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

Ecole Centrale de Lille

Activity Report 2014

Project-Team SEQUEL

Sequential Learning

IN COLLABORATION WITH: Laboratoire d'informatique fondamentale de Lille (LIFL), Laboratoire d'Automatique, de Génie Informatique et Signal (LAGIS)

RESEARCH CENTER
Lille - Nord Europe

THEME
Optimization, machine learning and statistical methods

Table of contents

Project-Team SEQUEL

Keywords: Machine Learning, Statistical Learning, Sequential Learning, Inference, Analysis Of Algorithms

Creation of the Project-Team: 2007 July 01.

1. Members

Research Scientists

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

Faculty Members

Philippe Preux [Team leader, Univ. Lille III, Professor, HdR]
Pierre Chainais [Ecole Centrale de Lille, Associate Professor, HdR]
Rémi Coulom [Univ. Lille III, Associate Professor, until Sep 2014]
Emmanuel Duflos [Ecole Centrale de Lille, Professor, HdR]
Romaric Gaudel [Univ. Lille III, Associate Professor]
Jérémie Mary [Univ. Lille III, Associate Professor]
Philippe Vanheeghe [Ecole Centrale de Lille, Professor, HdR]
Bilal Piot [Univ. Lille III, from Oct 2014]
Olivier Pietquin [IUF and Univ. Lille I, HdR]

PhD Students

Timothé Collet [Univ. Metz, from Mar 2014 until Nov 2014]
Layla El Asri [CIFRE Orange]
Marc Abeille [Univ. Lille I, from Sep 2014]
Boris Baldassari [Squoring, until Sep 2014]
Alexandre Berard [Univ. Lille I, from Oct 2014]
Daniele Calandriello [Inria]
Pratik Gajane [Orange Labs, from Oct 2014]
Hadrien Glaude [Thales, from Feb 2014]
Jean Bastien Grill [ENS Cachan, from Oct 2014]
Frédéric Guillou [Région Nord-Pas-de-Calais and TBS]
Adrien Hoarau [DGA]
Tomas Kocak [Inria]
Julien Perolat [Univ. Lille I, from Oct 2014]
Amir Sani [Région Nord-Pas-de-Calais and Inria]
Marta Soare [Région Nord-Pas-de-Calais and EU FP7 Complacs]
Olivier Nicol [Univ. Lille I and III, until Dec 2014]
Victor Gabillon [MENRT, Univ. Lille I, until June 2014]
Vincenzo Musco [Univ. Lille I and III]
Hong-Phuong Dang [Ecole Centrale Lille]
Clément Elvira [ANR BNPSI no ANR-13-BS-03-0006-01]

Post-Doctoral Fellows

Prashanth Lakshmanrao Anantha Padmanabha [Inria, until Oct 2014, granted by EU FP7 Complacs]
Gergely Neu [Inria, granted by ERCIM and Hermès]
Balázs Szörényi [Inria, granted by EU FP7 Complacs]

Visiting Scientists
Sergio Valcarcel Macua [Technical University of Madrid, from Feb 2014 until May 2014]
Jennifer Healey [Intel]
Julien Audiffren [Univ. Marseille]

Administrative Assistant
Amélie Supervielle [Inria]

Others
Nicolas Carion [ENS-Lyon, L3, from Jun 2014 until Jul 2014]
Jessica Chemali [Inria, Master, Carnegie Mellon University, from May 2014 until Aug 2014]
Valentin Owczarek [Univ. Lille I, L3, from Apr 2014 until Jun 2014]
Julien Rouse [Univ. Lille III, L3, from Mar 2014 until Jun 2014]
Mathias Sable Meyer [ENS-Cachan, L3, from Jun 2014 until Jul 2014]
Othmane Safsafi [ENS-Ulm, L3, from Jun 2014 until Aug 2014]

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 [52]). 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 [48].

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

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

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

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

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

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

where $\mathbb{E}$ is the expectation operator and $\gamma \in (0,1)$ is the discount factor. This value function V^π gives an evaluation of the performance of a given policy π. Other functionals of the sequence of future rewards may be considered, such as the undiscounted reward (see the stochastic shortest path problems [43]) and average reward settings. Note also that, here, we considered the problem of maximizing a reward functional, but a formulation in terms of minimizing some cost or risk functional would be equivalent.

In order to maximize a given functional in a sequential framework, one usually applies Dynamic Programming (DP) [41], 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 [42]. 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 ([54]):

- 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 [49], 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.* [40] introduced the algorithm UCB (Upper Confidence Bounds) that follows what is now called the "optimism in the face of uncertainty principle". Their algorithm works by computing upper confidence bounds for all the arms and then choosing the arm with the highest such bound. They proved that the expected regret of their algorithm increases at most

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

3.3. Statistical analysis of time series

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

3.3.1. Prediction of Sequences of Structured and Unstructured Data

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

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

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

3.3.2. Hypothesis testing

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

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

3.3.3. Change Point Analysis

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

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

3.3.4. Clustering Time Series, Online and Offline

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

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

3.3.5. Online Semi-Supervised Learning

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

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

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

3.4. Statistical Learning and Bayesian Analysis

Before detailing some issues in these fields, let us remind the definition of a few terms.

Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge. This capacity to learn from experience, analytical observation, and other means, results in a system that can continuously self-improve and thereby offer increased efficiency and effectiveness.

Statistical learning is an approach to machine intelligence that is based on statistical modeling of data. With a statistical model in hand, one applies probability theory and decision theory to get an algorithm. This is opposed to using training data merely to select among different algorithms or using heuristics/"common sense" to design an algorithm.

Bayesian Analysis applies to data that could be seen as observations in the more general meaning of the term. These data may not only come from classical sensors but also from any *device* recording information. From an operational point of view, like for statistical learning, uncertainty about the data is modeled by a probability measure thus defining the so-called likelihood functions. This last one depends upon parameters defining the state of the world we focus on for decision purposes. Within the Bayesian framework the uncertainty about these parameters is also modeled by probability measures, the priors that are subjective probabilities. Using probability theory and decision theory, one then defines new algorithms to estimate the parameters of interest and/or associated decisions. According to the International Society for Bayesian Analysis (source: http://bayesian.org), and from a more general point of view, this overall process could be

summarize as follows: one assesses the current state of knowledge regarding the issue of interest, gather new data to address remaining questions, and then update and refine their understanding to incorporate both new and old data. Bayesian inference provides a logical, quantitative framework for this process based on probability theory.

Kernel method. Generally speaking, a kernel function is a function that maps a couple of points to a real value. Typically, this value is a measure of dissimilarity between the two points. Assuming a few properties on it, the kernel function implicitly defines a dot product in some function space. This very nice formal property as well as a bunch of others have ensured a strong appeal for these methods in the last 10 years in the field of function approximation. Many classical algorithms have been "kernelized", that is, restated in a much more general way than their original formulation. Kernels also implicitly induce the representation of data in a certain "suitable" space where the problem to solve (classification, regression, ...) is expected to be simpler (non-linearity turns to linearity).

The fundamental tools used in SEQUEL come from the field of statistical learning [45]. We briefly present the most important for us to date, namely, kernel-based non parametric function approximation, and non parametric Bayesian models.

3.4.1. Non-parametric methods for Function Approximation

In statistics in general, and applied mathematics, the approximation of a multi-dimensional real function given some samples is a well-known problem (known as either regression, or interpolation, or function approximation, ...). Regressing a function from data is a key ingredient of our research, or to the least, a basic component of most of our algorithms. In the context of sequential learning, we have to regress a function while data samples are being obtained one at a time, while keeping the constraint to be able to predict points at any step along the acquisition process. In sequential decision problems, we typically have to learn a value function, or a policy.

Many methods have been proposed for this purpose. We are looking for suitable ones to cope with the problems we wish to solve. In reinforcement learning, the value function may have areas where the gradient is large; these are areas where the approximation is difficult, while these are also the areas where the accuracy of the approximation should be maximal to obtain a good policy (and where, otherwise, a bad choice of action may imply catastrophic consequences).

We particularly favor non parametric methods since they make quite a few assumptions about the function to learn. In particular, we have strong interests in l_1-regularization, and the (kernelized-)LARS algorithm. l_1-regularization yields sparse solutions, and the LARS approach produces the whole regularization path very efficiently, which helps solving the regularization parameter tuning problem.

