unconstrained-budget setting and the expensive setting, as well as a comparison with the DiRect algorithm to which SOO is mostly related. Overall, SOO is shown to perform rather poorly in the highest dimensions while agreeably exhibiting interesting performance for the most difficult functions, which is to be attributed to its global nature and to the fact that its design was guided by the goal of obtaining theoretically provable performance. The greedy exploration-exploitation sampling strategy underlying SOO design is also shown to be a viable alternative for the expensive setting which gives rooms for further improvements in this direction.

7.1.3. Recommendation systems

Bandits and Recommender Systems [23]

This paper addresses the on-line recommendation problem facing new users and new items; we assume that no information is available neither about users, nor about the items. The only source of information is a set of ratings given by users to some items. By on-line, we mean that the set of users, and the set of items, and the set of ratings is evolving along time and that at any moment, the recommendation system has to select items to recommend based on the currently available information, that is basically the sequence of past events. We also mean that each user comes with her preferences which may evolve along short and longer scales of time; so we have to continuously update their preferences. When the set of ratings is the only available source of information , the traditional approach is matrix factorization. In a decision making under uncertainty setting, actions should be selected to balance exploration with exploitation; this is best modeled as a bandit problem. Matrix factors provide a latent representation of users and items. These representations may then be used as contextual information by the bandit algorithm to select items. This last point is exactly the originality of this paper: the combination of matrix factorization and bandit algorithms to solve the on-line recommendation problem. Our work is driven by considering the recommendation problem as a feedback controlled loop. This leads to interactions between the representation learning, and the recommendation policy.

Collaborative Filtering as a Multi-Armed Bandit [35]

Recommender Systems (RS) aim at suggesting to users one or several items in which they might have interest. Following the feedback they receive from the user, these systems have to adapt their model in order to improve future recommendations. The repetition of these steps defines the RS as a sequential process. This sequential aspect raises an exploration-exploitation dilemma, which is surprisingly rarely taken into account for RS without contextual information. In this paper we present an explore-exploit collaborative filtering RS, based on Matrix Factor-ization and Bandits algorithms. Using experiments on artificial and real datasets, we show the importance and practicability of using sequential approaches to perform recommendation. We also study the impact of the model update on both the quality and the computation time of the recommendation procedure.

AUC Optimisation and Collaborative Filtering [39]

In recommendation systems, one is interested in the ranking of the predicted items as opposed to other losses such as the mean squared error. Although a variety of ways to evaluate rankings exist in the literature, here we focus on the Area Under the ROC Curve (AUC) as it widely used and has a strong theoretical underpinning. In practical recommendation, only items at the top of the ranked list are presented to the users. With this in mind, we propose a class of objective functions over matrix factorisations which primarily represent a smooth surrogate for the real AUC, and in a special case we show how to prioritise the top of the list. The objectives are differentiable and optimised through a carefully designed stochastic gradient-descent-based algorithm which scales linearly with the size of the data. In the special case of square loss we show how to improve computational complexity by leveraging previously computed measures. To understand theoretically the underlying matrix factorisation approaches we study both the consistency of the loss functions with respect to AUC, and generalisation using Rademacher theory. The resulting generalisation analysis gives strong motivation for the optimisation under study. Finally, we provide computation results as to the efficacy of the proposed method using synthetic and real data.

Collaborative Filtering with Localised Ranking [16]

In recommendation systems, one is interested in the ranking of the predicted items as opposed to other losses such as the mean squared error. Although a variety of ways to evaluate rankings exist in the literature, here we focus on the Area Under the ROC Curve (AUC) as it widely used and has a strong theoretical underpinning. In practical recommendation, only items at the top of the ranked list are presented to the users. With this in mind we propose a class of objective functions which primarily represent a smooth surrogate for the real AUC, and in a special case we show how to prioritise the top of the list. This loss is differentiable and is optimised through a carefully designed stochastic gradient-descent-based algorithm which scales linearly with the size of the data. We mitigate sample bias present in the data by sampling observations according to a certain power-law based distribution. In addition, we provide computation results as to the efficacy of the proposed method using synthetic and real data.

Collaborative Filtering with Stacked Denoising AutoEncoders and Sparse Inputs [36]

Neural networks have not been widely studied in Collaborative Filtering. For instance, no paper using neural networks was published during the Net-flix Prize apart from Salakhutdinov et al's work on Restricted Boltzmann Machine (RBM) [14]. While deep learning has tremendous success in image and speech recognition, sparse inputs received less attention and remains a challenging problem for neural networks. Nonetheless, sparse inputs are critical for collaborative filtering. In this paper, we introduce a neural network architecture which computes a non-linear matrix factorization from sparse rating inputs. We show experimentally on the movieLens and jester dataset that our method performs as well as the best collaborative filtering algorithms. We provide an implementation of the algorithm as a reusable plugin for Torch [4], a popular neural network framework.

7.1.4. Nonparametric statistics of time series

The Replacement Bootstrap for Dependent Data [31]

Applications that deal with time-series data often require evaluating complex statistics for which each time series is essentially one data point. When only a few time series are available, bootstrap methods are used to generate additional samples that can be used to evaluate empirically the statistic of interest. In this work a novel bootstrap method is proposed, which is shown to have some asymptotic consistency guarantees under the only assumption that the time series are stationary and ergodic. This contrasts previously available results that impose mixing or finite-memory assumptions on the data. Empirical evaluation on simulated and real data, using a practically relevant and complex extrema statistic is provided.

