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

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2016

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

1. Members **1**
2. Overall Objectives **3**
3. Research Program **3**
3.1. In Short 3
3.2. Decision-making Under Uncertainty 4
3.2.1. Reinforcement Learning 4
3.2.2. Multi-arm Bandit Theory 6
3.3. Statistical analysis of time series 6
3.3.1. Prediction of Sequences of Structured and Unstructured Data 6
3.3.2. Hypothesis testing 7
3.3.3. Change Point Analysis 7
3.3.4. Clustering Time Series, Online and Offline 7
3.3.5. Online Semi-Supervised Learning 8
3.3.6. Online Kernel and Graph-Based Methods 8
4. Application Domains **8**
5. Highlights of the Year **8**
6. New Software and Platforms **9**
6.1. BAC 9
6.2. Collaborative Filtering Network 9
7. New Results **9**
7.1. Decision-making Under Uncertainty 9
7.1.1. Reinforcement Learning 9
7.1.2. Multi-arm Bandit Theory 10
7.1.3. Recommendation systems 12
7.1.4. Nonparametric statistics of time series 13
7.1.5. Imitation and Inverse Reinforcement Learning 13
7.1.6. Stochastic Games 14
7.2. Statistical analysis of time series 14
7.2.1. Change Point Analysis 14
7.2.2. Clustering Time Series, Online and Offline 14
7.2.3. Automata Learning 15
7.2.4. Online Kernel and Graph-Based Methods 15
7.3. Statistical Learning and Bayesian Analysis 15
7.3.1. Non-parametric methods for Function Approximation 15
7.3.2. Non-parametric methods for functional supervised learning 15
7.3.3. Differential privacy 16
7.4. Applications 16
7.4.1. Spoken Dialogue Systems 16
7.4.2. Software development 17
8. Bilateral Contracts and Grants with Industry **18**
9. Partnerships and Cooperations **19**
9.1. National Initiatives 19
9.1.1. ANR BoB 19
9.1.2. ANR Badass 20
9.1.3. ANR ExTra-Learn 20
9.1.4. ANR KEHATH 21
9.1.5. ANR MaRDi 22
9.1.6. National Partners 22
9.2. European Initiatives 22

9.2.1. FP7 & H2020 Projects 22
9.2.2. Collaborations in European Programs, Except FP7 & H2020 23
9.3. International Initiatives 24
9.3.1. Inria Associate Teams Not Involved in an Inria International Labs 24
9.3.2. Inria International Partners 24
9.3.2.1. With CWI 24
9.3.2.2. With University of Leoben 25
9.3.2.3. Informal International Partners 25
9.4. International Research Visitors 25
9.4.1. Visits of International Scientists 25
9.4.2. Visits to International Teams 25
10. Dissemination **26**
10.1. Promoting Scientific Activities 26
10.1.1. Scientific Events Organisation 26
10.1.2. Scientific Events Selection 26
10.1.2.1. Member of the Conference Program Committees 26
10.1.2.2. Reviewer 26
10.1.3. Journal 26
10.1.3.1. Member of the Editorial Boards 26
10.1.3.2. Reviewer - Reviewing Activities 26
10.1.4. Invited Talks 27
10.1.5. Scientific Expertise 28
10.1.6. Research Administration 28
10.2. Teaching - Supervision - Juries 28
10.2.1. Teaching 28
10.2.2. Supervision 29
10.2.3. Juries 30
10.3. Popularization 31
11. Bibliography **31**

Project-Team SEQUEL

Creation of the Project-Team: 2007 July 01

Keywords:

Computer Science and Digital Science:

3. - Data and knowledge
3.1. - Data
3.1.1. - Modeling, representation
3.1.4. - Uncertain data
3.3. - Data and knowledge analysis
3.3.1. - On-line analytical processing
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
4.8. - Privacy-enhancing technologies
5.1. - Human-Computer Interaction
8. - Artificial intelligence
8.2. - Machine learning
8.3. - Signal analysis
8.7. - AI algorithmics

Other Research Topics and Application Domains:

5.8. - Learning and training
6.1. - Software industry
6.1.1. - Software engineering
6.1.2. - Software evolution, maintenance
9.1.1. - E-learning, MOOC
9.4. - Sciences
9.4.5. - Data science

1. Members

Research Scientists

Emilie Kaufmann [CNRS, Researcher]
Alessandro Lazaric [Inria, Researcher]
Odalric Maillard [Inria, Researcher, moved to SEQUEL on 11/1/2016]

Rémi Munos [secondment at Google/Deepmind, Senior Researcher, HDR]
Daniil Ryabko [Inria, Researcher, HDR]
Michal Valko [Inria, Researcher, HDR]

Faculty Members

Philippe Preux [Team leader, Univ. Lille III, Professor, HDR]
Christos Dimitrakakis [Univ. Lille III, Associate Professor, HDR]
Romaric Gaudel [Univ. Lille III, Associate Professor]
Jérémie Mary [Univ. Lille III, Associate Professor, HDR]
Bilal Piot [Univ. Lille I, Associate Professor]
Olivier Pietquin [Univ. Lille I, Professor, currently in secondment at Google/Deepmind since May 2016, HDR]

PhD Students

Marc Abeille [Univ. Lille I]
Merwan Barlier [Orange Labs, granted by CIFRE]
Alexandre Berard [Univ. Lille I]
Lilian Besson [ENS Cachan, from Oct 2016]
Daniele Calandriello [Inria]
Ronan Fruit [Inria]
Pratik Gajane [Orange Labs, granted by CIFRE]
Guillaume Gautier [Inria and CNRS]
Jean-Bastien Grill [granted by ENS Paris]
Frédéric Guillou [Inria]
Tomáš Kocák [Inria, until Nov 2016]
Julien Perolat [Univ. Lille I]
Florian Strub [Univ. Lille I]
Romain Warlop [55]

Post-Doctoral Fellow

James Ridgway [Inria, from Oct 2016]

Visiting Scientists

Maryam Aziz [Northeastern University, from May 2016 until Aug 2016]
Kamyar Azizzadenesheli [University of California at Irvine, from Aug 2016]
Cricia Zilda Felicio Paixao [University Uberlandia, Brasil, until Aug 2016]
Yao Ma [University of Tokyo, Japan, until Mar 2016]
Aristide Tossou [Chalmers University, Sweden]

Administrative Assistant

Amelie Supervielle [Inria]

Others

Mehdi Abbana Bennani [Univ. Lille III, intern, from Jun 2016 until Aug 2016]
Remi Bardenet [CNRS, Researcher]
Pierre Chainais [Ecole Centrale de Lille, Associate professor, HDR]
Pierre-Victor Chaumier [Inria, intern, from Feb 2016 until Jun 2016]
Quentin Coët [Univ. Lille I, intern, from Mar 2016 until Aug 2016]
Jean-Benoît Delbrouck [Univ. Lille I, intern, from Mar 2016 until Aug 2016]
Eddy El Khatib [Univ. Lille I, intern, from Apr 2016 until Jul 2016]

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 [67]). 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 [65].

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 [64]) 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.

[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.

In order to maximize a given functional in a sequential framework, one usually applies Dynamic Programming (DP) [62], 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 [63]. Such an optimal policy is written π^*.
- We say that a (deterministic stationary) policy π is greedy with respect to (w.r.t.) some function V (defined on $\mathcal{X}$) if, for all $x \in \mathcal{X}$,

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

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

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

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

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

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

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

In short, we would like to mention that most of the reinforcement learning methods developed so far are built on one (or both) of the two following approaches ([68]):

- 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 [66], 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.* [61] 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 $x^1 = (x_1^1, \cdots, x_{n_1}^1), \cdots, x^N = (x_N^1, \cdots, x_{n_N}^N)$, we wish to group similar objects together. While this is of course not a precise formulation, it can be made precise if we assume that the samples were generated by k different distributions.

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

3.3.5. Online Semi-Supervised Learning

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

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

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

3.3.6. Online Kernel and Graph-Based Methods

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

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

4. Application Domains

4.1. Sequential decision making under uncertainty and prediction

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

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

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

5. Highlights of the Year

5.1. Highlights of the Year

- Grill, Valko & Munos gave an oral presentation at NIPS. Oral presentations at NIPS are rare: out of 2500+ submissions, only 1.8% are presented orally.

- Using a deep learning approach (sparse denoising autoencoders), Strub, Mary & Gaudel have obtained the best ever published results on the data from the Netflix challenge on recommendation systems. 10 years ago, such an achievement was worth 1M$.

6. New Software and Platforms

6.1. BAC

Bayesian Policy Gradient and Actor-Critic Algorithms
KEYWORDS: Machine learning - Incremental learning - Policy Learning
FUNCTIONAL DESCRIPTION

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

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

6.2. Collaborative Filtering Network

KEYWORDS: Recommender system - Neural networks - Deep learning
FUNCTIONAL DESCRIPTION

Recommendation systems advise users on which items (movies, musics, books etc.) they are more likely to be interested in. A good recommendation system may dramatically increase the amount of sales of a firm or retain customers. For instance, 80% of movies watched on Netflix come from the recommender system of the company. Colaborative Filtering (CF) aims at recommending an item to a user by predicting how a user would rate this item. To do so, the feedback of one user on some items is combined with the feedback of all other users on all items to predict a new rating. For instance, if someone rated a few books, CF objective is to estimate the ratings he would have given to thousands of other books by using the ratings of all the other readers.