3.4.2. Nonparametric Bayesian Estimation

Numerous problems may be solved efficiently by a Bayesian approach. The use of Monte-Carlo methods allows us to handle non–linear, as well as non–Gaussian, problems. In their standard form, they require the formulation of probability densities in a parametric form. For instance, it is a common usage to use Gaussian likelihood, because it is handy. However, in some applications such as Bayesian filtering, or blind deconvolution, the choice of a parametric form of the density of the noise is often arbitrary. If this choice is wrong, it may also have dramatic consequences on the estimation quality. To overcome this shortcoming, one possible approach is to consider that this density must also be estimated from data. A general Bayesian approach then consists in defining a probabilistic space associated with the possible outcomes of the *object* to be estimated. Applied to density estimation, it means that we need to define a probability measure on the probability density of the noise: such a measure is called a *random measure*. The classical Bayesian inference procedures can then been used. This approach being by nature non parametric, the associated frame is called *Non Parametric Bayesian*.

In particular, mixtures of Dirichlet processes [44] provide a very powerful formalism. Dirichlet Processes are a possible random measure and Mixtures of Dirichlet Processes are an extension of well-known finite mixture models. Given a mixture density $f(x|\theta)$, and $G(d\theta) = \sum_{k=1}^{\infty} \omega_k \delta_{U_k}(d\theta)$, a Dirichlet process, we define a mixture of Dirichlet processes as:

$$F(x) = \int_{\Theta} f(x|\theta) G(d\theta) = \sum_{k=1}^{\infty} \omega_k f(x|U_k) \tag{4}$$

where $F(x)$ is the density to be estimated. The class of densities that may be written as a mixture of Dirichlet processes is very wide, so that they really fit a very large number of applications.

Given a set of observations, the estimation of the parameters of a mixture of Dirichlet processes is performed by way of a Monte Carlo Markov Chain (MCMC) algorithm. Dirichlet Process Mixture are also widely used in clustering problems. Once the parameters of a mixture are estimated, they can be interpreted as the parameters of a specific cluster defining a class as well. Dirichlet processes are well known within the machine learning community and their potential in statistical signal processing still need to be developed.

3.4.3. Random Finite Sets for multisensor multitarget tracking

In the general multi-sensor multi-target Bayesian framework, an unknown (and possibly varying) number of targets whose states $x_1, ...x_n$ are observed by several sensors which produce a collection of measurements $z_1, ..., z_m$ at every time step k. Well-known models to this problem are track-based models, such as the joint probability data association (JPDA), or joint multi-target probabilities, such as the joint multi-target probability density. Common difficulties in multi-target tracking arise from the fact that the system state and the collection of measures from sensors are unordered and their size evolve randomly through time. Vector-based algorithms must therefore account for state coordinates exchanges and missing data within an unknown time interval. Although this approach is very popular and has resulted in many algorithms in the past, it may not be the optimal way to tackle the problem, since the sate and the data are in fact *sets* and not vectors.

The random finite set theory provides a powerful framework to deal with these issues. Mahler's work on finite sets statistics (FISST) provides a mathematical framework to build multi-object densities and derive the Bayesian rules for state prediction and state estimation. Randomness on object number and their states are encapsulated into random finite sets (RFS), namely multi-target(state) sets $X = \{x_1, ..., x_n\}$ and multi-sensor (measurement) set $Zk = \{z_1, ..., z_m\}$. The objective is then to propagate the multitarget probability density $f_{k|k}(X|Z(k))$ by using the Bayesian set equations at every time step k:

$$\begin{aligned} f_{k+1|k}(X|Z^{(k)}) &= \int f_{k+1|k}(X|W) f_{k|k}(W|Z^{(k)}) \delta W \\ f_{k+1|k+1}(X|Z^{(k+1)}) &= \frac{f_{k+1}(Z_{k+1}|X) f_{k+1|k}(X|Z^{(k)})}{\int f_{k+1}(Z_{k+1}|W) f_{k+1|k}(W|Z^{(k)}) \delta W} \end{aligned} \tag{5}$$

where:

- $X = \{x_1, ..., x_n\}$ is a multi-target state, *i.e.* a finite set of elements x_i defined on the single-target space $\mathcal{X}$; [2]
- $Z_{k+1} = \{z_1, ..., z_m\}$ is the current multi-sensor observation, *i.e.* a collection of measures z_i produced at time $k+1$ by all the sensors;
- $Z^{(k)} = \bigcup_{t \leqslant k} Z_t$ is the collection of observations up to time k;
- $f_{k|k}(W|Z^{(k)})$ is the current multi-target posterior density in state W;
- $f_{k+1|k}(X|W)$ is the current multi-target Markov transition density, from state W to state X;
- $f_{k+1}(Z|X)$ is the current multi-sensor/multi-target likelihood function.

[2] The state x_i of a target is usually composed of its position, its velocity, etc.

Although equations (5) may seem similar to the classical single-sensor/single-target Bayesian equations, they are generally intractable because of the presence of the *set integrals*. For, a RFS Ξ is characterized by the family of its Janossy densities $j_{\Xi,1}(x_1)$, $j_{\Xi,2}(x_1,x_2)$... and not just by one density as it is the case with vectors. Mahler then introduced the PHD, defined on single-target state space. The PHD is the quantity whose integral on any region S is the expected number of targets inside S. Mahler proved that the PHD is the first-moment density of the multi-target probability density. Although defined on single-state space X, the PHD encapsulates information on both target number and states.

4. Application Domains

4.1. In Short

SEQUEL aims at solving problems of prediction, as well as problems of optimal and adaptive control. As such, the application domains are very numerous.

The application domains have been organized as follows:

- adaptive control,
- signal processing and functional prediction,
- web mining,
- computer games.

4.2. Adaptive Control

Adaptive control is an important application of the research being done in SEQUEL. Reinforcement learning (RL) precisely aims at controling the behavior of systems and may be used in situations with more or less information available. Of course, the more information, the better, in which case methods of (approximate) dynamic programming may be used [47]. But, reinforcement learning may also handle situations where the dynamics of the system is unknown, situations where the system is partially observable, and non stationary situations. Indeed, in these cases, the behavior is learned by interacting with the environment and thus naturally adapts to the changes of the environment. Furthermore, the adaptive system may also take advantage of expert knowledge when available.

Clearly, the spectrum of potential applications is very wide: as far as an agent (a human, a robot, a virtual agent) has to take a decision, in particular in cases where he lacks some information to take the decision, this enters the scope of our activities. To exemplify the potential applications, let us cite:

- game software: in the 1990's, RL has been the basis of a very successful Backgammon program, TD-Gammon [53] that learned to play at an expert level by basically playing a very large amount of games against itself. Today, various games are studied with RL techniques.
- many optimization problems that are closely related to operation research, but taking into account the uncertainty, and the stochasticity of the environment: see the job-shop scheduling, or the cellular phone frequency allocation problems, resource allocation in general [47]
- we can also foresee that some progress may be made by using RL to design adaptive conversational agents, or system-level as well as application-level operating systems that adapt to their users habits.

  More generally, these ideas fall into what adaptive control may bring to human beings, in making their life simpler, by being embedded in an environment that is made to help them, an idea phrased as "ambient intelligence".
- The sensor management problem consists in determining the best way to task several sensors when each sensor has many modes and search patterns. In the detection/tracking applications, the tasks assigned to a sensor management system are for instance:
  - detect targets,

  - track the targets in the case of a moving target and/or a smart target (a smart target can change its behavior when it detects that it is under analysis),
  - combine all the detections in order to track each moving target,
  - dynamically allocate the sensors in order to achieve the previous three tasks in an optimal way. The allocation of sensors, and their modes, thus defines the action space of the underlying Markov decision problem.

  In the more general situation, some sensors may be localized at the same place while others are dispatched over a given volume. Tasking a sensor may include, at each moment, such choices as where to point and/or what mode to use. Tasking a group of sensors includes the tasking of each individual sensor but also the choice of collaborating sensors subgroups. Of course, the sensor management problem is related to an objective. In general, sensors must balance complex trade-offs between achieving mission goals such as detecting new targets, tracking existing targets, and identifying existing targets. The word "target" is used here in its most general meaning, and the potential applications are not restricted to military applications. Whatever the underlying application, the sensor management problem consists in choosing at each time an action within the set of available actions.

- sequential decision processes are also very well-known in economy. They may be used as a decision aid tool, to help in the design of social helps, or the implementation of plants (see [51], [50] for such applications).

4.3. Signal Processing

Applications of sequential learning in the field of signal processing are also very numerous. A signal is naturally sequential as it flows. It usually comes from the recording of the output of sensors but the recording of any sequence of numbers may be considered as a signal like the stock-exchange rates evolution with respect to time and/or place, the number of consumers at a mall entrance or the number of connections to a web site. Signal processing has several objectives: predict , estimate, remove noise, characterize or classify. The signal is often considered as sequential: we want to predict, estimate or classify a value (or a feature) at time t knowing the past values of the parameter of interest or past values of data related to this parameter. This is typically the case in estimation processes arising in dynamical systems.