7.1.5. Imitation and Inverse Reinforcement Learning

Inverse Reinforcement Learning in Relational Domains [24]

In this work, we introduce the first approach to the Inverse Reinforcement Learning (IRL) problem in relational domains. IRL has been used to recover a more compact representation of the expert policy leading to better generalization performances among different contexts. On the other hand, rela-tional learning allows representing problems with a varying number of objects (potentially infinite), thus provides more generalizable representations of problems and skills. We show how these different formalisms allow one to create a new IRL algorithm for relational domains that can recover with great efficiency rewards from expert data that have strong generalization and transfer properties. We evaluate our algorithm in representative tasks and study the impact of diverse experimental conditions such as : the number of demonstrations, knowledge about the dynamics, transfer among varying dimensions of a problem, and changing dynamics.

Imitation Learning Applied to Embodied Conversational Agents [29]

Embodied Conversational Agents (ECAs) are emerging as a key component to allow human interact with machines. Applications are numerous and ECAs can reduce the aversion to interact with a machine by providing user-friendly interfaces. Yet, ECAs are still unable to produce social signals appropriately during their interaction with humans, which tends to make the interaction less instinctive. Especially, very little attention has been paid to the use of laughter in human-avatar interactions despite the crucial role played by laughter in human-human interaction. In this paper, methods for predicting when and how to laugh during an interaction for an ECA are proposed. Different Imitation Learning (also known as Apprenticeship Learning) algorithms are used in this purpose and a regularized classification algorithm is shown to produce good behavior on real data.

7.1.6. Stochastic Games

Optimism in Active Learning [3]

Active learning is the problem of interactively constructing the training set used in classification in order to reduce its size. It would ideally successively add the instance-label pair that decreases the classification error most. However, the effect of the addition of a pair is not known in advance. It can still be estimated with the pairs already in the training set. The online minimization of the classification error involves a tradeoff between exploration and exploitation. This is a common problem in machine learning for which multiarmed bandit, using the approach of Optimism int the Face of Uncertainty, has proven very efficient these last years. This paper introduces three algorithms for the active learning problem in classification using Optimism in the Face of Uncertainty. Experiments lead on built-in problems and real world datasets demonstrate that they compare positively to state-of-the-art methods.

Bayesian Credible Intervals for Online and Active Learning of Classification Trees [13]

Classification trees have been extensively studied for decades. In the online learning scenario, a whole class of algorithms for decision trees has been introduced, called incremental decision trees. In the case where subtrees may not be discarded, an incremental decision tree can be seen as a sequential decision process, consisting in deciding to extend the existing tree or not. This problem involves an trade-off between exploration and exploitation, which is addressed in recent work with the use of Hoeffding's bounds. This paper proposes to use Bayesian Credible Intervals instead, in order to get the most out of the knowledge of the output's distribution's shape. It also studies the case of Active Learning in such a tree following the Optimism in the Face of Uncertainty paradigm. Two novel algorithms are introduced for the online and active learning problems. Evaluations on real-world datasets show that these algorithms compare positively to state-of-the-art.

Optimism in Active Learning with Gaussian Processes [14]

In the context of Active Learning for classification, the classification error depends on the joint distribution of samples and their labels which is initially unknown. The minimization of this error requires estimating this distribution. Online estimation of this distribution involves a trade-off between exploration and exploitation. This is a common problem in machine learning for which multi-armed bandit theory, building upon Optimism in the Face of Uncertainty, has been proven very efficient these last years. We introduce two novel algorithms that use Optimism in the Face of Uncertainty along with Gaussian Processes for the Active Learning problem. The evaluation lead on real world datasets shows that these new algorithms compare positively to state-of-the-art methods.

Approximate Dynamic Programming for Two-Player Zero-Sum Markov Games [28]

This paper provides an analysis of error propagation in Approximate Dynamic Programming applied to zero-sum two-player Stochastic Games. We provide a novel and unified error propagation analysis in L p-norm of three well-known algorithms adapted to Stochastic Games (namely Approximate Value Iteration, Approximate Policy Iteration and Approximate Generalized Policy Iteratio,n). We show that we can achieve a stationary policy which is 2γ+ $(1–\gamma)$ 2-optimal, where is the value function approximation error and is the approximate greedy operator error. In addition , we provide a practical algorithm (AGPI-Q) to solve infinite horizon γ-discounted two-player zero-sum Stochastic Games in a batch setting. It is an extension of the Fitted-Q algorithm (which solves Markov Decisions Processes from data) and can be non-parametric. Finally, we demonstrate experimentally the performance of AGPI-Q on a simultaneous two-player game, namely Alesia.

7.2. Statistical analysis of time series

7.2.1. Automata Learning

Non-negative Spectral Learning for Linear Sequential Systems [18]

Method of moments (MoM) has recently become an appealing alternative to standard iterative approaches like Expectation Maximization (EM) to learn latent variable models. In addition, MoM-based algorithms come with global convergence guarantees in the form of finite sample bounds. However, given enough computation time, by using restarts and heuristics to avoid local optima, iterative approaches often achieve better performance. We believe that this performance gap is in part due to the fact that MoM-based algorithms can output negative probabilities. By constraining the search space, we propose a non-negative spectral algorithm (NNSpectral) avoiding computing negative probabilities by design. NNSpectral is compared to other MoM-based algorithms and EM on synthetic problems of the PAutomaC challenge. Not only, NNSpectral outperforms other MoM-based algorithms, but also, achieves very competitive results in comparison to EM.

Learning of scanning strategies for electronic support using predictive state representations [17]

In Electronic Support, a receiver must monitor a wide frequency spectrum in which threatening emitters operate. A common approach is to use sensors with high sensitivity but a narrow band-width. To maintain surveillance over the whole spectrum, the sensor has to sweep between frequency bands but requires a scanning strategy. Search strategies are usually designed prior to the mission using an approximate knowledge of illumination patterns. This often results in open-loop policies that cannot take advantage of previous observations. As pointed out in past researches, these strategies lack of robustness to the prior. We propose a new closed loop search strategy that learns a stochastic model of each radar using predic-tive state representations. The learning algorithm benefits from the recent advances in spectral learning and rank minimization using nuclear norm penalization.