The following module tackles Collaborative Filtering tasks by using a novel approach based on neural networks (sparse denoising autoencoders). In a few words, the module lets the user train neural networks to predict unknown entries in a history files.

The input files are classic csv files. The output files can either be the full matrix of ratings and/or the network weights. The root mean square error is computed to assess the quality of the training.

This module is based on Lua/Torch Framework. It works on both CPU/GPU and it is multithreaded.

- Contact: Florian Strub
- URL: https://github.com/fstrub95/Autoencoders_cf

7. New Results

7.1. Decision-making Under Uncertainty

7.1.1. Reinforcement Learning

Analysis of Classification-based Policy Iteration Algorithms, [20] We introduce a variant of the classification-based approach to policy iteration which uses a cost-sensitive loss function weighting each classification mistake by its actual regret, that is, the difference between the action-value of the greedy action and of the action chosen by the classifier. For this algorithm, we provide a full finite-sample analysis. Our results state a performance bound in terms of the number of policy improvement steps, the number of rollouts used in each iteration, the capacity of the considered policy space (classifier), and a capacity measure which indicates how well the policy space can approximate policies that are greedy with respect to any of its members. The analysis reveals a tradeoff between the estimation and approximation errors in this classification-based policy iteration setting. Furthermore it confirms the intuition that classification-based policy iteration algorithms could be favorably compared to value-based approaches when the policies can be approximated more easily than their corresponding value functions. We also study the consistency of the algorithm when there exists a sequence of policy spaces with increasing capacity.

Reinforcement Learning of POMDPs using Spectral Methods, [23]

We propose a new reinforcement learning algorithm for partially observable Markov decision processes (POMDP) based on spectral decomposition methods. While spectral methods have been previously employed for consistent learning of (passive) latent variable models such as hidden Markov models, POMDPs are more challenging since the learner interacts with the environment and possibly changes the future observations in the process. We devise a learning algorithm running through episodes, in each episode we employ spectral techniques to learn the POMDP parameters from a trajectory generated by a fixed policy. At the end of the episode, an optimization oracle returns the optimal memoryless planning policy which maximizes the expected reward based on the estimated POMDP model. We prove an order-optimal regret bound w.r.t. the optimal memoryless policy and efficient scaling with respect to the dimensionality of observation and action spaces.

Bayesian Policy Gradient and Actor-Critic Algorithms, [15]

Policy gradient methods are reinforcement learning algorithms that adapt a parameterized policy by following a performance gradient estimate. Many conventional policy gradient methods use Monte-Carlo techniques to estimate this gradient. The policy is improved by adjusting the parameters in the direction of the gradient estimate. Since Monte-Carlo methods tend to have high variance, a large number of samples is required to attain accurate estimates, resulting in slow convergence. In this paper, we first propose a Bayesian framework for policy gradient, based on modeling the policy gradient as a Gaussian process. This reduces the number of samples needed to obtain accurate gradient estimates. Moreover, estimates of the natural gradient as well as a measure of the uncertainty in the gradient estimates, namely, the gradient covariance, are provided at little extra cost. Since the proposed Bayesian framework considers system trajectories as its basic observable unit, it does not require the dynamics within trajectories to be of any particular form, and thus, can be easily extended to partially observable problems. On the downside, it cannot take advantage of the Markov property when the system is Markovian. To address this issue, we proceed to supplement our Bayesian policy gradient framework with a new actor-critic learning model in which a Bayesian class of non-parametric critics, based on Gaussian process temporal difference learning, is used. Such critics model the action-value function as a Gaussian process, allowing Bayes' rule to be used in computing the posterior distribution over action-value functions, conditioned on the observed data. Appropriate choices of the policy parameterization and of the prior covariance (kernel) between action-values allow us to obtain closed-form expressions for the posterior distribution of the gradient of the expected return with respect to the policy parameters. We perform detailed experimental comparisons of the proposed Bayesian policy gradient and actor-critic algorithms with classic Monte-Carlo based policy gradient methods, as well as with each other, on a number of reinforcement learning problems.

7.1.2. *Multi-arm Bandit Theory*

Improved Learning Complexity in Combinatorial Pure Exploration Bandits, [32]

We study the problem of combinatorial pure exploration in the stochastic multi-armed bandit problem. We first construct a new measure of complexity that provably characterizes the learning performance of the algorithms we propose for the fixed confidence and the fixed budget setting. We show that this complexity is never higher than the one in existing work and illustrate a number of configurations in which it can be significantly smaller.

While in general this improvement comes at the cost of increased computational complexity, we provide a series of examples , including a planning problem, where this extra cost is not significant.

Online learning with noisy side observations, [43]

We propose a new partial-observability model for online learning problems where the learner, besides its own loss, also observes some noisy feedback about the other actions, depending on the underlying structure of the problem. We represent this structure by a weighted directed graph, where the edge weights are related to the quality of the feedback shared by the connected nodes. Our main contribution is an efficient algorithm that guarantees a regret of $O(\sqrt{\alpha * T})$ after T rounds, where α * is a novel graph property that we call the effective independence number. Our algorithm is completely parameter-free and does not require knowledge (or even estimation) of α *. For the special case of binary edge weights, our setting reduces to the partial-observability models of Mannor & Shamir (2011) and Alon et al. (2013) and our algorithm recovers the near-optimal regret bounds.

Online learning with Erdös-Rényi side-observation graphs, [42]

We consider adversarial multi-armed bandit problems where the learner is allowed to observe losses of a number of arms beside the arm that it actually chose. We study the case where all non-chosen arms reveal their loss with an unknown probability rt, independently of each other and the action of the learner. Moreover, we allow rt to change in every round t, which rules out the possibility of estimating rt by a well-concentrated sample average. We propose an algorithm which operates under the assumption that rt is large enough to warrant at least one side observation with high probability. We show that after T rounds in a bandit problem with N arms, the expected regret of our algorithm is of order O(sqrt(sum(t=1)T (1/rt) log N)), given that rt less than log T / (2N-2) for all t. All our bounds are within logarithmic factors of the best achievable performance of any algorithm that is even allowed to know exact values of rt.

Revealing graph bandits for maximizing local influence, [27]

We study a graph bandit setting where the objective of the learner is to detect the most influential node of a graph by requesting as little information from the graph as possible. One of the relevant applications for this setting is marketing in social networks, where the marketer aims at finding and taking advantage of the most influential customers. The existing approaches for bandit problems on graphs require either partial or complete knowledge of the graph. In this paper, we do not assume any knowledge of the graph, but we consider a setting where it can be gradually discovered in a sequential and active way. At each round, the learner chooses a node of the graph and the only information it receives is a stochastic set of the nodes that the chosen node is currently influencing. To address this setting, we propose BARE, a bandit strategy for which we prove a regret guarantee that scales with the detectable dimension, a problem dependent quantity that is often much smaller than the number of nodes.

Algorithms for Differentially Private Multi-Armed Bandits, [50]

We present differentially private algorithms for the stochastic Multi-Armed Bandit (MAB) problem. This is a problem for applications such as adaptive clinical trials, experiment design, and user-targeted advertising where private information is connected to individual rewards. Our major contribution is to show that there exist (ϵ, δ) differentially private variants of Upper Confidence Bound algorithms which have optimal regret, $O(\epsilon^{-1} + \log T)$. This is a significant improvement over previous results, which only achieve poly-log regret $O(\epsilon^{-2} \log^2 T)$, because of our use of a novel interval-based mechanism. We also substantially improve the bounds of previous family of algorithms which use a continual release mechanism. Experiments clearly validate our theoretical bounds.

On the Complexity of Best Arm Identification in Multi-Armed Bandit Models, [17]

The stochastic multi-armed bandit model is a simple abstraction that has proven useful in many different contexts in statistics and machine learning. Whereas the achievable limit in terms of regret minimization is now well known, our aim is to contribute to a better understanding of the performance in terms of identifying the m best arms. We introduce generic notions of complexity for the two dominant frameworks considered in the literature: fixed-budget and fixed-confidence settings. In the fixed-confidence setting, we provide the first known distribution-dependent lower bound on the complexity that involves information-theoretic quantities and holds when m is larger than 1 under general assumptions. In the specific case of two armed-bandits, we derive refined lower bounds in both the fixed-confidence and fixed-budget settings, along with matching algorithms for Gaussian and Bernoulli bandit models. These results show in particular that the complexity of the fixed-budget setting may be smaller than the complexity of the fixed-confidence setting, contradicting the familiar behavior observed when testing fully specified alternatives. In addition, we also provide improved sequential stopping rules that have guaranteed error probabilities and shorter average running times. The proofs rely on two technical results that are of independent interest : a deviation lemma for self-normalized sums (Lemma 19) and a novel change of measure inequality for bandit models (Lemma 1).

Optimal Best Arm Identification with Fixed Confidence, [33]

We give a complete characterization of the complexity of best-arm identification in one-parameter bandit problems. We prove a new, tight lower bound on the sample complexity. We propose the 'Track-and-Stop' strategy, which we prove to be asymptotically optimal. It consists in a new sampling rule (which tracks the optimal proportions of arm draws highlighted by the lower bound) and in a stopping rule named after Chernoff, for which we give a new analysis.