Signals may be processed in several ways. One of the best–known way is the time-frequency analysis in which the frequencies of each signal are analyzed with respect to time. This concept has been generalized to the time-scale analysis obtained by a wavelet transform. Both analysis are based on the projection of the original signal onto a well-chosen function basis. Signal processing is also closely related to the probability field as the uncertainty inherent to many signals leads to consider them as stochastic processes: the Bayesian framework is actually one of the main frameworks within which signals are processed for many purposes. It is worth noting that Bayesian analysis can be used jointly with a time-frequency or a wavelet analysis. However, alternatives like belief functions came up these last years. Belief functions were introduced by Demspter few decades ago and have been successfully used in the few past years in fields where probability had, during many years, no alternatives like in classification. Belief functions can be viewed as a generalization of probabilities which can capture both imprecision and uncertainty. Belief functions are also closely related to data fusion.

4.4. Web Mining

We work on the news/ad recommendation. These online learning algorithms reached a critical importance over the last few years due to these major applications. After designing a new algorithm, it is critical to be able to evaluate it without having to plug it into the real application in order to protect user experiences or/and the company's revenue. To do this, people used to build simulators of user behaviors and try to achieve good performances against it. However designing such a simulator is probably much more difficult than designing the algorithm itself! An other common way to evaluate is to not consider the exploration/exploitation dilemma (also known as "Cold Start" for recommender systems). Lately data-driven methods have been developed.

We are working on building automatic replay methodology with some theoretical guarantees. This work also exhibits strong link with the choice of the number of contexts to use with recommender systems wrt your audience.

An other point is that web sites must forecast Web page views in order to plan computer resource allocation and estimate upcoming revenue and advertising growth. In this work, we focus on extracting trends and seasonal patterns from page view series. We investigate Holt-Winters/ARIMA like procedures and some regularized models for making short-term prediction (3-6 weeks) wrt to logged data of several big media websites. We work on some news event related webpages and we feel that kind of time series deserves a particular attention. Self-similarity is found to exist at multiple time scales of network traffic, and can be exploited for prediction. In particular, it is found that Web page views exhibit strong impulsive changes occasionally. The impulses cause large prediction errors long after their occurrences and can sometimes be predicted (*e.g.*, elections, sport events, editorial changes,holidays) in order to improve accuracies. It also seems that some promising model could arise from using global trends shift in the population.

4.5. Games

The problem of artificial intelligence in games consists in choosing actions of players in order to produce artificial opponents. Most games can be formalized as Markov decision problems, so they can be approached with reinforcement learning.

In particular, SEQUEL was a pioneer of Monte Carlo Tree Search, a technique that obtained spectacular successes in the game of Go. Other application domains include the game of poker and the Japanese card game of hanafuda.

5. New Software and Platforms

5.1. Computer Games

Participant: Rémi Coulom.

- ***Crazy Stone*** is a top-level Go-playing program that has been developed by Rémi Coulom since 2005. Crazy Stone won several major international Go tournaments in the past. In 2013, a new version was released in Japan. This new version won the 6th edition of the UEC Cup (the most important international computer-Go tournament). It also won the first edition of the Denseisen, by winning a 4-stone handicap game against 9-dan professional player Yoshio Ishida. It is distributed as a commercial product by *Unbalance Corporation* (Japan). 6-month work in 2013. URL: http://remi.coulom.free.fr/CrazyStone/
- ***Kifu Snap*** is an Android image-recognition app. It can automatically recognize a Go board from a picture, and analyze it with Crazy Stone. It was released on Google Play in November, 2013. 6-month work in 2013. URL: http://remi.coulom.free.fr/kifu-snap/

5.2. Function optimization

Participant: Philippe Preux.

5.2.1. yaStoSOO

We have worked on the efficient implementation of the StoSOO algorithm in order to have a software that can be used for real to optimize real functions, and to be able to experiment with the algorithm, and assess its practical usefulness. This led to yaStoSOO, an implementation in C available on the web at http://www.grappa.univ-lille3.fr/~ppreux/software/StoSOO/. The code is distributed under the GPL licence.

Thanks to this implementation, we were able to compete in the CEC'2014 competition on Real-Parameter Single Objective optimization at which we ranked honorably (10th out of 17 competitor algorithms). More experimental work is under-way.

6. New Results

6.1. Highlights of the Year

- New startup by Rémi Coulom on AI in games (go, chess, $\cdots$).
- Successful Collaboration with Deezer and the victory at the ACM RecSys Recommendation Systems Challenge
- We were selected and working on preparation of ICML 2015 in Lille. ICML is the most important conference in the field of machine learning. This is the first time after more than 30 years of existence, that this conference will be held in France.

6.2. Decision-making Under Uncertainty

6.2.1. Reinforcement Learning

Selecting Near-Optimal Approximate State Representations in Reinforcement Learning [23]

We consider a reinforcement learning setting where the learner does not have explicit access to the states of the underlying Markov decision process (MDP). Instead, she has access to several models that map histories of past interactions to states. Here we improve over known regret bounds in this setting, and more importantly generalize to the case where the models given to the learner do not contain a true model resulting in an MDP representation but only approximations of it. We also give improved error bounds for state aggregation.

Online Stochastic Optimization under Correlated Bandit Feedback [15]

In this paper we consider the problem of online stochastic optimization of a locally smooth function under bandit feedback. We introduce the high-confidence tree (HCT) algorithm, a novel anytime X -armed bandit algorithm, and derive regret bounds matching the performance of state-of-the-art algorithms in terms of the dependency on number of steps and the near-optimality di-mension. The main advantage of HCT is that it handles the challenging case of correlated bandit feedback (reward), whereas existing meth-ods require rewards to be conditionally indepen-dent. HCT also improves on the state-of-the-art in terms of the memory requirement, as well as requiring a weaker smoothness assumption on the mean-reward function in comparison with the existing anytime algorithms. Finally, we discuss how HCT can be applied to the problem of policy search in reinforcement learning and we report preliminary empirical results.

Sparse Multi-task Reinforcement Learning [9]

In multi-task reinforcement learning (MTRL), the objective is to simultaneously learn multiple tasks and exploit their similarity to improve the performance w.r.t. single-task learning. In this paper we investigate the case when all the tasks can be accurately represented in a linear approximation space using the same small subset of the original (large) set of features. This is equivalent to assuming that the weight vectors of the task value functions are *jointly sparse*, i.e., the set of their non-zero components is small and it is shared across tasks. Building on existing results in multi-task regression, we develop two multi-task extensions of the fitted Q-iteration algorithm. While the first algorithm assumes that the tasks are jointly sparse in the given representation, the second one learns a transformation of the features in the attempt of finding a more sparse representation. For both algorithms we provide a sample complexity analysis and numerical simulations.

6.2.2. Multi-arm Bandit Theory

Spectral Bandits for Smooth Graph Functions with Applications in Recommender Systems [20]

Smooth functions on graphs have wide applications in manifold and semi-supervised learning. In this paper, we study a bandit problem where the payoffs of arms are smooth on a graph. This framework is suitable for solving online learning problems that involve graphs, such as content-based recommendation. In this problem, each recommended item is a node and its expected rating is similar to its neighbors. The goal is to recommend items that have high expected ratings. We aim for the algorithms where the cumulative regret would not scale poorly with the number of nodes. In particular, we introduce the notion of an effective dimension, which is small in real-world graphs, and propose two algorithms for solving our problem that scale linearly in this dimension. Our experiments on real-world content recommendation problem show that a good estimator of user preferences for thousands of items can be learned from just tens nodes evaluations.

Online combinatorial optimization with stochastic decision sets and adversarial losses [21]

Most work on sequential learning assumes a fixed set of actions that are available all the time. However, in practice, actions can consist of picking subsets of readings from sensors that may break from time to time, road segments that can be blocked or goods that are out of stock. In this paper we study learning algorithms that are able to deal with stochastic availability of such unreliable composite actions. We propose and analyze algorithms based on the Follow-The-Perturbed-Leader prediction method for several learning settings differing in the feedback provided to the learner. Our algorithms rely on a novel loss estimation technique that we call Counting Asleep Times. We deliver regret bounds for our algorithms for the previously studied full information and (semi-)bandit settings, as well as a natural middle point between the two that we call the restricted information setting. A special consequence of our results is a significant improvement of the best known performance guarantees achieved by an efficient algorithm for the sleeping bandit problem with stochastic availability. Finally, we evaluate our algorithms empirically and show their improvement over the known approaches.