Spectral learning with proper probabilities for finite state automation [19]

Probabilistic Finite Automaton (PFA), Probabilistic Finite State Transducers (PFST) and Hidden Markov Models (HMM) are widely used in Automatic Speech Recognition (ASR), Text-to-Speech (TTS) systems and Part Of Speech (POS) tagging for language mod-eling. Traditionally, unsupervised learning of these latent variable models is done by Expectation-Maximization (EM)-like algorithms, as the Baum-Welch algorithm. In a recent alternative line of work, learning algorithms based on spectral properties of some low order moments matrices or tensors were proposed. In comparison to EM, they are orders of magnitude faster and come with theoretical convergence guarantees. However, returned models are not ensured to compute proper distributions. They often return negative values that do not sum to one, limiting their applicability and preventing them to serve as an initialization to EM-like algorithms. In this paper, we propose a new spectral algorithm able to learn a large range of models constrained to return proper distributions. We assess its performances on synthetic problems from the PAutomaC challenge and real datasets extracted from Wikipedia. Experiments show that it outperforms previous spectral approaches as well as the Baum-Welch algorithm with random restarts, in addition to serve as an efficient initialization step to EM-like algorithms.

7.3. Statistical Learning and Bayesian Analysis

7.3.1. Prediction of Sequences of Structured and Unstructured Data

Operator-valued Kernels for Learning from Functional Response Data [6]

In this paper we consider the problems of supervised classification and regression in the case where attributes and labels are functions: a data is represented by a set of functions, and the label is also a function. We focus on the use of reproducing kernel Hilbert space theory to learn from such functional data. Basic concepts and properties of kernel-based learning are extended to include the estimation of function-valued functions. In this setting, the representer theorem is restated, a set of rigorously defined infinite-dimensional operator-valued kernels that can be valuably applied when the data are functions is described, and a learning algorithm for nonlinear functional data analysis is introduced. The methodology is illustrated through speech and audio signal processing experiments.

7.4. Applications

7.4.1. Software development

An Experimental Protocol for Analyzing the Accuracy of Software Error Impact Analysis [25]

In software engineering, error impact analysis consists in predicting the software elements (e.g. modules, classes, methods) potentially impacted by a change. Impact analysis is required to optimize the testing effort. In this paper we present a new protocol to analyze the accuracy of impact analysis. This protocol uses mutation testing to simulate changes that introduce errors. To this end, we introduce a variant of call graphs we name the ”use graph” of a software which may be computed efficiently. We apply this protocol to two open-source projects and correctly predict the impact of 30

A Learning Algorithm for Change Impact Prediction: Experimentation on 7 Java Applications [41]

Change impact analysis consists in predicting the impact of a code change in a software application. In this paper, we take a learning perspective on change impact analysis and consider the problem formulated as follows. The artifacts that are considered are methods of object-oriented software; the change under study is a change in the code of the method, the impact is the test methods that fail because of the change that has been performed. We propose an algorithm, called LCIP that learns from past impacts to predict future impacts. To evaluate our system, we consider 7 Java software applications totaling 214,000+ lines of code. We simulate 17574 changes and their actual impact through code mutations, as done in mutation testing. We find that LCIP can predict the impact with a precision of 69

7.4.2. *Spoken Dialogue Systems*

Human-Machine Dialogue as a Stochastic Game [10]

In this paper, an original framework to model human-machine spoken dialogues is proposed to deal with co-adaptation between users and Spoken Dialogue Systems in non-cooperative tasks. The conversation is modeled as a Stochastic Game: both the user and the system have their own preferences but have to come up with an agreement to solve a non-cooperative task. They are jointly trained so the Dialogue Manager learns the optimal strategy against the best possible user. Results obtained by simulation show that non-trivial strategies are learned and that this framework is suitable for dialogue modeling.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

- Jeremie Mary got a contract with Nuukik on the use of seasonality to improve recommender systems for e-commerce. This work won the price of the “Best data analysis” at “La nuit du commerce connecté” - http://www.retail-network.fr", 1500 participants, 80 projects in 5 categories.

8.2. Bilateral Grants with Industry

- Romain Warlop obtains a CIFRE grant with the start-up Fifty-Five and started his PhD in July under the supervision of Alessandro Lazaric, Jérémie Mary and Philippe Preux. The PhD is on the use of tensor and bandits techniques for recommender systems with a special focus on the cold start problem, and the non-stationarity of the environment.
- Nicolas Carrara obtains a CIFRE grant with Orange Labs and started his PhD in October under the supervision of Olivier Pietquin. The PhD topic is on transfer learning for fast adaption of spoken dialogue systems.

9. Partnerships and Cooperations

9.1. Regional Initiatives

Participant: Olivier Pietquin.

- *Title*: Sniper, Guerrilla, Shark, Razor et les autres
- *Type*: PICTANOVO
- *Coordinator*: Association P.A.S. (Emmanuelle Grangier)
- *Duration*: 2015
- *Abstract*:

  “Sniper, Guerrilla, Shark et les autres” is an interactive physical setting as well as a choreographic performance for four dancers /performers and two types of robots behaving as a swarm (some of them flying, others being on the floor). The context is high frequency trading from which emerges a world where human performers and non-humanoid robots live together. Their behaviour are depending on the same basic rules working at a non-temporal scale and a macro-temporal scale of share prices fluctuation.