On Explore-Then-Commit Strategies, [35]

We study the problem of minimising regret in two-armed bandit problems with Gaussian rewards. Our objective is to use this simple setting to illustrate that strategies based on an exploration phase (up to a stopping time) followed by exploitation are necessarily suboptimal. The results hold regardless of whether or not the difference in means between the two arms is known. Besides the main message, we also refine existing deviation inequalities, which allow us to design fully sequential strategies with finite-time regret guarantees that are (a) asymptotically optimal as the horizon grows and (b) order-optimal in the minimax sense. Furthermore we provide empirical evidence that the theory also holds in practice and discuss extensions to non-gaussian and multiple-armed case.

7.1.3. Recommendation systems

Scalable explore-exploit Collaborative Filtering, [39]

Recommender Systems (RS) aim at suggesting to users one or several items in which they might have interest. These systems have to update themselves as users provide new ratings, but also as new users/items enter the system. While this adaptation makes recommendation an intrinsically sequential task, most researches about RS based on Collaborative Filtering are omitting this fact, as well as the ensuing exploration/exploitation dilemma: should the system recommend items which bring more information about the users (explore), or should it try to get an immediate feedback as high as possible (exploit)? Recently, a few approaches were proposed to solve that dilemma, but they do not meet requirements to scale up to real life applications which is a crucial point as the number of items available on RS and the number of users in these systems explode. In this paper, we present an explore-exploit Collaborative Filtering RS which is both efficient and scales well. Extensive experiments on some of the largest available real-world datasets show that the proposed approach performs accurate personalized recommendations in less than a millisecond per recommendation, which makes it a good candidate for true applications.

Large-scale Bandit Recommender System, [38]

The main target of Recommender Systems (RS) is to propose to users one or several items in which they might be interested. However, as users provide more feedback, the recommendation process has to take these new data into consideration. The necessity of this update phase makes recommendation an intrinsically sequential task. A few approaches were recently proposed to address this issue, but they do not meet the need to scale up to real life applications. In this paper , we present a Collaborative Filtering RS method based on Matrix Factorization and Multi-Armed Bandits. This approach aims at good recommendations with a narrow computation time. Several experiments on large datasets show that the proposed approach performs personalized recommendations in less than a millisecond per recommendation.

Sequential Collaborative Ranking Using (No-)Click Implicit Feedback, [40] We study Recommender Systems in the context where they suggest a list of items to users. Several crucial issues are raised in such a setting: first, identify the relevant items to recommend; second, account for the feedback given by the user after he clicked and rated an item; third, since new feedback arrive into the system at any moment, incorporate such information to improve future recommendations. In this paper, we take these three aspects into consideration and present an approach handling click/no-click feedback information. Experiments on real-world datasets show that our approach outperforms state of the art algorithms.

Hybrid Recommender System based on Autoencoders, [49]

A standard model for Recommender Systems is the Matrix Completion setting: given partially known matrix of ratings given by users (rows) to items (columns), infer the unknown ratings. In the last decades, few attempts where done to handle that objective with Neural Networks, but recently an architecture based on Autoencoders proved to be a promising approach. In current paper, we enhanced that architecture (i) by using a loss function adapted to input data with missing values, and (ii) by incorporating side information. The experiments demonstrate that while side information only slightly improve the test error averaged on all users/items, it has more impact on cold users/items.

Compromis exploration-exploitation pour système de recommandation à grande échelle, [53]

Les systèmes de recommandation recommandent à des utilisateurs un ou des produits qui pourraient les intéresser. La recommandation se fonde sur les retours des utilisateurs par le passé, lors des précédentes recommandations. La recommandation est donc un problème séquentiel et le système de recommandation recommande (i) pour obtenir une bonne récompense, mais aussi (ii) pour mieux cerné l'utilisateur/les produits et ainsi obtenir de meilleures récompenses par la suite. Quelques approches récentes ciblent ce double objectif mais elles sont trop gourmandes en temps de calcul pour s'appliquer à certaines applications de la vie réelle. Dans cet article, nous présentons un système de recommandation fondé sur la factorisation de matrice et les bandits manchots. Plusieurs expériences sur de grandes base de données montrent que l'approche proposée fournit de bonnes recommendations en moins d'une milli-seconde par recommandation.

Filtrage Collaboratif Hybride avec des Auto-encodeurs, [54]

Le filtrage collaboratif (CF) exploite les retours des utilisateurs pour leur fournir des recommandations personnalisées. Lorsque ces algorithmes ont accès à des informations complémentaires, ils ont de meilleurs résultats et gèrent plus efficacement le démarrage à froid. Bien que les réseaux de neurones (NN) remportent de nombreux succès en traitement d'images, ils ont reçu beaucoup moins d'attention dans la communauté du CF. C'est d'autant plus surprenant que les NN apprennent comme les algorithme de CF une représentation latente des données. Dans cet article, nous introduisons une architecture de NN adaptée au CF (nommée CFN) qui prend en compte la parcimonie des données et les informations complémentaires. Nous montrons empiriquement sur les bases de données MovieLens et Douban que CFN bât l'état de l'art et profite des informations complémentaires. Nous fournissons une implémentation de l'algorithme sous forme d'un plugin pour Torch.

7.1.4. Nonparametric statistics of time series

Things Bayes can't do, [48]

The problem of forecasting conditional probabilities of the next event given the past is consideredin a general probabilistic setting. Given an arbitrary (large, uncountable) set C of predictors, we would like to construct a single predictor that performs asymptotically as well as the best predictor in C, on any data. Here we show that there are sets C for which such predictors exist, but none of them is a Bayesian predictor with a prior concentrated on C.In other words, there is a predictor with sublinear regret, but every Bayesian predictor must have a linear regret. This negative finding is in sharp contrast with previous resultsthat establish the opposite for the case when one of the predictors in C achieves asymptotically vanishing error.In such a case, if there is a predictor that achieves asymptotically vanishing error for any measure in C, then there is a Bayesian predictor that also has this property, and whose prior is concentrated on (a countable subset of) C.

7.1.5. Imitation and Inverse Reinforcement Learning

Score-based Inverse Reinforcement Learning, [29] This paper reports theoretical and empirical results obtained for the score-based Inverse Reinforcement Learning (IRL) algorithm. It relies on a non-standard setting for IRL consisting of learning a reward from a set of globally scored trajec-tories. This allows using any type of policy (optimal or not) to generate trajectories without prior knowledge during data collection. This way, any existing database (like logs of systems in use) can be scored a posteriori by an expert and used to learn a reward function. Thanks to this reward function, it is shown that a near-optimal policy can be computed. Being related to least-square regression, the algorithm (called SBIRL) comes with theoretical guarantees that are proven in this paper. SBIRL is compared to standard IRL algorithms on synthetic data showing that annotations do help under conditions on the quality of the trajectories. It is also shown to be suitable for real-world applications such as the optimisation of a spoken dialogue system.

7.1.6. Stochastic Games

Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning, [37]

You are a robot and you live in a Markov decision process (MDP) with a finite or an infinite number of transitions from state-action to next states. You got brains and so you plan before you act. Luckily, your roboparents equipped you with a generative model to do some Monte-Carlo planning. The world is waiting for you and you have no time to waste. You want your planning to be efficient. Sample-efficient. Indeed, you want to exploit the possible structure of the MDP by exploring only a subset of states reachable by following near-optimal policies. You want guarantees on sample complexity that depend on a measure of the quantity of near-optimal states. You want something, that is an extension of Monte-Carlo sampling (for estimating an expectation) to problems that alternate maximization (over actions) and expectation (over next states). But you do not want to StOP with exponential running time, you want something simple to implement and computationally efficient. You want it all and you want it now. You want TrailBlazer.

Maximin Action Identification: A New Bandit Framework for Games, [34]

We study an original problem of pure exploration in a strategic bandit model motivated by Monte Carlo Tree Search. It consists in identifying the best action in a game, when the player may sample random outcomes of sequentially chosen pairs of actions. We propose two strategies for the fixed-confidence setting: Maximin-LUCB, based on lower-and upper-confidence bounds; and Maximin-Racing, which operates by successively eliminating the sub-optimal actions. We discuss the sample complexity of both methods and compare their performance empirically. We sketch a lower bound analysis, and possible connections to an optimal algorithm.

7.2. Statistical analysis of time series

7.2.1. Change Point Analysis

Nonparametric multiple change point estimation in highly dependent time series, [18]

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.

7.2.2. Clustering Time Series, Online and Offline

Consistent Algorithms for Clustering Time Series, [19]

The problem of clustering is considered for the case where every point is a time series. The time series are either given in one batch (offline setting), or they are allowed to grow with time and new time series can be added along the way (online setting). We propose a natural notion of consistency for this problem, and show that there are simple, com-putationally efficient algorithms that are asymptotically consistent under extremely weak assumptions on the distributions that generate the data. The notion of consistency is as follows. A clustering algorithm is called consistent if it places two time series into the same cluster if and only if the distribution that generates them is the same. In the considered framework the time series are allowed to be highly dependent, and the dependence can have arbitrary form. If the number of clusters is known, the only assumption we make is that the (marginal) distribution of each time series is stationary ergodic. No parametric, memory or mixing assumptions are made. When the number of clusters is unknown, stronger assumptions are provably necessary, but it is still possible to devise nonparametric algorithms that are consistent under very general conditions. The theoretical findings of this work are illustrated with experiments on both synthetic and real data.