Extreme bandits [10]

In many areas of medicine, security, and life sciences, we want to allocate limited resources to different sources in order to detect extreme values. In this paper, we study an efficient way to allocate these resources sequentially under limited feedback. While sequential design of experiments is well studied in bandit theory, the most commonly optimized property is the regret with respect to the maximum mean reward. However, in other problems such as network intrusion detection, we are interested in detecting the most extreme value output by the sources. Therefore, in our work we study extreme regret which measures the efficiency of an algorithm compared to the oracle policy selecting the source with the heaviest tail. We propose the ExtremeHunter algorithm, provide its analysis, and evaluate it empirically on synthetic and real-world experiments.

Efficient learning by implicit exploration in bandit problems with side observations [18]

We consider online learning problems under a a partial observability model capturing situations where the information conveyed to the learner is between full information and bandit feedback. In the simplest variant, we assume that in addition to its own loss, the learner also gets to observe losses of some other actions. The revealed losses depend on the learner's action and a directed observation system chosen by the environment. For this setting, we propose the first algorithm that enjoys near-optimal regret guarantees without having to know the observation system before selecting its actions. Along similar lines, we also define a new partial information setting that models online combinatorial optimization problems where the feedback received by the learner is between semi-bandit and full feedback. As the predictions of our first algorithm cannot be always computed efficiently in this setting, we propose another algorithm with similar properties and with the benefit of always being computationally efficient, at the price of a slightly more complicated tuning mechanism. Both algorithms rely on a novel exploration strategy called implicit exploration, which is shown to be more efficient both computationally and information-theoretically than previously studied exploration strategies for the problem.

Best-Arm Identification in Linear Bandits [29]

We study the best-arm identification problem in linear bandit, where the rewards of the arms depend linearly on an unknown parameter θ^* and the objective is to return the arm with the largest reward. We characterize the complexity of the problem and introduce sample allocation strategies that pull arms to identify the best arm with a fixed confidence, while minimizing the sample budget. In particular, we show the importance of exploiting the global linear structure to improve the estimate of the reward of near-optimal arms. We analyze the proposed strategies and compare their empirical performance. Finally, we point out the connection to the G-optimality criterion used in optimal experimental design.

Exploiting easy data in online optimization [28]

We consider the problem of online optimization, where a learner chooses a decision from a given decision set and suffers some loss associated with the decision and the state of the environment. The learner's objective is to minimize its cumulative regret against the best fixed decision in hindsight. Over the past few decades numerous variants have been considered, with many algorithms designed to achieve sub-linear regret in the worst case. However, this level of robustness comes at a cost. Proposed algorithms are often over-conservative, failing to adapt to the actual complexity of the loss sequence which is often far from the worst case. In this paper we introduce a general algorithm that, provided with a "safe" learning algorithm and an opportunistic "benchmark", can effectively combine good worst-case guarantees with much improved performance on "easy" data. We derive general theoretical bounds on the regret of the proposed algorithm and discuss its implementation in a wide range of applications, notably in the problem of learning with shifting experts (a recent COLT open problem). Finally, we provide numerical simulations in the setting of prediction with expert advice with comparisons to the state of the art.

Spectral Bandits for Smooth Graph Functions [32]

Smooth functions on graphs have wide applications in manifold and semi-supervised learning. In this paper, we study a bandit problem where the payoffs of arms are smooth on a graph. This framework is suitable for solving online learning problems that involve graphs, such as content-based recommendation. In this problem, each item we can recommend is a node and its expected rating is similar to its neighbors. The goal is to recommend items that have high expected ratings. We aim for the algorithms where the cumulative regret with respect to the optimal policy would not scale poorly with the number of nodes. In particular, we introduce the notion of an effective dimension, which is small in real-world graphs, and propose two algorithms for solving our problem that scale linearly and sublinearly in this dimension. Our experiments on real-world content recommendation problem show that a good estimator of user preferences for thousands of items can be learned from just tens of nodes evaluations.

Regret bounds for restless Markov bandits [5]

We consider the restless Markov bandit problem, in which the state of each arm evolves according to a Markov process independently of the learner's actions. We suggest an algorithm, that first represents the setting as an MDP which exhibits some special structural properties. In order to grasp this information we introduce the notion of ϵ-structured MDPs, which are a generalization of concepts like (approximate) state aggregation and MDP homomorphisms. We propose a general algorithm for learning ϵ-structured MDPs and show regret bounds that demonstrate that additional structural information enhances learning. Applied to the restless bandit setting, this algorithm achieves after any T steps regret of order $\widetilde{O}(T^{1/2})$ with respect to the best policy that knows the distributions of all arms. We make no assumptions on the Markov chains underlying each arm except that they are irreducible. In addition, we show that index-based policies are necessarily suboptimal for the considered problem.

Spectral Thompson Sampling [19]

Thompson Sampling (TS) has surged a lot of interest due to its good empirical performance, in particular in the computational advertising. Though successful, the tools for its performance analysis appeared only recently. In this paper, we describe and analyze SpectralTS algorithm for a bandit problem, where the payoffs of the choices are smooth given an underlying graph. In this setting, each choice is a node of a graph and the expected payoffs of the neighboring nodes are assumed to be similar. Although the setting has application

both in recommender systems and advertising, the traditional algorithms would scale poorly with the number of choices. For that purpose we consider an effective dimension d, which is small in real-world graphs. We deliver the analysis showing that the regret of SpectralTS scales as $d(T\ln N)^{1/2}$ with high probability, where T is the time horizon and N is the number of choices. Since a d sqrt(T ln N) regret is comparable to the known results, SpectralTS offers a computationally more efficient alternative. We also show that our algorithm is competitive on both synthetic and real-world data.

6.2.3. Recommendation systems

User Engagement as Evaluation: a Ranking or a Regression Problem? [39]

In this paper, we describe the winning approach used on the RecSys Challenge 2014 which focuses on employing user en-gagement as evaluation of recommendations. On one hand, we regard the challenge as a ranking problem and apply the LambdaMART algorithm, which is a listwise model special-ized in a Learning To Rank approach. On the other hand, after noticing some specific characteristics of this challenge, we also consider it as a regression problem and use pointwise regression models such as Random Forests. We compare how these different methods can be modified or combined to improve the accuracy and robustness of our model and we draw the advantages or disadvantages of each approach.

Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques [22]

In many recommendation applications such as news recommendation, the items that can be recommended come and go at a very fast pace. This is a challenge for recommender systems (RS) to face this setting. Online learning algorithms seem to be the most straight forward solution. The contextual bandit framework was introduced for that very purpose. In general the evaluation of a RS is a critical issue. Live evaluation is often avoided due to the potential loss of revenue, hence the need for offline evaluation methods. Two options are available. Model based meth- ods are biased by nature and are thus difficult to trust when used alone. Data driven methods are therefore what we consider here. Evaluat- ing online learning algorithms with past data is not simple but some methods exist in the litera- ture. Nonetheless their accuracy is not satisfac- tory mainly due to their mechanism of data re- jection that only allow the exploitation of a small fraction of the data. We precisely address this issue in this paper. After highlighting the limita- tions of the previous methods, we present a new method, based on bootstrapping techniques. This new method comes with two important improve- ments: it is much more accurate and it provides a measure of quality of its estimation. The latter is a highly desirable property in order to minimize the risks entailed by putting online a RS for the first time. We provide both theoretical and ex- perimental proofs of its superiority compared to state-of-the-art methods, as well as an analysis of the convergence of the measure of quality.

Bandits Warm-up Cold Recommender Systems [35]

We address the cold start problem in recommendation systems assuming no contextual information is available neither about users, nor items. We consider the case in which we only have access to a set of ratings of items by users. Most of the existing works consider a batch setting, and use cross-validation to tune parameters. The classical method consists in minimizing the root mean square error over a training subset of the ratings which provides a factorization of the matrix of ratings, interpreted as a latent representation of items and users. Our contribution in this paper is 5-fold. First, we explicit the issues raised by this kind of batch setting for users or items with very few ratings. Then, we propose an online setting closer to the actual use of recommender systems; this setting is inspired by the bandit framework. The proposed methodology can be used to turn any recommender system dataset (such as Netflix, MovieLens,...) into a sequential dataset. Then, we explicit a strong and insightful link between contextual bandit algorithms and matrix factorization; this leads us to a new algorithm that tackles the exploration/exploitation dilemma associated to the cold start problem in a strikingly new perspective. Finally, experimental evidence confirm that our algorithm is effective in dealing with the cold start problem on publicly available datasets. Overall, the goal of this paper is to bridge the gap between recommender systems based on matrix factorizations and those based on contextual bandits.