9.2. National Initiatives

9.2.1. ANR ExTra-Learn

Participants: Alessandro Lazaric, Jérémie Mary, Rémi Munos, 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 focused on how to effectively transfer knowledge from an external expert as in apprenticeship learning. This is an important step towards automatic transfer because it digs into the problem of how knowledge of an expert can be integrated into the learning process. This investigation led to the publication of two papers at IJCAI’15. In 2015 a number of activities has also started. Ronan Fruit has been recruited for a PhD started in December. The main focus of the PhD will be related to transfer in multi-armed bandit, in particular in systems which are non-stationary where the task can change multiple times. Pierre-Victor Chaumier will start a long internship on transfer in RL with focus on applications to Atari games. Romain Warlop started in July a Cifre PhD (co-supervised by A. Lazaric, J. Mary, and Ph. Preux) with focus on how to use transfer learning in recommendation systems. We expect these activities to significantly advance the research in the project within 2016.

9.2.2. ANR KEHATH

Participant: Olivier Pietquin.

- *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.

9.2.3. ANR MaRDi

Participants: Olivier Pietquin, Bilal Piot.

- *Acronym*: MaRDi
- *Title*: Man-Robot Dialogue
- *Type*: ANR
- *Coordinator*: Univ. Lille 1 (Olivier Pietquin)
- *Duration*: 2012-2016
- *Other partners*: Laboratoire d'Informatique d'Avignon (LIA), CNRS - LAAS (Toulouse), Acapela group (Toulouse)
- *Abstract*: In the MaRDi project, we study the interaction between humans and machines as a situated problem in which human users and machines share the same environment. Especially, we investigate how the physical environment of robots interacting with humans can be used to improve the performance of spoken interaction which is known to be imperfect and sensible to noise. To achieve this objectif, we study three main problems. First, how to interactively build a multimodal representation of the current dialogue context from perception and proprioception signals. Second, how to automatically learn a strategy of interaction using methods such as reinforcement learning. Third, how to provide expressive feedbacks to users about how the machine is confident about its behaviour and to reflect its current state (also the physical state).

9.2.4. National Partners

- Inria Bordeaux - Sud-Ouest
  - B.Piot and O.Pietquin worked with T.Munzer and M.Lopes on Inverse Reinforcement Learning with Relational Domains. It led to a publication in IJCAI 2015 [24].
- CentraleSupélec
  - B.Piot and O.Pietquin worked with M.Geist on Inverse Reinforcement Learning with Relational Domains and Dialogue Management. It led to a conference publication in IJCAI 2015 [24] and a workshop publication in MLIS 2015 [29].
- Inria Nancy - Grand Est
  - J.Perolat, B.Piot and O.Pietquin worked with Bruno Scherrer on Stochastic Games. It led to a conference publication in ICML 2015 [28].
- CMLA - ENS Cachan.
  - Julien Audiffren *Collaborator*
    M. Valko, A. Lazaric, and M. Ghavamzadeh work with Julien on Semi-Supervised Apprenticeship Learning. We finalized and published a max-entropy algorithm that outperforms the approach without unlabeled data.
- LTCI, Institut Télécom-ParisTech, France.
  - Charanpal Dhanjal, Stefan Clemençon*Collaborator*
    Romaric Gaudel collaborates with Charanpal and Stefan since 2010 on topics related to *Matrix Factorization*. In the past we applied our work to sequential recommendation and to sequential clustering. This year, the collaboration has led to a publication in AAAI'15 conference [16].

9.3. European Initiatives

9.3.1. Collaborations in European Programs, except FP7 & H2020

9.3.1.1. CHIST-ERA IGLU

Participants: Olivier Pietquin, Bilal Piot, Jérémie Mary.

Program: CHIST-ERA

Project acronym: IGLU

Project title: Interactive Grounding of Language Generation

Duration: 10/2015 - 9/2018

Coordinator: Jean-Rouat (Univ. Sherbrooke)

Other partners: Univ. Lille, CRIStAL (France) - Inria, Flowers (France) - UMONS, Numédiart (Belgium) - KTH, TMH (Sweden) - Universidad de Zaragoza, I3A (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.

9.4. International Initiatives

9.4.1. Inria Associate Teams not involved in an Inria International Labs

9.4.1.1. CWI

In the end of 2015 SEQUEL started an Inria Associate team with CWI, Amsterdam. This project is called "Universal algorithms for sequential forecasting and bandit problems" and is led by Daniil Ryabko from the SEQUEL side, and by Peter Grunwald from the CWI side.

9.4.1.2. EduBand

Title: Educational Bandits

International Partner (Institution - Laboratory - Researcher):

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

Inria investigators: A. Lazaric, M. Valko

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.

9.4.2. Inria International Partners

9.4.2.1. Declared Inria International Partners

9.4.2.1.1. Montanuniverstat Leoben

Montanuniverstat Leoben (MUL), Austria, is an international partner of SEQUEL. The work in 2015 has been mostly on representation learning in reinforcement learning. The partnership involves Ronald Ortner and Peter Auer on the MUL side.

9.4.2.2. Informal International Partners

- University of California Irvine (USA)

  Anima Anandkumar *Collaborator*

  A. Lazaric collaborates with A. Anandkumar on the use of spectral methods for reinforcement learning.

- Politecnico di Milano (Italy)

  Nicola Gatti *Collaborator*

  A. Lazaric finalized a work with N. Gatti on the application of MAB on sponsored search auctions and mechanism design.

- Universität Potsdam (Germany)

  Alexandra Carpentier *Collaborator*

  M. Valko collaborates with A. Carpentier on scaling bandits to large dimensions and structures.

- Adobe Research, California

  Branislav Kveton *Collaborator*

  M. Valko and B. Kveton collaboration for sequential learning at recommendation for the entertainment content that features diversity.