7.2.3. Automata Learning

PAC learning of Probabilistic Automaton based on the Method of Moments, [36]

Probabilitic Finite Automata (PFA) are gener-ative graphical models that define distributions with latent variables over finite sequences of symbols, a.k.a. stochastic languages. Traditionally , unsupervised learning of PFA is performed through algorithms that iteratively improves the likelihood like the Expectation-Maximization (EM) algorithm. Recently, learning algorithms based on the so-called Method of Moments (MoM) have been proposed as a much faster alternative that comes with PAC-style guarantees. However, these algorithms do not ensure the learnt automata to model a proper distribution , limiting their applicability and preventing them to serve as an initialization to iterative algorithms. In this paper, we propose a new MoM-based algorithm with PAC-style guarantees that learns automata defining proper distributions. We assess its performances on synthetic problems from the PAutomaC challenge and real datasets extracted from Wikipedia against previous MoM-based algorithms and EM algorithm.

7.2.4. Online Kernel and Graph-Based Methods

Analysis of Nyström method with sequential ridge leverage score sampling, [26]

Large-scale kernel ridge regression (KRR) is limited by the need to store a large kernel matrix Kt. To avoid storing the entire matrix Kt, Nystro¨m methods subsample a subset of columns of the kernel matrix, and efficiently find an approximate KRR solution on the reconstructed Kt . The chosen subsampling distribution in turn affects the statistical and computational tradeoffs. For KRR problems, [15, 1] show that a sampling distribution proportional to the ridge leverage scores (RLSs) provides strong reconstruction guarantees for Kt. While exact RLSs are as difficult to compute as a KRR solution, we may be able to approximate them well enough. In this paper, we study KRR problems in a sequential setting and introduce the INK-ESTIMATE algorithm, that incrementally computes the RLSs estimates. INK-ESTIMATE maintains a small sketch of Kt, that at each step is used to compute an intermediate estimate of the RLSs. First, our sketch update does not require access to previously seen columns, and therefore a single pass over the kernel matrix is sufficient. Second, the algorithm requires a fixed, small space budget to run dependent only on the effective dimension of the kernel matrix. Finally, our sketch provides strong approximation guarantees on the distance $||Kt - Kt||^2$, and on the statistical risk of the approximate KRR solution at any time, because all our guarantees hold at any intermediate step.

7.3. Statistical Learning and Bayesian Analysis

7.3.1. Non-parametric methods for Function Approximation

Pliable rejection sampling, [30]

Rejection sampling is a technique for sampling from difficult distributions. However, its use is limited due to a high rejection rate. Common adaptive rejection sampling methods either work only for very specific distributions or without performance guarantees. In this paper, we present pliable rejection sampling (PRS), a new approach to rejection sampling, where we learn the sampling proposal using a kernel estimator. Since our method builds on rejection sampling, the samples obtained are with high probability i.i.d. and distributed according to f. Moreover, PRS comes with a guarantee on the number of accepted samples.

7.3.2. Non-parametric methods for functional supervised learning

Operator-valued Kernels for Learning from Functional Response Data, [16] 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.

7.3.3. Differential privacy

On the Differential Privacy of Bayesian Inference, [51]

We study how to communicate findings of Bayesian inference to third parties, while preserving the strong guarantee of differential privacy. Our main contributions are four different algorithms for private Bayesian inference on proba-bilistic graphical models. These include two mechanisms for adding noise to the Bayesian updates, either directly to the posterior parameters, or to their Fourier transform so as to preserve update consistency. We also utilise a recently introduced posterior sampling mechanism, for which we prove bounds for the specific but general case of discrete Bayesian networks; and we introduce a maximum-a-posteriori private mechanism. Our analysis includes utility and privacy bounds, with a novel focus on the influence of graph structure on privacy. Worked examples and experiments with Bayesian naïve Bayes and Bayesian linear regression illustrate the application of our mechanisms.

Algorithms for Differentially Private Multi-Armed Bandits, [50]

We present differentially private algorithms for the stochastic Multi-Armed Bandit (MAB) problem. This is a problem for applications such as adaptive clinical trials, experiment design, and user-targeted advertising where private information is connected to individual rewards. Our major contribution is to show that there exist (ϵ, δ) differentially private variants of Upper Confidence Bound algorithms which have optimal regret, $O(\epsilon^{-1} + \log T)$. This is a significant improvement over previous results, which only achieve poly-log regret $O(\epsilon^{-2} \log^2 T)$, because of our use of a novel interval-based mechanism. We also substantially improve the bounds of previous family of algorithms which use a continual release mechanism. Experiments clearly validate our theoretical bounds.

7.4. Applications

7.4.1. Spoken Dialogue Systems

Compact and Interpretable Dialogue State Representation with Genetic Sparse Distributed Memory, [28]

t User satisfaction is often considered as the objective that should be achieved by spoken dialogue systems. This is why, the reward function of Spoken Dialogue Systems (SDS) trained by Reinforcement Learning (RL) is often designed to reflect user satisfaction. To do so, the state space representation should be based on features capturing user satisfaction characteristics such as the mean speech recognition confidence score for instance. On the other hand, for deployment in industrial systems, there is a need for state representations that are understandable by system engineers. In this paper, we propose to represent the state space using a Genetic Sparse Distributed Memory. This is a state aggregation method computing state prototypes which are selected so as to lead to the best linear representation of the value function in RL. To do so, previous work on Genetic Sparse Distributed Memory for classification is adapted to the Reinforcement Learning task and a new way of building the prototypes is proposed. The approach is tested on a corpus of dialogues collected with an appointment scheduling system. The results are compared to a grid-based linear parametrisation. It is shown that learning is accelerated and made more memory efficient. It is also shown that the framework is calable in that it is possible to include many dialogue features in the representation, interpret the resulting policy and identify the most important dialogue features.

A Stochastic Model for Computer-Aided Human-Human Dialogue, [24] In this paper we introduce a novel model for computer-aided human-human dialogue. In this context, the computer aims at improving the outcome of a human-human task-oriented dialogue by intervening during the course of the interaction. While dialogue state and topic tracking in human-human dialogue have already been studied, few work has been devoted to the sequential part of the problem, where the impact of the system's actions on the future of the conversation is taken into account. This paper addresses this issue by first modelling human-human dialogue as a Markov Reward Process. The task of purposely taking part into the conversation is then optimised within the Linearly Solvable Markov Decision Process framework. Utterances of the Conversational Agent are seen as perturbations in this process, which aim at satisfying the user's long-term goals while keeping the conversation natural. Finally, results obtained by simulation suggest that such an approach is suitable for computer-aided human-human dialogue and is a first step towards three-party dialogue.

Learning Dialogue Dynamics with the Method of Moments, [25]

In this paper, we introduce a novel framework to encode the dynamics of dialogues into a probabilistic graphical model. Traditionally, Hidden Markov Models (HMMs) would be used to address this problem, involving a first step of hand-crafting to build a dialogue model (e.g. defining potential hidden states) followed by applying expectation-maximisation (EM) algorithms to refine it. Recently, an alternative class of algorithms based on the Method of Moments (MoM) has proven successful in avoiding issues of the EM-like algorithms such as convergence towards local optima, tractability issues, initialization issues or the lack of theoretical guarantees. In this work, we show that dialogues may be modeled by SP-RFA, a class of graphical models efficiently learnable within the MoM and directly usable in planning algorithms (such as reinforcement learning). Experiments are led on the Ubuntu corpus and dialogues are considered as sequences of dialogue acts, represented by their Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). We show that a MoM-based algorithm can learn a compact model of sequences of such acts.

7.4.2. Software development

Mutation-Based Graph Inference for Fault Localization, [45]

We present a new fault localization algorithm, called Vautrin, built on an approximation of causality based on call graphs. The approximation of causality is done using software mutants. The key idea is that if a mutant is killed by a test, certain call graph edges within a path between the mutation point and the failing test are likely causal. We evaluate our approach on the fault localization benchmark by Steimann et al. totaling 5,836 faults. The causal graphs are extracted from 88,732 nodes connected by 119,531 edges. Vautrin improves the fault localization effectiveness for all subjects of the benchmark. Considering the wasted effort at the method level, a classical fault localization evaluation metric, the improvement ranges from 3

A Large-scale Study of Call Graph-based Impact Prediction using Mutation Testing, [21]

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

A Learning Algorithm for Change Impact Prediction, [44]

Change impact analysis (CIA) consists in predicting the impact of a code change in a software application. In this paper, the artifacts that are considered for CIA 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 LCIP, a learning algorithm that learns from past impacts to predict future impacts. To evaluate LCIP, we consider Java software applications that are strongly tested. We simulate 6000 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 74

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

- contract with “500px”; PI: Romaric Gaudel.

  Title: Recommender System for Photos

  Duration: May 2016 – Oct. 2016 (6 months)

  Abstract: Recommender Systems aim at recommending items to users. Advances in that field are targeting more and more personalized recommendation. From a recommendation based on market segment to a recommendation based on individual user taste. From a recommendation based on user’s information to a recommendation based on any feedback from any user. From a recommendation based on logged data to a recommendation including latest trends... 500px is a Canadian company which is part of this trend. 500px offers solutions to store pictures online, to share pictures, and to browse among pictures exhibited by other users. Given the huge amount of pictures stored by 500px, users need help to find pictures which corresponds to their tastes. 500px offers several tools to filter the content presented to users. But the tools allowing exploration of the pictures landscape are not personalized, the selection is mostly based on the popularity of pictures/galleries. The most personalized recommendations are obtained by following other users: you see recent pictures of that users. But such recommendations requires you (i) to discover by yourself relevant users, (ii) to explicitly tag these users. The aim of the project is to scan state of the art in Collaborative Filtering and to design a tool which recommends pictures to users based on their implicit actions: given the list of followed users, famed pictures, commented pictures, browsed pictures, ..., infer user’s tastes and recommend to that user pictures and/or other user to look at. The system would also make use of informations on the pictures and of user profiles.