6.2.4. Nonparametric statistics of time series

Uniform hypothesis testing for finite-valued stationary processes [6]

Given a discrete-valued sample $X_1, \cdots, X_n$ we wish to decide whether it was generated by a distribution belonging to a family H_0, or it was generated by a distribution belonging to a family H_1. In this work we assume that all distributions are stationary ergodic, and do not make any further assumptions (e.g. no independence or mixing rate assumptions). We would like to have a test whose probability of error (both Type I and Type II) is uniformly bounded. More precisely, we require that for each ϵ there exist a sample size n such that probability of error is upper-bounded by ϵ for samples longer than n. We find some necessary and some sufficient conditions on H_0 and H_1 under which a consistent test (with this notion of consistency) exists. These conditions are topological, with respect to the topology of distributional distance.

Asymptotically consistent estimation of the number of change points in highly dependent time series [17]

The problem of change point estimation is considered in a general framework where the data are generated by arbitrary unknown stationary ergodic process distributions. This means that the data may have long-range dependencies of an arbitrary form. In this context the consistent estimation of the number of change points is provably impossible. A formulation is proposed which overcomes this obstacle: it is possible to find the correct number of change points at the expense of introducing the additional constraint that the correct number of process distributions that generate the data is provided. This additional parameter has a natural interpretation in many real-world applications. It turns out that in this formulation change point estimation can be reduced to time series clustering. Based on this reduction, an algorithm is proposed that finds the number of change points and locates the changes. This algorithm is shown to be asymptotically consistent. The theoretical results are complemented with empirical evaluations.

6.3. Statistical Learning and Bayesian Analysis

6.3.1. Prediction of Sequences of Structured and Unstructured Data

Statistical performance analysis of a fast super-resolution technique using noisy translations [38]

It is well known that the registration process is a key step for super-resolution reconstruction. In this work, we propose to use a piezoelectric system that is easily adaptable on all microscopes and telescopes for controlling accurately their motion (down to nanometers) and therefore acquiring multiple images of the same scene at different controlled positions. Then a fast super-resolution algorithm can be used for efficient super-resolution reconstruction. In this case, the optimal use of r2 images for a resolution enhancement factor r is generally not enough to obtain satisfying results due to the random inaccuracy of the positioning system. Thus we propose to take several images around each reference position. We study the error produced by the super-resolution algorithm due to spatial uncertainty as a function of the number of images per position. We obtain a lower bound on the number of images that is necessary to ensure a given error upper bound with probability higher than some desired confidence level.

Quantitative control of the error bounds of a fast super-resolution technique for microscopy and astronomy [11]

While the registration step is often problematic for super-resolution, many microscopes and telescopes are now equipped with a piezoelectric mechanical system which permits to ac-curately control their motion (down to nanometers). There-fore one can use such devices to acquire multiple images of the same scene at various controlled positions. Then a fast super-resolution algorithm [1] can be used for efficient super-resolution. However the minimal use of r 2 images for a resolution enhancement factor r is generally not sufficient to obtain good results. We propose to take several images at po-sitions randomly distributed close to each reference position. We study the number of images necessary to control the error resulting from the super-resolution algorithm by [1] due to the uncertainty on positions. The main result is a lower bound on the number of images to respect a given error upper bound with probability higher than a desired confidence level.

6.3.2. Statistical analysis of superresolution

A diffusion strategy for distributed dictionary learning [12]

We consider the problem of a set of nodes which is required to collectively learn a common dictionary from noisy measurements. This distributed dictionary learning approach may be useful in several contexts including sensor networks. Dif-fusion cooperation schemes have been proposed to estimate a consensus solution to distributed linear regression. This work proposes a diffusion-based adaptive dictionary learning strategy. Each node receives measurements which may be shared or not with its neighbors. All nodes cooperate with their neighbors by sharing their local dictionary to estimate a common representa-tion. In a diffusion approach, the resulting algorithm corresponds to a distributed alternate optimization. Beyond dictionary learn-ing, this strategy could be adapted to many matrix factorization problems in various settings. We illustrate its efficiency on some numerical experiments, including the difficult problem of blind hyperspectral images unmixing.

6.4. Miscellaneous

6.4.1. Miscellaneous

Online Matrix Completion Through Nuclear Norm Regularisation [14]

It is the main goal of this paper to propose a novel method to perform matrix completion on-line. Motivated by a wide variety of applications, ranging from the design of recommender systems to sensor network localization through seismic data reconstruction, we consider the matrix completion problem when entries of the matrix of interest are observed gradually. Precisely, we place ourselves in the situation where the predictive rule should be refined incrementally, rather than recomputed from scratch each time the sample of observed entries increases. The extension of existing matrix completion methods to the sequential prediction context is indeed a major issue in the Big Data era, and yet little addressed in the literature. The algorithm promoted in this article builds upon the Soft Impute approach introduced in Mazumder et al. (2010). The major novelty essentially arises from the use of a randomised technique for both computing and updating the Singular Value Decomposition (SVD) involved in the algorithm. Though of disarming simplicity, the method proposed turns out to be very efficient, while requiring reduced computations. Several numerical experiments based on real datasets illustrating its performance are displayed, together with preliminary results giving it a theoretical basis.

Synthèse en espace et temps du rayonnement acoustique d'une paroi sous excitation turbulente par synthèse spectrale 2D+T et formulation vibro-acoustique directe [33]

Une méthode directe pour simuler les vibrations et le rayonnement acoustique d'une paroi soumise à un écoulement subsonique est proposée. Tout d'abord, en adoptant l'hypothèse d'un écoulement homogène et stationnaire, on montre qu'une méthode de synthèse spectrale en espace et temps (2D+t) est suffisante pour obtenir explicitement une réalisation d'un champ de pression pariétale excitatrice p(x,y,t) dont les propriétés inter-spectrales sont prescrites par un modèle empirique de Chase. Cette pression turbulente p(x,y,t) est obtenue explicitement et permet de résoudre le problème vibro-acoustique de la paroi dans une formulation directe. La méthode proposée fournit ainsi une solution complète du problème dans le domaine spatio-temporel : pression excitatrice, déplacement en flexion et pression acoustique rayonnée par la paroi. Une caractéristique de la méthode proposée est un cout de calcul qui s'avère similaire aux formulations inter-spectrales majoritairement utilisées dans la littérature. En particulier, la synthèse permet de prendre en compte l'intégralité des échelles spatio-temporelles du problème : échelles turbulentes, vibratoires et acoustiques. A titre d'exemple, la pression aux oreilles d'un auditeur suite à l'excitation turbulente de la paroi est synthétisée.

Bandits attack function optimization [27]

We consider function optimization as a sequential decision making problem under the budget constraint. Such constraint limits the number of objective function evaluations allowed during the optimization. We consider an algorithm inspired by a continuous version of a multi-armed bandit problem which attacks this optimization problem by solving the tradeoff between exploration (initial quasi-uniform search of the domain)

and exploitation (local optimization around the potentially global maxima). We introduce the so-called Simultaneous Optimistic Optimization (SOO), a deterministic algorithm that works by domain partitioning. The benefit of such an approach are the guarantees on the returned solution and the numerical eficiency of the algorithm. We present this machine learning rooted approach to optimization, and provide the empirical assessment of SOO on the CEC'2014 competition on single objective real-parameter numerical optimization testsuite.

Optimistic planning in Markov decision processes using a generative model [30]

We consider the problem of online planning in a Markov decision process with discounted rewards for any given initial state. We consider the PAC sample com-plexity problem of computing, with probability 1-δ, an -optimal action using the smallest possible number of calls to the generative model (which provides reward and next-state samples). We design an algorithm, called StOP (for Stochastic-Optimistic Planning), based on the "optimism in the face of uncertainty" princi-ple. StOP can be used in the general setting, requires only a generative model, and enjoys a complexity bound that only depends on the local structure of the MDP.