- Boston University, USA

  Venkatesh Saligrama *Collaborator*

  M. Valko, R. Munos collaborated with V. Saligrama and M. Hanawal, on cost-effective spectral sensing, useful in radars.

9.5. International Research Visitors

9.5.1. Visits to International Teams

9.5.1.1. Sabbatical programme

Ryabko Daniil

Date: Jan 2014 - Jan 2015

Institution: CMM (Chile)

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific events organisation

10.1.1.1. Member of the organizing committees

Participation to the organization of the 32nd International Conference on Machine Learning:

- Philippe Preux, local chair organization of ICML
- Jérémie Mary, conference webmaster
- Romaric Gaudel, local volunteer chair

Co-organization of ICML workshops:

- 12th European Workshop on Reinforcement Learning (EWRL)
- 4th Workshop on Machine Learning for Interactive Systems
- Jeremie Mary was Co-organizer of the workshop “Offline and Online Evaluation of Web-based Services” at WWW’15 with Lihong Li (MSR) as main organizer.

10.1.2. Scientific events selection

10.1.2.1. Member of the conference program committees

- International Joint Conference on Artificial Intelligence (IJCAI 2015)
- International Conference on Pattern Recognition Applications and Methods (ICPRAM 2015)
- Approximate Dynamic Programing and Reinforcement Learning (ADPRL 2015)
- International Conference on Machine Learning (ICML 2015)
- Annual Conference on Neural Information Processing Systems (NIPS 2015)

French-speaking conferences:

- French Conference on Planning, Decision-making, and Learning in Control Systems (JFPDA 2015)
- Extraction et Gestion des Connaissances (EGC 2015)
- XXIIè rencontres de la société francophone de classification

10.1.2.2. Reviewer

- International Conference on Pattern Recognition Applications and Methods (ICPRAM 2015)
- Algorithmic Learning Theory (ALT 2015)
- AAAI Conference on Artificial Intelligence (AAAI 2015)
- Conference on Learning Theory (COLT 2015)
- European Workshop on Reinforcement Learning (EWRL 2015)
- Annual Conference on Neural Information Processing Systems (NIPS 2015)
- International Conference on Artificial Intelligence and Statistics (AISTATS 2015)
- European Conference on Machine Learning (ECML 2015)
- International Conference on Machine Learning (ICML 2015)
- International Joint Conferences on Artificial Intelligence (IJCAI 2015)
- Reinforcement Learning and Decision Making (RLDM 2015)
- International Conference on Uncertainty in Artificial Intelligence (UAI 2015)
- International Conference on Autonmous Agents and Multiagent Systems (AAMAS 2015)
- International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)

French-speaking conferences:

- French Conference on Planning, Decision-making, and Learning in Control Systems (JFPDA 2015)
- Conférence francophone sur l'Apprentissage Automatique (CAP 2015)

10.1.3. Journal

10.1.3.1. Member of the editorial boards

- Neurocomputing

10.1.3.2. Reviewer - Reviewing activities

- IEEE Signal Processing Letters
- IEEE Transactions on Information Theory
- IEEE Transactions on Neural Networks and Learning Systems
- Scandinavian Journal of Statistics
- Speech Communication
- Journal of Machine Learning Research
- Artificial Intelligence Journal
- Machine Learning Journal

- Journal of Artificial Intelligence Research

10.1.4. Invited talks

- Gergely Neu, invited talk at the “Data, Learning, and Inference” (DALI) workshop on Learning Theory, Spain, April 2015.
- Gergely Neu, invited talk at the “Learning Faster from Easy Data” NIPS workshop, Montreal, December 2015.
- Olivier Pietquin, NIPS Workshop on Spoken Language Understanding (SLU NIPS 2015).
- Michal Valko, invited talk at “LIX, École Polytechnique” April 2015
- A. Lazaric, *Open Questions in Transfer in RL*, “Machine Learning with Interdependent and Non-identically Distributed Data”, Dagsthul, Germany, April 2015.
- A. Lazaric, *Exploiting Easy Data in Online Optimization*, “Modal Seminar Series”, Inria Lille, May 2015.
- A. Lazaric, *Policy Search in Reinforcement Learning*, Criteo, Paris, June 2015.
- A. Lazaric, *Transfer in Multi-Armed Bandit*, Aston University, Birmingham, July 2015.
- A. Lazaric, *Transfer in Reinforcement Learning*, “Promotion et Developpement de l’Intelligence Artificielle”, Paris, October 2015.
- A. Lazaric, *The Hidden World of Bandits*, “Workshop on Sequential Learning and Applications”, Toulouse, November 2015.
- J. Mary, Invited talk at Euratechnologies on *recent advances of machine learning and deep learning for Sequential data*, Lille, November 2015.
- J. Mary, Invited talk at Recommender Days organized by CRITEO http://recommenders.fr/, December 2015.

10.1.5. Scientific expertise

- Agence Nationale pour la Recherche (ANR)
- Fonds National pour la Recherche Scientifique (FNRS), Belgium
- Olivier Pietquin and Philippe Preux are expert for H2020 European Program
- *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 Nancy (2015)
  - Selection committee for Inria award for scientific excellence (2015)
  - Selection committee for CR promotions (2015)
- Jérémie Mary is expert for the Research Council of Norway.
- *A. Lazaric* is a member of the committee for research evaluation (CER) at Inria Lille.
- *A. Lazaric* was a member of the hiring committee for junior researchers at Inria Lille (2015).