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

  Title: Sequential Learning and Decision Making under Partial Monitoring

  Duration: Oct. 2014 – Sep. 2017

  Abstract: In applications such as recommendation systems, or computational advertising, the return collected from the user is partial: (s)he clicks on one item, or no item at all. We study this setting in which only a “partial” information is gathered in particular how to learn to behave optimaly in such a setting.

- contract with “55”; PI: Jérémie Mary

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

  Duration: Oct. 2015 – Sep. 2018

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

- contract with “What a nice place” ; PI: Jérémie Mary

  Title: Deduplication of pictures

  Duration: Mar. 2016 – Jan. 2017

  Abstract: “What is nice place” is a start up which aggregates products from different sources in order to provide some home staging advises. Uniqueness of presence for the items in their database can be hard to achieve because of the differences over names and variations of a product. Here we build a classification and deduplication system based on deep neural networks. In this contract we received support from Inria Tech and transferred them some knowledge about deep neural networks.

- contract with “What a nice place” and “Leroy Merlin”; PI: Jérémie Mary

  Title: New Shopping Experience - Virtual Coach

  Duration: Jun. 2016 – Fev. 2017

  Abstract: The goal of this project is to use pictures of house interiors in order to propose automatically some products which would fit in nicely. The relations are learnt automatically using deep neural networks and recommendation systems techniques. We made a first version which focuses on lamps which is available for demonstration at https://whataniceplace.leroymerlin.fr/

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. ANR BoB

Participant: Michal Valko.

- *Title*: Bayesian statistics for expensive models and tall data
- *Type*: National Research Agency
- *Coordinator*: CNRS (R. Bardenet)
- *Duration*: 2016-2020
- *Abstract*:

  Bayesian methods are a popular class of statistical algorithms for updating scientific beliefs. They turn data into decisions and models, taking into account uncertainty about models and their parameters. This makes Bayesian methods popular among applied scientists such as biologists, physicists, or engineers. However, at the heart of Bayesian analysis lie 1) repeated sweeps over the full dataset considered, and 2) repeated evaluations of the model that describes the observed physical process. The current trends to large-scale data collection and complex models thus raises two main issues. Experiments, observations, and numerical simulations in many areas of science nowadays generate terabytes of data, as does the LHC in particle physics for instance. Simultaneously, knowledge creation is becoming more and more data-driven, which requires new paradigms addressing how data are captured, processed, discovered, exchanged, distributed, and analyzed. For statistical algorithms to scale up, reaching a given performance must require as few iterations and as little access to data as possible. It is not only experimental measurements that are growing at a rapid pace. Cell biologists tend to have scarce data but large-scale models of tens of nonlinear differential equations to describe complex dynamics. In such settings, evaluating the model once requires numerically solving a large system of differential equations, which may take minutes for some tens of differential equations on today’s hardware. Iterative statistical processing that requires a million sequential runs of the model is thus out of the question. In this project, we tackle the fundamental cost-accuracy trade-off for Bayesian methods, in order to produce generic inference algorithms that scale favourably with the number of measurements in an experiment and the number of runs of a statistical model. We propose a collection of objectives with different risk-reward trade-offs to tackle these two goals. In particular, for experiments with large numbers of measurements, we further develop existing subsampling-based Monte Carlo methods, while developing a novel decision theory framework that includes data constraints. For expensive models, we build an ambitious programme around Monte Carlo methods that leverage determinantal processes, a rich class of probabilistic tools that lead to accurate inference with limited model evaluations. In short, using innovative techniques such as subsampling-based Monte Carlo and determinantal point processes, we propose in this project to push the boundaries of the applicability of Bayesian inference.

9.1.2. ANR Badass

Participants: Odalric Maillard, Emilie Kaufmann.

- *Title*:
- *Type*: National Research Agency
- *Coordinator*: Inria Lille (O. Maillard)
- *Duration*: 2016-2020
- *Abstract*: Motivated by the fact that a number of modern applications of sequential decision making require developing strategies that are especially robust to change in the stationarity of the signal, and in order to anticipate and impact the next generation of applications of the field, the BADASS project intends to push theory and application of MAB to the next level by incorporating non-stationary observations while retaining near optimality against the best not necessarily constant decision strategy. Since a non-stationary process typically decomposes into chunks associated with some possibly hidden variables (states), each corresponding to a stationary process, handling non-stationarity crucially requires exploiting the (possibly hidden) structure of the decision problem. For the same reason, a MAB for which arms can be arbitrary non-stationary processes is powerful enough to capture MDPs and even partially observable MDPs as special cases, and it is thus important to jointly address the issue of non-stationarity together with that of structure. In order to advance these two nested challenges from a solid theoretical standpoint, we intend to focus on the following objectives: *(i)* To broaden the range of optimal strategies for stationary MABs: current strategies are only known to be provably optimal in a limited range of scenarios for which the class of distribution (structure) is perfectly known; also, recent heuristics possibly adaptive to the class need to be further analyzed. *(ii)* To strengthen the literature on pure sequential prediction (focusing on a single arm) for non-stationary signals via the construction of adaptive confidence sets and a novel measure of complexity: traditional approaches consider a worst-case scenario and are thus overly conservative and non-adaptive to simpler signals. *(iii)* To embed the low-rank matrix completion and spectral methods in the context of reinforcement learning, and further study models of structured environments: promising heuristics in the context of e.g. contextual MABs or Predictive State Representations require stronger theoretical guarantees.

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

9.1.3. ANR ExTra-Learn

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

9.1.4. ANR KEHATH

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.

9.1.5. 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).

9.1.6. National Partners

- CentraleSupélec
  - J.Perolat, B.Piot and O.Pietquin worked with M.Geist on Stochastic Games. it led to a conference publication in ICML 2016.
- 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 AISTATS 2016 [47] and ICML 2016.
- Institut de Mathématiques de Toulouse
  - É. Kaufmann had publications at COLT, ALT and NIPS with Aurélie Garivier.

9.2. European Initiatives

9.2.1. FP7 & H2020 Projects

Program: H2020

Project acronym: BabyRobot

Project title: Child-Robot Communication and Collaboration

Duration: 01/2016 - 12/2018

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

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

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

9.2.2. Collaborations in European Programs, Except FP7 & H2020

Program: CHIST-ERA

Project acronym: IGLU

Project title: Interactively Grounded Language Understanding

Duration: 11/2015 - 10/2018

Coordinator: Jean Rouat (Université de Sherbrooke, Canada)

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

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

9.3. International Initiatives

9.3.1. Inria Associate Teams Not Involved in an Inria International Labs

9.3.1.1. EduBand

Title: Educational Bandits

International Partner (Institution - Laboratory - Researcher):

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

Start year: 2015

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

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

9.3.2. Inria International Partners

9.3.2.1. With CWI

Title: Learning theory

“North-European Associate Team”

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

Duration: 2016 - 2018

Start year: 2016

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

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

This collaboration aims at developing through the Inria International Laboratory with CWI.

9.3.2.2. With University of Leoben

Title: The multi-armed bandit problem

International Partner (Institution - Laboratory - Researcher):

University of Leoben (Austria) - Peter Auer

Duration: 2016 - 2016

Start year: 2016

ABSTRACT: Study of the multi-armed bandit problem.

9.3.2.3. 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.
- University of Lancaster (UK)

  Borja Balle *Collaborator*

  O-A. Maillard collaborates with B. Balle on concentration inequalities for Hankel matrices.

9.4. International Research Visitors

9.4.1. Visits of International Scientists

9.4.1.1. Internships

- Cricia Zilda Felicio Paixao, University Uberlandia, Brasil, Sep. 2015-Jul. 2016, working on recommendation systems in collaboration with Philippe Preux
- Maryam Aziz, Northeastern University, May-Aug. 2016, working on multi-armed bandits for clinical trials in collaboration with Emilie Kaufmann
- Kamyar Azizzadenesheli, University of California at Irvine, Aug-Oct. 2016, working on latent variable models for reinforcement learning in collaboration with Alessandro Lazaric
- Pierre-Victor Chaumier, Ecole Polytechnique, Jan-Jun. 2016, working on transfer learning in collaboration with Alessandro Lazaric
- Firas Jarboui, ENSTA ParisTech, France, May-July.@ 2016, working on Human-AI co-operation, in collaboration with Christos Dimitrakakis.

9.4.2. Visits to International Teams

9.4.2.1. Research Stays Abroad

- Christos Dimitrakakis visited SEAS, Harvard University, USA in the context of a Swedish/EU project “Market Mechanisms for Multiple Minds”, and the future of life institute project “Mechanism Design for Multiple AIs”, May-June, September-December 2016.
- Christos Dimitrakakis visited ETHZ, Switzerland, in the context of the Swiss SNSF project “Differential Privacy and Approximate Decision Making”, July-September 2016.

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. Member of the Organizing Committees

- C. Dimitrakakis, ICML Workshop on the theory and practice of differential privacy.
- Ph. Preux, "Big Data : Modelisation, Estimation and Selection", June 2016, Villeneuve d'Ascq.