Near-Optimal Rates for Limited-Delay Universal Lossy Source Coding [3]

We consider the problem of limited-delay lossy coding of individual sequences. Here, the goal is to design (fixed-rate) compression schemes to minimize the normalized expected distortion redundancy relative to a reference class of coding schemes, measured as the difference between the average distortion of the algorithm and that of the best coding scheme in the reference class. In compressing a sequence of length T, the best schemes available in the literature achieve an $O(T^{-1/3})$ normalized distortion redundancy relative to finite reference classes of limited delay and limited memory, and the same redundancy is achievable, up to logarithmic factors, when the reference class is the set of scalar quantizers. It has also been shown that the distortion redundancy is at least of order $T^{-1/2}$ in the latter case, and the lower bound can easily be extended to sufficiently powerful (possibly finite) reference coding schemes. In this paper, we narrow the gap between the upper and lower bounds, and give a compression scheme whose normalized distortion redundancy is $O(ln(T)/T^{1/2})$ relative to any finite class of reference schemes, only a logarithmic factor larger than the lower bound. The method is based on the recently introduced shrinking dartboard prediction algorithm, a variant of exponentially weighted average prediction. The algorithm is also extended to the problem of joint source-channel coding over a (known) stochastic noisy channel and to the case when side information is also available to the decoder (the Wyner–Ziv setting). The same improvements are obtained for these settings as in the case of a noiseless channel. Our method is also applied to the problem of zero-delay scalar quantization, where $O(ln(T)/T^{1/2})$ normalized distortion redundancy is achieved relative to the (infinite) class of scalar quantizers of a given rate, almost achieving the known lower bound of order $1/T^{-1/2}$. The computationally efficient algorithms known for scalar quantization and the Wyner–Ziv setting carry over to our (improved) coding schemes presented in this paper.

Online Markov Decision Processes Under Bandit Feedback [4]

Software systems are composed of many interacting elements. A natural way to abstract over software systems is to model them as graphs. In this paper we consider software dependency graphs of object-oriented software and we study one topological property: the degree distribution. Based on the analysis of ten software systems written in Java, we show that there exists completely different systems that have the same degree distribution. Then, we propose a generative model of software dependency graphs which synthesizes graphs whose degree distribution is close to the empirical ones observed in real software systems. This model gives us novel insights on the potential fundamental rules of software evolution.

A Generative Model of Software Dependency Graphs to Better Understand Software Evolution [37]

Software systems are composed of many interacting elements. A natural way to abstract over software systems is to model them as graphs. In this paper we consider software dependency graphs of object-oriented software and we study one topological property: the degree distribution. Based on the analysis of ten software systems written in Java, we show that there exists completely different systems that have the same degree distribution. Then, we propose a generative model of software dependency graphs which synthesizes graphs whose degree

distribution is close to the empirical ones observed in real software systems. This model gives us novel insights on the potential fundamental rules of software evolution.

Preference-Based Rank Elicitation using Statistical Models: The Case of Mallows [8]

We address the problem of rank elicitation as-suming that the underlying data generating pro-cess is characterized by a probability distribu-tion on the set of all rankings (total orders) of a given set of items. Instead of asking for complete rankings, however, our learner is only allowed to query pairwise preferences. Using information of that kind, the goal of the learner is to reliably predict properties of the distribution, such as the most probable top-item, the most probable rank-ing, or the distribution itself. More specifically, learning is done in an online manner, and the goal is to minimize sample complexity while guaran-teeing a certain level of confidence.

Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm [1]

We introduce a novel approach to preference-based reinforcement learn-ing, namely a preference-based variant of a direct policy search method based on evolutionary optimization. The core of our approach is a preference-based racing algorithm that selects the best among a given set of candidate policies with high probability. To this end, the algorithm operates on a suitable ordinal preference structure and only uses pairwise comparisons between sample rollouts of the policies. Embedding the racing algorithm in a rank-based evolutionary search procedure, we show that approxima-tions of the so-called Smith set of optimal policies can be produced with certain theoretical guarantees. Apart from a formal performance and complexity analysis, we present first experimental studies showing that our approach performs well in practice.

Biclique Coverings, Rectifier Networks and the Cost of ε-Removal [16]

We relate two complexity notions of bipartite graphs: the minimal weight biclique covering number Cov(G) and the minimal rec-tifier network size Rect(G) of a bipartite graph G. We show that there exist graphs with Cov(G) $\geq$ Rect(G) 3/2-ǫ . As a corollary, we estab-lish that there exist nondeterministic finite automata (NFAs) with ε-transitions, having n transitions total such that the smallest equivalent ε-free NFA has Ω(n 3/2-ǫ) transitions. We also formulate a version of previous bounds for the weighted set cover problem and discuss its con-nections to giving upper bounds for the possible blow-up.

Efficient Eigen-updating for Spectral Graph Clustering [2]

Partitioning a graph into groups of vertices such that those within each group are more densely connected than vertices assigned to different groups, known as graph clustering, is often used to gain insight into the organisation of large scale networks and for visualisation purposes. Whereas a large number of dedicated techniques have been recently proposed for static graphs, the design of on-line graph clustering methods tailored for evolving networks is a challenging problem, and much less documented in the literature. Motivated by the broad variety of applications concerned, ranging from the study of biological networks to the analysis of networks of scientific references through the exploration of communications networks such as the World Wide Web, it is the main purpose of this paper to introduce a novel, computationally efficient, approach to graph clustering in the evolutionary context. Namely, the method promoted in this article can be viewed as an incremental eigenvalue solution for the spectral clustering method described by Ng. et al. (2001). The incremental eigenvalue solution is a general technique for finding the approximate eigenvectors of a symmetric matrix given a change. As well as outlining the approach in detail, we present a theoretical bound on the quality of the approximate eigenvectors using perturbation theory. We then derive a novel spectral clustering algorithm called Incremental Approximate Spectral Clustering (IASC). The IASC algorithm is simple to implement and its efficacy is demonstrated on both synthetic and real datasets modelling the evolution of a HIV epidemic, a citation network and the purchase history graph of an e-commerce website.

From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning [36]

This work covers several aspects of the optimism in the face of uncertainty principle applied to large scale optimization problems under finite numerical budget. The initial motivation for the research reported here originated from the empirical success of the so-called Monte-Carlo Tree Search method popularized in computer-go and further extended to many other games as well as optimization and planning problems. Our objective is to contribute to the development of theoretical foundations of the field by characterizing the complexity of the underlying optimization problems and designing efficient algorithms with performance guarantees. The main idea presented here is that it is possible to decompose a complex decision making problem (such as an optimization problem in a large search space) into a sequence of elementary decisions, where each decision of the sequence is solved using a (stochastic) multi-armed bandit (simple mathematical model for decision making in stochastic environments). This so-called hierarchical bandit approach (where the reward observed by a bandit in the hierarchy is itself the return of another bandit at a deeper level) possesses the nice feature of starting the exploration by a quasi-uniform sampling of the space and then focusing progressively on the most promising area, at different scales, according to the evaluations observed so far, and eventually performing a local search around the global optima of the function. The performance of the method is assessed in terms of the optimality of the returned solution as a function of the number of function evaluations. Our main contribution to the field of function optimization is a class of hierarchical optimistic algorithms designed for general search spaces (such as metric spaces, trees, graphs, Euclidean spaces, ...) with different algorithmic instantiations depending on whether the evaluations are noisy or noiseless and whether some measure of the ”smoothness” of the function is known or unknown. The performance of the algorithms depend on the local behavior of the function around its global optima expressed in terms of the quantity of near-optimal states measured with some metric. If this local smoothness of the function is known then one can design very efficient optimization algorithms (with convergence rate independent of the space dimension), and when it is not known, we can build adaptive techniques that can, in some cases, perform almost as well as when it is known.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contracts with Industry

- Deezer, 2013-2014
  Participants: Jérémie Mary, Philippe Preux, Romaric Gaudel.

  A research project has started on June 2013 in collaboration with the Deezer company. The goal is to build a system which automatically recommends music to users. That goal is an extension of the bandit setting to the Collaborative Filtering problem.

- Nuukik, 2013-2014
  Participant: Jérémie Mary.

  Nuukik is a start-up from Hub Innovation in Lille. It proposes a recommender systems for e-commerce based on matrix factorization. We worked with them specifically on the cold start problem (*i.e* when you have absolutely no data on a product or a customer). This led to promising result and allowed us to close the gap between bandits and matrix factorization. This work led to a patent submission in december 2013.

- Squoring Technologies, 2011-2014
  Participants: Boris Baldassari, Philippe Preux.

  Boris Baldassari has been hired by Squoring Technologies (Toulouse) as a PhD student in May 2011. He works on the use of machine learning to improve the quality of the software development process. During his first year as a PhD student, Boris investigated the existing norms and measures of quality of software development process. He also dedicated some time to gather some relevant datasets, which are made of either the sequence of source code releases over a multi-years period, or all the versions stored on an svn repository (svn or alike). Information from mailing-lists (bugs,

support, ...) may also be part of these datasets. Tools in machine learning capable of dealing with this sort of data have also been investigated. Goals that may be reached in this endeavor have also been precised.