10.1.6. Research administration

- Philippe Preux is:
  - head of the DatInG (Data Intelligence Group) thematic group at CRIStAL that gathers 4 research groups, totaling more than 70 people,
  - member of the scientific committee of CRIStAL,
  - member of the Bureau du Comité des Projets at Inria Lille.
- Romaric Gaudel is:
  - board member of CRIStAL

  - manager of proml mailing list. This mailing list gathers French-speaking researchers from Machine Learning community.
- Olivier Pietquin is:
  - board member of CRIStAL
  - board member of the IEEA faculty at Univ. Lille 1
  - member of the computer science department of the Ecole Doctorale SPI
  - in charge of research and innovation for the computer science department of Univ. Lille 1

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Licence: R. Gaudel, 2015/2016 Spring: programmation R pour statistiques et sociologie quantitative, 28h eqTD, L1, université Lille 3, France

Licence: R. Gaudel, 2015/2016 Fall: préparation au C2i niveau 1, 24h eqTD, L1-3, université Lille 3, France

Licence: R. Gaudel, 2015/2016 Fall: travail collaboratif et à distance dans un monde numérique, 13h eqTD, L1-3 (enseignement à distance), université Lille 3, France

Master: M. Valko, 2014/2015 Spring: Graphs in Machine Learning, 27h eqTD, M2, ENS Cachan

Master: M. Valko, 2015/2016 Fall: Graphs in Machine Learning, 27h eqTD, M2, ENS Cachan

Master : A. Lazaric, Reinforcement Learning, 25h eqTD, M2, ENS Cachan, France

Master : A. Lazaric, Reinforcement Learning, 25h eqTD, M2, Ecole Centrale Lille, France

Summer school : A. Lazaric, Reinforcement Learning, 8h eqTD, Toulouse, France

Master: J. Mary, 2015/2016 Fall: Machine learning with R, 20h eqTD, M2, Ecole Centrale de Lille.

E-learning

SPOC: R. Gaudel, Marc Tommasi and Alain Preux, culture numérique S3, 8 semaines, Moodle, université Lille 3, licence (L1), formation initiale, tous les étudiants (> 3 000).

Ph. Preux:

- modeling and simulation of the dynamics of behavior, Master 1 in Psychology & master in Cognitive Science, Université de Lille 3
- Formal neural networks, Master 1 in Cognitive Science, Université de Lille 3
- Supervised Learning, Licence 3 MIASHS, Université de Lille 3
- Advanced Data Mining, master 2 MIASHS, Université de Lille 3
- Unsupervised learning, master 1 MIASHS, Université de Lille 3

C. Dimitrakakis:

- Web Fundamentals, Licence MIASHS, Université de Lille 3
- Supervised Learning, Master MIASHS, Université de Lille 3

B. Piot:

- Networks, Master SID, Université de Lille 3
- Databases, Licence MIASHS, Université de Lille 3
- Excel, Licence MIASHS, Université de Lille 3
- Databases, Master SIAD, Université de Lille 1
- Networks, Master SIAD, Université de Lille 1
- UML, Université de Lille 1

O. Pietquin:

- Machine learning, Master Informatique, Université Lille 1
- Machine learning and decision making, Master Informatique, Université Lille 1
- Bayesian signal processing, Engineering degree, Université de Mons (Belgique)

10.2.2. Supervision

Supervision of PhD:

- HDR: Jérémie Mary, Université de Lille 3, defended Nov 2015
- PhD: Amir Sani, Université de Lille 1, defended May 2015, Munos, Lazaric
- PhD: Marta Soare, Université de Lille 1, defended Dec 2015, Munos, Lazaric
- PhD in progress: Marc Abeille, since Sept. 2014, Munos, Lazaric
- PhD in progress: Merwan Barlier, since oct. 2014, Pietquin
- PhD in progress: Alexandre Bérard, since Oct. 2014, Pietquin
- PhD in progress: Daniele Calandriello, since Oct. 2014, Preux, Lazaric, Valko
- PnD in progress: Nicolas Carrara, since Oct. 2015, Pietquin
- PhD in progress: Ronan Fruit, since Dec. 2015, Ryabko, Lazaric
- PhD in progress: Pratik Gajane, since oct. 2014, Preux
- PhD in progress: Hadrien Glaude, since Feb. 2014, Pietquin
- PhD in progress: Jean-Bastien Grill, since Oct. 2014, Munos, Valko
- PhD in progress: Frédéric Guillou, since Oct. 2013, Preux, Mary, Gaudel
- PhD in progress: Tomáš Kocák, since Oct. 2013, Munos, Valko
- PhD in progress: Vincenzo Musco, since Nov. 2013, Preux, Monperrus
- PhD in progress: Julien Perolat, since Oct. 2014, Pietquin
- PhD in progress: Florian Strub, since Jan. 2016, Pietquin, Mary
- PhD in progress: Romain Warlop, since Sep. 2015, Preux, Mary, Lazaric

Management of diplomas:

- Ph. Preux is the head of the master in computer science “machine learning and data science”, Université de Lille 3.
- J. Mary is the head of the “Web analyst” track in master MIASHS, Université de Lille 3.
- head of the MoCAD master at Université Lille 1.

10.2.3. Juries

Ph. Preux has been member of the PhD juries:

- Manel Tagorti, Université de Lorraine,
- Yacine Nair Benrekia, Université de Nantes,
- El Mehdi Rochd, Université de Marseille.

Ph. Preux has been member of the HdR juries:

- Jérémie Mary, Université de Lille 3.

A. Lazaric has been member of the PhD juries:

- Rodrigue Talla Kuate, Aston University, Birmingham, UK.
- Kamyar Azzizade (PhD qualification), University of California Irvine, USA.

O. Pietquin has been member of the PhD juries:

- Nicolas Galichet, Université Paris-Saclay,
- Alaedine Mihoub, Université Grenoble-Alpes,
- Emmanuel Ferreira, Université d’Avignon et des Pays du Vaucluse.

10.3. Popularization

- Ph. Preux participates to a radio program on machine learning.
- Ph. Preux co-authors two papers on “Le Monde” binaire blog [38].
- Inria interview with N. Vayatis and M. Valko about teaching machine learning at ENS, July 2015.
- Rue89 interviewed M. Valko about machine learning at Inria, June 2015.
- Intel advertising face recognition software (that included the work of M. Valko), February 2015.
- M. Valko volunteered in teaching mathematics in “Association de la Clé”, that helps students from underprivileged backgrounds.
- Jérémie Mary was interviewed by "Ca m’intéresse" for a special issue on artificial intelligence.

11. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] A. SANI. *Machine Learning for Decision Making*, Université de Lille 1, May 2015, https://tel.archives-ouvertes.fr/tel-01256178

[2] M. SOARE. *Sequential Resource Allocation in Linear Stochastic Bandits* , Université Lille 1 - Sciences et Technologies, December 2015, https://hal.archives-ouvertes.fr/tel-01249224

Articles in International Peer-Reviewed Journals

[3] T. COLLET, O. PIETQUIN. *Optimism in Active Learning*, in "Computational Intelligence and Neuroscience", August 2015, https://hal.inria.fr/hal-01225798

[4] L. DEVROYE, G. LUGOSI, G. NEU. *Random-Walk Perturbations for Online Combinatorial Optimization*, in "IEEE Transactions on Information Theory", June 2015, vol. 61, n^o 7, pp. 4099 - 4106 [*DOI :* 10.1109/TIT.2015.2428253], https://hal.inria.fr/hal-01214987

[5] 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

[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)", 2015, https://hal.archives-ouvertes.fr/hal-01221329

[7] A. KHALEGHI, D. RYABKO. *Nonparametric multiple change point estimation in highly dependent time series*, in "Theoretical Computer Science", November 2015 [*DOI :* 10.1016/J.TCS.2015.10.041], https://hal.inria.fr/hal-01235330

[8] B. SCHERRER, M. GHAVAMZADEH, V. GABILLON, B. LESNER, M. GEIST. *Approximate Modified Policy Iteration and its Application to the Game of Tetris*, in "Journal of Machine Learning Research", 2015, vol. 16, 1629-1676 p. , A paraître, https://hal.inria.fr/hal-01091341

International Conferences with Proceedings

[9] J. AUDIFFREN, M. VALKO, A. LAZARIC, M. GHAVAMZADEH. *Maximum Entropy Semi-Supervised Inverse Reinforcement Learning*, in "International Joint Conference on Artificial Intelligence", Bueons Aires, Argentina, July 2015, https://hal.inria.fr/hal-01146187

[10] M. BARLIER, J. PEROLAT, R. LAROCHE, O. PIETQUIN. *Human-Machine Dialogue as a Stochastic Game*, in "16th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL 2015)", Prague, Czech Republic, September 2015, https://hal.inria.fr/hal-01225848

[11] A. CARPENTIER, M. VALKO. *Simple regret for infinitely many armed bandits*, in "International Conference on Machine Learning", Lille, France, July 2015, https://hal.inria.fr/hal-01153538

[12] J. CHEMALI, A. LAZARIC. *Direct Policy Iteration with Demonstrations*, in "IJCAI - 24th International Joint Conference on Artificial Intelligence", Buenos Aires, Argentina, July 2015, https://hal.inria.fr/hal-01237659

[13] T. COLLET, O. PIETQUIN. *Bayesian Credible Intervals for Online and Active Learning of Classification Trees*, in "ADPRL 2015 - Symposium on Adaptive Dynamic Programming and Reinforcement Learning", Cape Town, South Africa, Proceedings of the Symposium Series on Computational Intelligence, IEEE, December 2015, https://hal.inria.fr/hal-01225850

[14] T. COLLET, O. PIETQUIN. *Optimism in Active Learning with Gaussian Processes*, in "22nd International Conference on Neural Information Processing (ICONIP2015)", Istanbul, Turkey, November 2015, https://hal.inria.fr/hal-01225826

[15] B. DERBEL, P. PREUX. *Simultaneous Optimistic Optimization on the Noiseless BBOB Testbed*, in "The 17th IEEE Congress on Evolutionary Computation (CEC)", Sendai, Japan, May 2015, https://hal.inria.fr/hal-01246420

[16] C. DHANJAL, R. GAUDEL, S. CLÉMENÇON. *Collaborative Filtering with Localised Ranking*, in "Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI'15)", Austin, United States, January 2015, 7 p. , https://hal.inria.fr/hal-01255890

[17] H. GLAUDE, C. ENDERLI, J.-F. GRANDIN, O. PIETQUIN. *Learning of scanning strategies for electronic support using predictive state representations*, in "International Workshop on Machine Learning for Signal Processing (MLSP 2015)", Boston, United States, September 2015, https://hal.inria.fr/hal-01225807

[18] H. GLAUDE, C. ENDERLI, O. PIETQUIN. *Non-negative Spectral Learning for Linear Sequential Systems*, in "22nd International Conference on Neural Information Processing (ICONIP2015)", Istanbul, Turkey, November 2015, https://hal.inria.fr/hal-01225838

[19] H. GLAUDE, C. ENDERLI, O. PIETQUIN. *Spectral learning with proper probabilities for finite state automation*, in "ASRU 2015 - Automatic Speech Recognition and Understanding Workshop", Scottsdale, United States, Proceedings of the Automatic Speech Recognition and Understanding Workshop, IEEE, December 2015, https://hal.inria.fr/hal-01225810

[20] J.-B. GRILL, M. VALKO, R. MUNOS. *Black-box optimization of noisy functions with unknown smoothness*, in "Neural Information Processing Systems", Montréal, Canada, December 2015, https://hal.inria.fr/hal-01222915

[21] M. K. H. Hanawal, V. Saligrama, M. Valko, R. Munos. *Cheap Bandits*, in "International Conference on Machine Learning", Lille, France, 2015, https://hal.inria.fr/hal-01153540