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

- Conference on Learning Theory (COLT)
- International Joint Conference on Artificial Intelligence (IJCAI)
- European Conference on Machine Learning (ECML)
- ICPRAM
- French conferences:
  - Extraction et Gestion de Conaissances (EGC),
  - Journées Francophones de Planification, Décision, Apprentissage (JFPDA),
  - Apprentissage Automatique et Fouille de Données & Société Française de Classification

10.1.2.2. Reviewer

- Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)
- Conference on Learning Theory (COLT 2016)
- European Workshop on Reinforcement Learning (EWRL 2016)
- European Conference on Machine Learning (ECML 2016)
- International Conference on Machine Learning (ICML 2016)
- Neural Information Processing Systems (NIPS 2016)
- International Joint Conference on Artificial Intelligence (IJCAI 2016)
- Conference on Autonomous Agents and Multia-Agent Systems (AAMAS 2016)
- International Conference on Artificial Intelligence and Statistics (AISTATS 2016)
- French conferences:
  - Extraction et Gestion de Conaissances (EGC),
  - Journées Francophones de Planification, Décision, Apprentissage (JFPDA),
  - conférence francophone sur l'Apprentissage Automatique (CAp),
  - Apprentissage Automatique et Fouille de Données & Société Française de Classification
  - Conférence Nationale d'Intelligence Artificielle (CNIA)

10.1.3. Journal

10.1.3.1. Member of the Editorial Boards

- Journal of Games.
- Neurocomputing.
- Revue d'Intelligence Artificielle

10.1.3.2. Reviewer - Reviewing Activities

- Automatica

- Artificial Intelligence Journal
- Machine Learning Journal
- Journal of Artificial Intelligence Research
- Journal of Machine Learning Research
- AMS Mathematical Review
- IEEE Transaction on Signal Processing
- IEEE Tansaction on Cybernetics

10.1.4. Invited Talks

- R. Gaudel, *From Bandits to Recommender Systems*, Presented on September 29th, 2016 at ENSAI, Rennes, France
- R. Gaudel, *Recommendation as a sequential process*, Presented on December 12th, 2016 at CMLA Mathématiques Appliquées, Cachan, France
- E. Kaufmann, *The information complexity of best arm identification*, Multi-armed Bandit Workshop 2016 at STOR-i, Lancaster University, UK, January 2016.
- E. Kaufmann, *The information complexity of sequential resource allocation*, seminar of the Collegio Carlo Alberto, Turin, March 2016.
- E. Kaufmann, *Optimal Best Arm Identification with Fixed Confidence*, Workshop on Computational and Statistical Trade-offs in Learning , Institut des Hautes Etudes Scientifiques (Orsay), March 2016.
- E. Kaufmann, *The information complexity of sequential resource allocation*, Stalab seminar, University of Cambridge, UK, April 2016.
- E. Kaufmann, *Stratégies bayésiennes et fréquentistes dans un modèle de bandit*, 1er congrès de la Société Mathématique de France, Tours, June 2016.
- E. Kaufmann, *Stratégies bayésiennes et fréquentistes dans un modèle de bandit*, Journées MAS, Grenoble, August 2016.
- E. Kaufmann, *Revisiting the Exploration-Exploitation Tradeoff in Bandit Models*, Workshop on Optimization and Decision-Making in Uncertainty, Simons Institute, Berkeley, September 2016.
- A. Lazaric, *Spectral Methods for Learning in POMDPs*, University of Liège, Belgium, February 2016.
- A. Lazaric, *Spectral Methods for Learning in POMDPs*, CMLA Mathématiques Appliquées, Cachan, France, February 2016.
- A. Lazaric, *Incremental Kernel Regression with Ridge Leverage Score Sampling*, “Data Learning and Inference” (DALI), Sestri Levante, Italy, April 2016.
- A. Lazaric, *Optimism and Randomness in Linear Multi-armed Bandit*, “International Conference on Monte-Carlo Techniques”, July 2016.
- J. Mary, *Structured Bandits*, “University of Strasbourg”, May. 2016.
- J. Mary, *Tutorial on Deep Neural Networks*, “Journées Big Data”, by the Laboratoire Painlevé. Jun. 2016.
- J. Mary, *Machine Learning and AI*, “EDF Seminar”, Dec. 2016.
- O. Pietquin, *Closing the Interaction Loop with (Inverse) Reinforcement Learning*, Presented on November 15, 2016 at AWRL, Hamilton, New-Zealand
- O. Pietquin, *Challenges of End-to-End Spoken Dialogue Systems*, Presented on December 10, 2016 at FILM@NIPS Workshop, Barcelona, Spain
- O. Pietquin, *Keeping the Human in the Loop: Challenges for Machine Learning*, Presented on March 10, 2016 at Xerox Research Center in Europe, Grenoble, France

- M. Valko, *Spectral Methods for Learning in POMDPs*, University of Liège, Belgium, February 2016.
- M. Valko, *Where is Justin Bieber?*, Presented on September 22nd, 2016 at Comenius University in Bratislava, Slovakia (*FMFI 2016*)
- M. Valko, *Bandit learning*, Presented on September 15–19th, 2016 at Information technologies - Applications and Theory, at Tatranské Matliare, High Tatras, Slovakia (*ITAT 2016*)
- M. Valko, *Decision-making on graphs without graphs*, Presented on June 16-17th, 2016 at Graph-based Learning and Graph Mining workshop, at Inria Lille, France (*GBLGM 2016*)
- M. Valko, *Sequential learning on graphs with limited feedback*, Presented on May 11–13th, 2016 at Data Driven Approach to Networks and Language, at ENS Lyon, France (*NETSpringLyon 2016*)
- M. Valko, *Benefits of Graphs in Bandit Settings*, Presented on January 11–12th, 2016 at Multi-armed Bandit Workshop 2016 at STOR-i, Lancaster University, UK (*STOR-i 2016*)

10.1.5. Scientific Expertise

- Agence Nationale pour la Recherche (ANR)
- ANRT
- D2RT Ile de France
- Institut National de Recherche en Agronomie (INRA)
- Fonds National pour la Recherche Scientifique (FNRS), Belgium
- H2020 program
- *A. Lazaric* was a member of the hiring committee for junior researchers at Inria Lille (2016).
- *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 Sophia Antipolis (2016)
  - Selection committee for Inria award for scientific excellence, junior and senior (2016)
  - Selection committee for CR promotions (2016)
- Ph. Preux has chaired the hiring committee for an associate professor position at Université de Lille 3
- J. Mary was webpage chair for ICML'2016 in NYC

10.1.6. Research Administration

- Philippe Preux is:
  - Délégué Scientifique Adjoint (DSA) at Inria Lille
  - member of the Evaluation Committee (CE) at Inria
  - member of the Project Committee Board (BCP, Bureau du Comité des Projets) at Inria Lille
  - head of the “Data Intelligence” (DatInG) thematic group at CRIStAL.
  - member of the Scientific Committee of CRIStAL.
- R. Gaudel is member of the board of CRIStAL.
- R. Gaudel is manager of proml mailing list. This mailing list gathers French-speaking researchers from Machine Learning community.
- J. Mary is member of the "Commission Développement Technologique" at Inria Lille.

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Licence: C. Dimitrakakis, C2i, 25h eqTD, L1-2, Université de Lille 3, France.

Licence: C. Dimitrakakis, Traitement de données, Université de Lille 3, France.

Licence: C. Dimitrakakis, Modélisation de bases de données, Université de Lille 3, France.

Licence: C. Dimitrakakis, Fonctionnement des réseaux, Université de Lille 3, France.

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

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

Master: Ph. Preux, Advanced data mining, 30h eqTD, M2, Université de Lille 3, France

Master: Ph. Preux, Fundamental algorithms for data mining, 30h eqTD, M1, Université de Lille 3, France

Licence: Ph. Preux, Neural Networks, 28h eqTD, L3, Université de Lille 3, France

Licence: Ph. Preux, Graph Theory, 28h eqTD, L3, Université de Lille 3, France

Licence: Ph. Preux, C2i, 25h eqTD, L1-2, Université de Lille 3, France.

Master: M. Valko, 2016/2017 Fall: Graphs in Machine Learning, 27h eqTD, M2, ENS Cachan

Licence: R. Gaudel, 2016/2017 Spring: programmation R pour statistiques et sociologie quantitative, 44h eqTD, L1, université Lille 3, France

Licence: R. Gaudel, 2016/2017 Fall: préparation au C2i niveau 1, 30h eqTD, L1-3, université Lille 3, France

Licence: R. Gaudel, 2016/2017 Spring: préparation au C2i niveau 1, 25h eqTD, L1-3, université Lille 3, France

Licence: R. Gaudel, 2016/2017 Fall: travail collaboratif et à distance dans un monde numérique, 13h eqTD, L1-3 (enseignement à distance), université Lille 3, France

Licence: R. Gaudel, 2016/2017 Fall: algorithmes fondamentaux de la fouille de données, 30h eqTD, M1, université Lille 3, France

Licence: R. Gaudel, 2016/2017 Fall: Fouille de données avancée, 30h eqTD, M2, université Lille 3, France

Master: B. Piot, 2016/2017 Spring: Web Design, 60h eqTD, M2, Univ. Lille, France

Master: B. Piot, 2016/2017 Spring: Object Programming, 70h eqTD, M2, Univ. Lille, France

Master: B. Piot, 2016/2017 Fall: Web Design, 30h eqTD, M2, Univ. Lille, France

Master: B. Piot, 2016/2017 Fall: Object Programming, 22h eqTD, M2, Univ. Lille, France