7.2. Bilateral Grants with Industry

- INTEL Corp., 2013 - 2014
  Participants: Philippe Preux, Michal Valko, Rémi Munos, Adrien Hoarau.

  This is a research project on Algorithmic Determination of IoT Edge Analytics Requirements. We are attempting to solve the problem of how to automatically predict the system requirements for edge node analytics in the Internet of Things (IoT). We envision that a flexible extensible system of edge analytics can be created for IoT management; however, edge nodes can be very different in terms of the systems requirements around: processing capability, wireless communication, security/cryptography, guaranteed responsiveness, guaranteed quality of service and on-board memory requirements. One of the challenges of managing a heterogeneous Internet of Things is determining the systems requirements at each edge node in the network.

  We suggest exploiting opportunity of being able to automatically customize large scale IoT systems that could comprise heterogeneous edge nodes and allow a flexible and scalable component and firmware SoC systems to be matched to the individual need of enterprise/ government level IoT customers. We propose using large scale sequential decision learning algorithms, particularly contextual bandit modeling to automatically determine the systems requirements for edge analytics. These algorithms have an adaptive property that allows for the addition of new nodes and the re-evaluation of existing nodes under dynamic and potentially adversarial conditions.

8. Partnerships and Cooperations

8.1. Regional Initiatives

Pierre Chainais and Hong-Phuong Dang are part of the ARCIR project *REPAR*, PARcimonious REpresentations, which is funded by the Region Nord-Pas de Calais for 2 years. This project is focused on sparsity based methods for signal and image processing. It has permitted to hire 1 postdoc for 1 year (2014-2015) who works on the use of sparse representation for video-tracking. The targetted application is in biological microscopy to track cellular vesiculas (collab. Laurent Héliot, Aymeric Leray, Univ. Lille 1).

8.2. National Initiatives

8.2.1. ANR BNPSI

Participants: Pierre Chainais, Hong-Phuong Dang, Clément Elvira, Emmanuel Duflos, Philippe Vanheeghe.

- *Title*: Bayesian Non Parametric approaches for Signal and Image Processing
- *Type*: National Research Agency no ANR-13-BS-03-0006-01
- *Coordinator*: Ecole Centrale Lille, LAGIS (P. Chainais)
- *Duration*: 2014-2018
- *Other Partners*: Inria Bordeaux, team ALEA, Université de Bordeaux, IMS, Institut de Recherche en Indormatique de Toulouse (IRIT), CEA-LIST Saclay.

- *Abstract*: Statistical methods have become more and more popular in signal and image processing over the past decades. These methods have been able to tackle various applications such as speech recognition, object tracking, image segmentation or restoration, classification, clustering, etc. We propose here to investigate the use of Bayesian nonparametric methods in statistical signal and image processing. Similarly to Bayesian parametric methods, this set of methods is concerned with the elicitation of prior and computation of posterior distributions, but now on infinite-dimensional parameter spaces. Although these methods have become very popular in statistics and machine learning over the last 15 years, their potential is largely underexploited in signal and image processing. The aim of the overall project, which gathers researchers in applied probabilities, statistics, machine learning and signal and image processing, is to develop a new framework for the statistical signal and image processing communities. Based on results from statistics and machine learning we aim at defining new models, methods and algorithms for statistical signal and image processing. Applications to hyperspectral image analysis, image segmentation, GPS localization, image restoration or space-time tomographic reconstruction will allow various concrete illustrations of the theoretical advances and validation on real data coming from realistic contexts.
- *Activity Report*: This ANR Project was accepted in 2013. It has started in february 2014 on a new area of research for signal and image processing and is supervised by Pierre Chainais. Three meetings have taken place in Lille (in February), Toulouse (in June) and Bordeaux (in November). One special session on Bayesian non parametric approaches has been submitted and accepted to the international conference EUSIPCO 2015. We have also been selected by the Franch National Signal & Image Processing Society (GRETSI) to organize the Peyresq 2016 Signal processing summer school. Two PhD students have been recruited in october 2014 thanks to this project: Clément Elvira works in Lille is co-supervised by P. Chainais and N. Dobigeon (Toulouse), Jessica Sodjo works in Bordeaux and is co-supervised by A. Giremus (IMS), N. Dobigeon (Toulouse) and F. Caron (Oxford). Moreover, Hong-Phuong Dang (PhD, 2nd year) has obtained new results on BNP for dictionary learning. The Indian Buffet Process permits to propose a method to learn a dictionary of which size automatically adapts to data. Several publications are in preparation. François Caron who is co-leading this project with Pierre Chainais has moved to Oxford University as an Assistant Professor so that we will benefit from strong connections with the Statistics Departmnt in Oxford University.

8.2.2. 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*: ExTra-Learn started officially in October and one paper has been published at NIPS'14 and in the workshop on "Transfer and Multi-task Learning" at NIPS'14.

8.2.3. National Partners

- Laboratoire Paul Painlevé Université des Sciences et Technologies de Lille, France
  - Mylène Maïda *Collaborator*

    Ph. Preux has collaborated with M. Maïda and co-advised a student of the École Centrale de Lille. The motivation of this collaboration is the study of random matrices and the potential use of this theory in machine learning.
- CMLA - ENS Cachan.
  - Julien Audiffren *Collaborator*
    M. Valko, A. Lazaric, and M. Ghavamzadeh work with Julien on Semi-Supervised Apprenticeship Learning. We work on a maximum entropy algorithm that outperforms the approach without unlabeled data.
- Laboratoire Lagrange, Université de Nice, France.
  - Cédric Richard *Collaborator*
    We have had collaboration on the topic of *dictionary learning over a sensor network*.
- Laboratoire de Mécanique de Lille, Université de Lille 1, France.
  - Jean-Philippe Laval *Collaborator*
    We co-supervise a starting PhD student (Linh Van Nguyen) on the topic of *high resolution field reconstruction from low resolution measurements in turbulent flows*.
- Institut Carnot de Bourgogne, CNRS UMR 6303, Université de Bourgogne, Dijon, France.
  - Aymeric Leray *Collaborator*
    P. Chainais and A. Leray have written an article on the topic of *quantitative guarantees of a super resolution method via concentration inequalities*. A paper has been published in ICASSP 2014 proceedings and a journal article is submitted to IEEE Transactions on Image Processing.
- LAGIS (CRIStAL), Ecole Centrale Lille - Université de Lille 1, France.
  - Patrick Bas *Collaborator*
    P. Chainais and P. Bas have a collaboration on the topic of *adaptive quantization to optimize classification from histrograms of features with an application to the steganalysis of textured images*.
- University of Oxford (Great-Britain)
  - Dr. François Caron *Collaborators*

  - P. Chainais is co-leading the ANR BNPSI in collaboration with François Caron. Note that Rémi Bardenet will arrive in Lille as a CNRS researcher in feb. 2015 after a post-doc at Oxford University.

- LTCI, Institut Télécom-ParisTech, France.
  - Charanpal Dhanjal*Collaborator*
    We have a collaboration on the topic of *Matrix Factorization update* with application to sequential recommendation and sequential clustering. This collaboration has leaded to two publications this year: one in Neurocomputing journal [2], one at SDM'14 conference [14].

8.3. European Initiatives

8.3.1. FP7 & H2020 Projects

8.3.1.1. CompLACS

Type: FP7

Defi: Cognitive Systems, Interaction, Robotics

Instrument: Specific Targeted Research Project

Objectif: Cognitive Systems and Robotics

Duration: March 2011 - February 2015

Coordinator: John Shaw-Taylor

Partner: University College London, University of Bristol, Royal Holloway, University of London, Radboud Universiteit Nijmegen, Technische Universitat Berlin, Montanuniversitat Leoben, Institut National de Recherche en Informatique et en Automatique, Technische Universität Darmstadt

Inria contact: Rémi MUNOS

Abstract: One of the aspirations of machine learning is to develop intelligent systems that can address a wide variety of control problems of many different types. However, although the community has developed successful technologies for many individual problems, these technologies have not previously been integrated into a unified framework. As a result, the technology used to specify, solve and analyse one control problem typically cannot be reused on a different problem. The community has fragmented into a diverse set of specialists with particular solutions to particular problems. The purpose of this project is to develop a unified toolkit for intelligent control in many different problem areas. This toolkit will incorporate many of the most successful approaches to a variety of important control problems within a single framework, including bandit problems, Markov Decision Processes (MDPs), Partially Observable MDPs (POMDPs), continuous stochastic control, and multi-agent systems. In addition, the toolkit will provide methods for the automatic construction of representations and capabilities, which can then be applied to any of these problem types. Finally, the toolkit will provide a generic interface to specifying problems and analysing performance, by mapping intuitive, human-understandable goals into machine-understandable objectives, and by mapping algorithm performance and regret back into human-understandable terms.