[22] K. Lakshmanan, R. Ortner, D. Ryabko. *Improved Regret Bounds for Undiscounted Continuous Reinforcement Learning*, in "International Conference on Machine Learning (ICML)", Lille, France, July 2015, https://hal.inria.fr/hal-01165966

[23] J. Mary, R. Gaudel, P. Preux. *Bandits and Recommender Systems*, in "First International Workshop on Machine Learning, Optimization, and Big Data (MOD'15)", Taormina, Italy, Lecture Notes in Computer Science, Springer International Publishing, July 2015, vol. 9432, pp. 325-336 [*DOI :* 10.1007/978-3-319-27926-8_29], https://hal.inria.fr/hal-01256033

[24] T. Munzer, B. Piot, M. Geist, O. Pietquin, M. Lopes. *Inverse Reinforcement Learning in Relational Domains*, in "International Joint Conferences on Artificial Intelligence", Buenos Aires, Argentina, July 2015, https://hal.archives-ouvertes.fr/hal-01154650

[25] V. Musco, M. Monperrus, P. Preux. *An Experimental Protocol for Analyzing the Accuracy of Software Error Impact Analysis*, in "Tenth IEEE/ACM International Workshop on Automation of Software Test", Florence, Italy, May 2015, https://hal.inria.fr/hal-01120913

[26] G. Neu. *Explore no more: Improved high-probability regret bounds for non-stochastic bandits*, in "Advances on Neural Information Processing Systems 28 (NIPS 2015)", Montreal, Canada, December 2015, pp. 3150-3158, https://hal.inria.fr/hal-01223501

[27] G. Neu. *First-order regret bounds for combinatorial semi-bandits*, in "Proceedings of the 28th Annual Conference on Learning Theory (COLT)", Paris, France, JMLR Workshop and Conference Proceedings, July 2015, vol. 40, pp. 1360-1375, https://hal.inria.fr/hal-01215001

[28] J. Perolat, B. Scherrer, B. Piot, O. Pietquin. *Approximate Dynamic Programming for Two-Player Zero-Sum Markov Games*, in "International Conference on Machine Learning (ICML 2015)", Lille, France, July 2015, https://hal.inria.fr/hal-01153270

[29] B. Piot, M. Geist, O. Pietquin. *Imitation Learning Applied to Embodied Conversational Agents*, in "4th Workshop on Machine Learning for Interactive Systems (MLIS 2015)", Lille, France, J. Workshop, C. Proceedings (editors), July 2015, vol. 43, https://hal.inria.fr/hal-01225816

[30] D. Ryabko, B. Ryabko. *Predicting the outcomes of every process for which an asymptotically accurate stationary predictor exists is impossible*, in "International Symposium on Information Theory", Hong Kong, Hong Kong SAR China, IEEE, June 2015, pp. 1204-1206, https://hal.inria.fr/hal-01165876

[31] A. Sani, A. Lazaric, D. Ryabko. *The Replacement Bootstrap for Dependent Data*, in "Proceedings of the IEEE International Symposium on Information Theory", Hong Kong, Hong Kong SAR China, June 2015, https://hal.inria.fr/hal-01144547

[32] B. Szorenyi, R. Busa-Fekete, P. Weng, E. Hüllermeier. *Qualitative Multi-Armed Bandits: A Quantile-Based Approach*, in "Proceedings of The 32nd International Conference on Machine Learning, pp. 1660–1668, 2015", Lille, France, July 2015, https://hal.inria.fr/hal-01204708

[33] A. C. Y. Tossou, C. Dimitrakakis. *Algorithms for Differentially Private Multi-Armed Bandits*, in "AAAI 2016", Phoenix, Arizona, United States, February 2016, https://hal.inria.fr/hal-01234427

[34] Z. Zhang, B. Rubinstein, C. Dimitrakakis. *On the Differential Privacy of Bayesian Inference*, in "AAAI 2016", Phoenix, Arizona, United States, February 2016, https://hal.inria.fr/hal-01234215

Conferences without Proceedings

[35] F. Guillou, R. Gaudel, P. Preux. *Collaborative Filtering as a Multi-Armed Bandit*, in "NIPS'15 Workshop: Machine Learning for eCommerce", Montréal, Canada, December 2015, https://hal.inria.fr/hal-01256254

[36] F. Strub, J. Mary. *Collaborative Filtering with Stacked Denoising AutoEncoders and Sparse Inputs*, in "NIPS Workshop on Machine Learning for eCommerce", Montreal, Canada, December 2015, https://hal.inria.fr/hal-01256422

Scientific Books (or Scientific Book chapters)

[37] *A Relative Exponential Weighing Algorithm for Adversarial Utility-based Dueling Bandits*, 2015, vol. 37, pp. 218–227, https://hal.inria.fr/hal-01225614

Scientific Popularization

[38] P. Philippe, M. Tommasi, T. Vieville, C. De La Higuera. *L'apprentissage automatique : le diable n'est pas dans l'algorithme*, June 2015, Article sur http://binaire.blog.lemonde.fr, https://hal.inria.fr/hal-01246178

Other Publications

[39] C. Dhanjal, R. Gaudel, S. Clemencon. *AUC Optimisation and Collaborative Filtering*, August 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01185836

[40] E. Kaufmann. *On Bayesian index policies for sequential resource allocation*, January 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01251606

[41] V. Musco, A. Carette, M. Monperrus, P. Preux. *A Learning Algorithm for Change Impact Prediction: Experimentation on 7 Java Applications*, December 2015, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01248241

References in notes

[42] 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

[43] R. Bellman. *Dynamic Programming*, Princeton University Press, 1957

[44] D. Bertsekas, S. Shreve. *Stochastic Optimal Control (The Discrete Time Case)*, Academic Press, New York, 1978

[45] D. Bertsekas, J. Tsitsiklis. *Neuro-Dynamic Programming*, Athena Scientific, 1996

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

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

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

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