Master: B. Piot, 2016/2017 Fall: Databases, 30h eqTD, M1, Univ. Lille, France

Master: J. Mary, 2016/2017 Fall: algorithmes fondamentaux de la fouille de données, 30h eqTD, M1, Univ. Lille, France

Master: J. Mary, 2016/2017 Fall: Programmation Web Avancée, 30h eqTD, M2, Univ. Lille, France

Master: J. Mary, 2016/2017 Fall: Reinforcement Learning, 16h eqTD, M2, Univ. Lille, France

10.2.2. Supervision

HdR: Michal Valko, Bandits on graphs and structures, ENS-Cachan, June 15th, 2016

PhD: Frédéric Guillou, On recommendation systems in sequential context, University of Lille, Dec. 2nd, 2016, advisors: Philippe Preux, Romaric Gaudel, Jérémie Mary

PhD: Tomas Kocak, Apprentissage séquentiel avec similitudes, University of Lille, Nov. 28th, 2016, advisor: Michal Valko, Rémi Munos

PhD: Hadrien Glaude, Méthodes des moments pour l'inférence de systèmes séquentiels linéaires rationnels, University of Lille, July. 8th, 2016, advisor: Olivier Pietquin

PhD in progress: Pratik Gajane, Sequential Learning and Decision Making under Partial Monitoring, University of Lille, started Oct. 2014, advisor: Philippe Preux

PhD in progress: Marc Abeille, Randomized Exploration-exploration Stretegies, University of Lille, started Oct. 2014, advisor: Alessandro Lazaric

PhD in progress: Merwan Barlier, Dialogues intelligents basés sur l'écoute de conversations homme/homme, University of Lille, started Oct. 2014, advisor: Olivier Pietquin

PhD in progress: Alexandre Berard, Learning from post-editing for machine translation, University of Lille, started Oct. 2014, advisor: Olivier Pietquin

PhD in progress: Lilian Besson, Apprentissage séquentiel multi-joueurs pour la radio intelligente, CentraleSupélec Rennes, started Oct. 2016, advisor: Emilie Kaufmann

PhD in progress: Reda Alami, Bandit à Mémoire pour la prise de décision en environnement dynamique, Orange LABS, University of Paris-Saclay, started Oct. 2016, advisor: Odalric-Ambrym Maillard, Raphaël Feraud

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

PhD in progress: Ronan Fruit, Transfer in Hierarchical Reinforcement Learning, University of Lille, started Dec. 2015, advisor: Alessandro Lazaric

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

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

PhD in progress: Julien Perolat, Reinforcement learning: the 2-player case, University of Lille, started Oct. 2014, advisor: Olivier Pietquin, Bilal Piot

PhD in progress: Florian Strub, Deep sequential learning and its application to human-robot interaction, University of Lille, started Jan. 2016, advisor: Olivier Pietquin, Jérémie Mary

PhD in progress: Romain Warlop, Novel Learning and Exploration-Exploitation Methods for Effective Recommender Systems, University of Lille, started Sep. 2015, advisor: Jérémie Mary

PhD in progress: Aristide Tossou, Privacy in Sequential Decision Making (provional), Chalmers, started Feb. 2015, advisor: Christos Dimitrakakis

10.2.3. Juries

PhD and hdr juries:

- E. Kaufmann: Marie-Liesse Cauwet, LRI, Orsay.
- A. Lazaric: Matteo Pirotta, Politecnico di Milano, Italy.
- J. Mary: examinator for Raphël Puget, université Paris 6.
- J. Mary: reviewer for Robin Allesiardo, université Paris Saclay
- Ph. Preux: reviewer for Hongliang Zhong, Laboratoire d'Informatique Fondamentale, Marseille
- Ph. Preux: president of the defense jury of the HDR of Matthieu Geist, Université de Lille
- O. Pietquin: advisor of the HDR of Matthieu Gesit, Université de Lille
- O. Pietquin: Hatim Khouzami, University of Avignon

PhD mid-term evaluation:

- A. Lazaric: Claire Vernade (mid-term evaluation), Telecom ParisTech, France.
- E. Kaufmann: opponent for the licenciate thesis of Stefan Magureanu, KTH, Stockholm, Sweden.

10.3. Popularization

- A. Lazaric was interviewed for Inria-Lille magazine (December issue).
- J. Mary gave a 55 min talk in front of students of Lycée Sainte-Famille at Amiens.
- Ph. Preux gave 2 talks on “Artificial Intelligence” in a high-school in Villeneuve d’Ascq within the “Fête de la science”.
- O. Pietquin was interviewed by France Culture (Supersonic) on March 23rd, 2016

11. Bibliography

Major publications by the team in recent years

[1] O. Cappé, A. Garivier, O.-A. Maillard, R. Munos, G. Stoltz. *Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation*, in "Annals of Statistics", 2013, vol. 41, n$^{\text{o}}$ 3, pp. 1516-1541, Accepted, to appear in Annals of Statistics, https://hal.archives-ouvertes.fr/hal-00738209

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

[3] 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

[4] M. Ghavamzadeh, Y. Engel, M. Valko. *Bayesian Policy Gradient and Actor-Critic Algorithms*, in "Journal of Machine Learning Research", January 2016, vol. 17, n$^{\text{o}}$ 66, pp. 1-53, https://hal.inria.fr/hal-00776608

[5] H. Kadri, E. Duflos, P. Preux, S. Canu, A. Rakotomamonjy, J. Audiffren. *Operator-valued Kernels for Learning from Functional Response Data*, in "Journal of Machine Learning Research (JMLR)", 2016, https://hal.archives-ouvertes.fr/hal-01221329

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

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

[8] R. Munos. *From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning*, 2014, 130 pages, https://hal.archives-ouvertes.fr/hal-00747575

[9] R. Ortner, D. Ryabko, P. Auer, R. Munos. *Regret bounds for restless Markov bandits*, in "Journal of Theoretical Computer Science (TCS)", 2014, vol. 558, pp. 62-76 [*DOI :* 10.1016/j.tcs.2014.09.026], https://hal.inria.fr/hal-01074077

[10] D. Ryabko, J. Mary. *A Binary-Classification-Based Metric between Time-Series Distributions and Its Use in Statistical and Learning Problems*, in "Journal of Machine Learning Research", 2013, vol. 14, pp. 2837-2856, https://hal.inria.fr/hal-00913240

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] H. GLAUDE. *Learning rational linear sequential systems using the method of moments*, Université de Lille 1 - Sciences et Technologies, July 2016, https://tel.archives-ouvertes.fr/tel-01374080

[12] F. GUILLOU. *On Recommendation Systems in a Sequential Context*, Université Lille 3, December 2016, https://tel.archives-ouvertes.fr/tel-01407336

[13] V. MUSCO. *Propagation Analysis based on Software Graphs and Synthetic Data*, Université Lille 3, November 2016, https://tel.archives-ouvertes.fr/tel-01398903

[14] M. VALKO. *Bandits on graphs and structures*, École normale supérieure de Cachan - ENS Cachan, June 2016, Habilitation à diriger des recherches, https://hal.inria.fr/tel-01359757

Articles in International Peer-Reviewed Journals

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

[16] H. KADRI, E. DUFLOS, P. PREUX, S. CANU, A. RAKOTOMAMONJY, J. AUDIFFREN. *Operator-valued Kernels for Learning from Functional Response Data*, in "Journal of Machine Learning Research (JMLR)", 2016, https://hal.archives-ouvertes.fr/hal-01221329

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

[18] A. KHALEGHI, D. RYABKO. *Nonparametric multiple change point estimation in highly dependent time series*, in "Theoretical Computer Science", 2016, vol. 620, pp. 119-133 [*DOI :* 10.1016/J.TCS.2015.10.041], https://hal.inria.fr/hal-01235330

[19] A. KHALEGHI, D. RYABKO, J. MARY, P. PREUX. *Consistent Algorithms for Clustering Time Series*, in "Journal of Machine Learning Research", 2016, vol. 17, n^o 3, pp. 1 - 32, https://hal.inria.fr/hal-01399613

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

[21] V. MUSCO, M. MONPERRUS, P. PREUX. *A Large-scale Study of Call Graph-based Impact Prediction using Mutation Testing*, in "Software Quality Journal", 2016 [*DOI :* 10.1007/S11219-016-9332-8], https://hal.inria.fr/hal-01346046

[22] G. NEU, B. GÁBOR. *Importance Weighting Without Importance Weights: An Efficient Algorithm for Combinatorial Semi-Bandits*, in "Journal of Machine Learning Research", August 2016, vol. 17, n^o 154, pp. 1 - 21, https://hal.archives-ouvertes.fr/hal-01380278

International Conferences with Proceedings

[23] K. Azizzadenesheli, A. Lazaric, A. Anandkumar. *Reinforcement Learning of POMDPs using Spectral Methods*, in "Proceedings of the 29th Annual Conference on Learning Theory (COLT2016)", New York City, United States, June 2016, https://hal.inria.fr/hal-01322207

[24] M. Barlier, R. Laroche, O. Pietquin. *A Stochastic Model for Computer-Aided Human-Human Dialogue*, in "Interspeech 2016", San Francisco, United States, September 2016, vol. 2016, pp. 2051 - 2055, https://hal.inria.fr/hal-01406894

[25] M. Barlier, R. Laroche, O. Pietquin. *Learning Dialogue Dynamics with the Method of Moments*, in "Workshop on Spoken Language Technologie (SLT 2016)", San Diego, United States, December 2016, https://hal.inria.fr/hal-01406904

[26] D. Calandriello, A. Lazaric, M. Valko. *Analysis of Nyström method with sequential ridge leverage score sampling*, in "Uncertainty in Artificial Intelligence", New York City, United States, June 2016, https://hal.inria.fr/hal-01343674

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

[28] L. El Asri, R. Laroche, O. Pietquin. *Compact and Interpretable Dialogue State Representation with Genetic Sparse Distributed Memory*, in "7th International Workshop on Spoken Dialogue Systems (IWSDS 2016)", Saariselka, Finland, January 2016, https://hal.inria.fr/hal-01406873

[29] L. El Asri, B. Piot, M. Geist, R. Laroche, O. Pietquin. *Score-based Inverse Reinforcement Learning*, in "International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)", Singapore, Singapore, May 2016, https://hal.inria.fr/hal-01406886

[30] A. Erraqabi, M. Valko, A. Carpentier, O.-A. Maillard. *Pliable rejection sampling*, in "International Conference on Machine Learning", New York City, United States, June 2016, https://hal.inria.fr/hal-01322168

[31] C. Z. Felício, K. V. R. Paixão, C. A. Z. Barcelos, P. Preux. *Preference-like Score to Cope with Cold-Start User in Recommender Systems*, in "28th International Conference on Tools with Artificial Intelligence (ICTAI)", San Jose, United States, Proceedings of the IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), November 2016, https://hal.inria.fr/hal-01390762

[32] V. Gabillon, A. Lazaric, M. Ghavamzadeh, R. Ortner, P. Bartlett. *Improved Learning Complexity in Combinatorial Pure Exploration Bandits*, in "Proceedings of the 19th International Conference on Artificial Intelligence (AISTATS)", Cadiz, Spain, May 2016, https://hal.inria.fr/hal-01322198

[33] A. Garivier, E. Kaufmann. *Optimal Best Arm Identification with Fixed Confidence*, in "29th Annual Conference on Learning Theory (COLT)", New York, United States, JMLR Workshop and Conference Proceedings, June 2016, vol. 49, https://hal.archives-ouvertes.fr/hal-01273838

[34] A. Garivier, E. Kaufmann, W. M. Koolen. *Maximin Action Identification: A New Bandit Framework for Games*, in "29th Annual Conference on Learning Theory (COLT)", New-York, United States, JMLR Workshop and Conference Proceedings, June 2016, vol. 49, https://hal.archives-ouvertes.fr/hal-01273842

[35] A. GARIVIER, E. KAUFMANN, T. LATTIMORE. *On Explore-Then-Commit Strategies*, in "NIPS", Barcelona, Spain, Advances in Neural Information Processing Systems (NIPS), December 2016, vol. 29, https://hal.archives-ouvertes.fr/hal-01322906

[36] H. GLAUDE, O. PIETQUIN. *PAC learning of Probabilistic Automaton based on the Method of Moments*, in "International Conference on Machine Learning (ICML 2016)", New York, United States, June 2016, https://hal.inria.fr/hal-01406889

[37] J.-B. GRILL, M. VALKO, R. MUNOS. *Blazing the trails before beating the path: Sample-efficient Monte-Carlo planning*, in "NIPS 2016 - Thirtieth Annual Conference on Neural Information Processing Systems", Barcelona, Spain, December 2016, https://hal.inria.fr/hal-01389107

[38] F. GUILLOU, R. GAUDEL, P. PREUX. *Large-scale Bandit Recommender System*, in "the 2nd International Workshop on Machine Learning, Optimization and Big Data (MOD'16)", Volterra, Italy, August 2016, https://hal.inria.fr/hal-01406389

[39] F. GUILLOU, R. GAUDEL, P. PREUX. *Scalable explore-exploit Collaborative Filtering*, in "Pacific Asia Conference on Information Systems (PACIS'16)", Chiayi, Taiwan, 2016, https://hal.inria.fr/hal-01406418

[40] F. GUILLOU, R. GAUDEL, P. PREUX. *Sequential Collaborative Ranking Using (No-)Click Implicit Feedback*, in "The 23rd International Conference on Neural Information Processing (ICONIP'16)", Kyoto, Japan, Lecture Notes in Computer Science, October 2016, vol. 9948, pp. 288 - 296 [*DOI :* 10.1007/978-3-319-46672-9_33], https://hal.inria.fr/hal-01406338

[41] E. KAUFMANN, T. BONALD, M. LELARGE. *A Spectral Algorithm with Additive Clustering for the Recovery of Overlapping Communities in Networks*, in "ALT 2016 - Algorithmic Learning Theory", Bari, Italy, R. ORTNER, H. U. SIMON, S. ZILLES (editors), Lecture Notes in Computer Science, Springer, October 2016, vol. 9925, pp. 355-370 [*DOI :* 10.1007/978-3-319-46379-7_24], https://hal.archives-ouvertes.fr/hal-01163147

[42] T. KOCÁK, G. NEU, M. VALKO. *Online learning with Erdős-Rényi side-observation graphs*, in "Uncertainty in Artificial Intelligence", New York City, United States, June 2016, https://hal.inria.fr/hal-01320588

[43] T. KOCÁK, G. NEU, M. VALKO. *Online learning with noisy side observations*, in "International Conference on Artificial Intelligence and Statistics", Seville, Spain, May 2016, https://hal.inria.fr/hal-01303377

[44] V. MUSCO, A. CARETTE, M. MONPERRUS, P. PREUX. *A Learning Algorithm for Change Impact Prediction*, in "5th International Workshop on Realizing Artificial Intelligence Synergies in Software Engineering", Austin, United States, May 2016, https://hal.inria.fr/hal-01279620

[45] V. MUSCO, M. MONPERRUS, P. PREUX. *Mutation-Based Graph Inference for Fault Localization*, in "International Working Conference on Source Code Analysis and Manipulation", Raleigh, United States, October 2016, https://hal.inria.fr/hal-01350515

[46] J. PÉROLAT, B. PIOT, M. GEIST, B. SCHERRER, O. PIETQUIN. *Softened Approximate Policy Iteration for Markov Games*, in "ICML 2016 - 33rd International Conference on Machine Learning", New York City, United States, June 2016, https://hal.inria.fr/hal-01393328

[47] J. PÉROLAT, B. PIOT, B. SCHERRER, O. PIETQUIN. *On the Use of Non-Stationary Strategies for Solving Two-Player Zero-Sum Markov Games*, in "19th International Conference on Artificial Intelligence and Statistics (AISTATS 2016)", Cadiz, Spain, Proceedings of the International Conference on Artificial Intelligences and Statistics, May 2016, https://hal.inria.fr/hal-01291495

[48] D. RYABKO. *Things Bayes can't do*, in "Proceedings of the 27th International Conference on Algorithmic Learning Theory (ALT'16)", Bari, Italy, October 2016, vol. LNCS, n^{O} 9925, pp. 253-260 [*DOI :* 10.1007/978-3-319-46379-7_17], https://hal.inria.fr/hal-01380063

[49] F. STRUB, R. GAUDEL, J. MARY. *Hybrid Recommender System based on Autoencoders*, in "the 1st Workshop on Deep Learning for Recommender Systems", Boston, United States, September 2016, pp. 11 - 16 [*DOI :* 10.1145/2988450.2988456], https://hal.inria.fr/hal-01336912

[50] 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

[51] 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

[52] A. BÉRARD, C. SERVAN, O. PIETQUIN, L. BESACIER. *MultiVec: a Multilingual and Multilevel Representation Learning Toolkit for NLP*, in "The 10th edition of the Language Resources and Evaluation Conference (LREC)", Portoroz, Slovenia, May 2016, https://hal.archives-ouvertes.fr/hal-01335930

[53] F. GUILLOU, R. GAUDEL, P. PREUX. *Compromis exploration-exploitation pour système de recommandation à grande échelle*, in "Conférence francophone sur l'Apprentissage Automatique (CAp'16)", Marseille, France, July 2016, https://hal.inria.fr/hal-01406439

[54] F. STRUB, J. MARY, R. GAUDEL. *Filtrage Collaboratif Hybride avec des Auto-encodeurs*, in "Conférence francophone sur l'Apprentissage Automatique (CAp'16)", Marseille, France, July 2016, https://hal.inria.fr/hal-01406432

Research Reports

[55] B. DANGLOT, P. PREUX, B. BAUDRY, M. MONPERRUS. *Correctness Attraction: A Study of Stability of Software Behavior Under Runtime Perturbation*, HAL, 2016, n^{O} hal-01378523, https://hal.archives-ouvertes.fr/hal-01378523

Other Publications

[56] S. BUBECK, R. ELDAN, J. LEHEC. *Sampling from a log-concave distribution with Projected Langevin Monte Carlo*, January 2017, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01428950

[57] C. DIMITRAKAKIS, F. JARBOUI, D. PARKES, L. SEEMAN. *Multi-view Sequential Games: The Helper-Agent Problem*, December 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01408294

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

[59] A. R. LUEDTKE, E. KAUFMANN, A. CHAMBAZ. *Asymptotically Optimal Algorithms for Multiple Play Bandits with Partial Feedback*, June 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01338733

[60] F. STRUB, J. MARY, R. GAUDEL. *Hybrid Collaborative Filtering with Autoencoders*, July 2016, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01281794

References in notes

[61] 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

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

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

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

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

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

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

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