8.4. International Initiatives

8.4.1. Inria International Partners

- Inria International partnership with Leoben, Austria; starting October 2014; duration: 4 years.
  - Ronald Ortner and Peter Auer: Montanuniversität Leoben (Austria).

– Reinforcement learning (RL) deals with the problem of interacting with an unknown stochastic environment that occasionally provides rewards, with the goal of maximizing the cumulative reward. The problem is well-understood when the unknown environment is a finite-state Markov process. This collaboration is centered around reducing the general RL problem to this case.

  In particular, the following problems are considered: representation learning, learning in continuous-state environments, bandit problems with dependent arms, and pure exploration in bandit problems. On each of these problems we have successfully collaborated in the past, and plan to sustain this collaboration possibly extending its scopes.

8.4.1.1. Informal International Partners

- Technion - Israel Institute of Technology, Haifa, Israel.
  - Odalric-Ambrym Maillard *Collaborator*
    Daniil Ryabko has worked with Odalric Maillard on representation learning for reinforcement learning problems. It led to a paper in AISTATS [46].
- School of Computer Science, Carnegie Mellon University, USA.
  - Prof. Emma Brunskill *Collaborator*
  - Mohammad Gheshlaghi Azar, (now at Northwestern University in Chicago) *Collaborator*
    A. Lazaric continued his collaboration on transfer in multi-arm bandit and reinforcement learning which led to one publication at ICML'14. We have submitted an associate team project with E. Brunskill on the topic of multi-arm bandit applied to education.
- Technicolor Research, Palo Alto.
  - Branislav Kveton *Collaborator*
    Michal Valko and Rémi Munos worked with Branislav on Spectral Bandits aimed at recommendation for the entertainment content recommendation. Michal continued the ongoing research on online semi-supervised learning and this year delivered the algorithm for a challenging single picture per person setting. Victor Gabillon has spent 6 month at Technicolor as an intern to work on the sequential learning with submodularity, which resulted in 1 accepted paper at NIPS, 1 in ICML, and 1 in AAAI.
- University of Cambridge (UK)
  - Alexandra Carpentier *Collaborator*
  - Michal Valko collaborates with A. Carpentier on extreme event detection (such as network intrusion) with limited allocation capabilities.
- Politecnico di Milano (Italy)
  - Prof. Marcello Restelli and Prof. Nicola Gatti *Collaborators*
  - A. Lazaric continued his collaboration on transfer in reinforcement learning which leads to a publication in NIPS'14. Furthermore, we have submitted a journal version of an application of multi-arm bandit in sponsored search auctions which is currently under review.

8.5. International Research Visitors

8.5.1. Visits of International Scientists

8.5.1.1. Internships

- Daniele Calandriello, student at Politecnico di Milano, Italy

  Period: April 2013 to May 2014.

  He was working with A. Lazaric on multi-task reinforcement learning.
- Jessica Chemali, Master, Carnegie Mellon University, May-August 2014

8.5.2. Visits to International Teams

8.5.2.1. Sabbatical programme

Ryabko Daniil

Date: Jan 2014 - Jan 2015

Institution: Centro de Modelamiento Matematico (Chile)

8.5.2.2. Research stays abroad

Munos Rémi

Date: Jul 2013 - June 2014

Institution: Microsoft Research New England (USA)

Munos Rémi

Date: October 2014 - now

Institution: Google Deepmind (UK)

Ghavamzadeh Mohammad

Date: September 2013 - now

Institution: Adobe Research (USA)

9. Dissemination

9.1. Promoting Scientific Activities

9.1.1. Scientific events organisation

9.1.1.1. general chair, scientific chair

- P. Chainais has co-organized with Z. Harchaoui the GDR ISIS 1 day workshop on "Learning adapted representations for signal and image processing" in Paris on Feb. 4th, 2014, see http://www.gdr-isis.fr/index.php?page=reunion&idreunion=234.
- P. Chainais has led the application of Lille to the organization of the french national Signal Processing conference (GRETSI 2017) : Marrakech won, but we had a good feedback in view of 2019.

9.1.1.2. member of the conference program committee

- AAAI Conference on Artificial Intelligence (AAAI 2014)
- IEEE Approximate Dynamic Programming and Reinforcement Learning (ADPRL 2014)
- French Conference on Planning, Decision-making, and Learning in Control Systems (JFPDA 2014)
- Conférence Apprentissage Automatique (CAP)
- Extraction et Gestion des Connaissances (EGC)

9.1.1.3. reviewer

- International Conference on Pattern Recognition Applications and Methods (ICPRAM 2014)
- Algorithmic Learning Theory (ALT 2014)
- AAAI Conference on Artificial Intelligence (AAAI 2014)
- Conference on Learning Theory (COLT 2014)
- European Workshop on Reinforcement Learning (EWRL 2014)
- Annual Conference on Neural Information Processing Systems (NIPS 2014)
- International Conference on Artificial Intelligence and Statistics (AISTATS 2014)
- European Conference on Machine Learning (ECML 2014)

- International Conference on Machine Learning (ICML 2014)
- International Conference on Uncertainty in Artificial Intelligence (UAI 2014)
- IEEE Congress on Evolutionary Computation (CEC)
- French Conference on Planning, Decision-making, and Learning in Control Systems (JFPDA 2014)
- IEEE FUSION 2014
- IEEE Approximate Dynamic Programming and Reinforcement Learning (ADPRL 2014)

9.1.2. Journal

9.1.2.1. reviewer

- IEEE Transactions on Image Processing
- Journal of Statistical Physics
- Digital Signal Processing
- IEEE Transactions on Information Theory
- IEEE Statistical Signal Processing SSP'2013
- European Signal Processing Conference EUSIPCO 2013
- 10th International Conference on Sampling Theory and Applications (SampTA 2013)
- IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2013 & 2014)
- Annual Conference on Neural Information Processing Systems (NIPS 2013)
- International Conference on Machine Learning (ICML 2013)
- European Conference on Machine Learning (ECML 2013)
- Uncertainty in Artificial Intelligence (UAI 2013)
- Machine Learning Journal (MLJ)
- Journal of Machine Learning Research (JMLR)
- Journal of Artificial Intelligence Research (JAIR)
- IEEE Transactions on Automatic Control (TAC)
- IEEE Transactions of Signal Processing
- Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)
- Mathematics of Operations Research (MOR)

9.1.3. Invited Talks

- Alessandro Lazaric gave an invited talk on "Approximate Dynamic Programming meets Statistical Learning Theory" at CNRS Journée Des D''ecollements in Orsay (November 2014)
- Alessandro Lazaric gave a talk on "Transfer in Reinforcement Learning" at the "30 minutes of sciences" seminars at Inria Lille (December 2014)
- Michal Valko gave a talk "Bandits on Graphs" at CMLA group at ENS Cachan (December 2014)
- Michal Valko gave a talk "Optimistic Optimization" at CMLA group and at MIST conference, Slovakia (January 2014)
- Ph. Preux gave a talk "décision adaptative face au Big Data", colloque AAFD, Institut Galilée (April 2014).

9.1.4. Evaluation activities, expertise

- *P. Chainais* is a grant proposal reviewer for the ANR.
- *M. Ghavamzadeh* is in the Editorial Board Member of Machine Learning Journal (MLJ, 2011-present).

- *M. Ghavamzadeh* is in the Steering Committee Member of the European Workshop on Reinforcement Learning (EWRL, 2011-present).
- *P. Preux* and *J. Mary* are experts for *Crédit Impôt Recherche* (CIR).
- *P. Preux* is expert for ANR, ANRT, AERES, FNRS. He was member on the visitng committee of the Laboratoire d'Informatique de Grenoble (LIG)
- *E. Duflos* is a project proposal reviewer for ANR.
- *A. Lazaric* is a project proposal reviewer for ANR.
- *A. Lazaric* is a the main organizer of the European Workshop in Reinforcement Learning in 2015.
- *A. Lazaric, J. Mary, R. Munos, O. Pietquin, and M. Valko* are members of the Belgium Commission Evaluation F.R.S-FNRS, 2014.
- *M. Valko* is an elected member of the evaluation committee and participates in the hiring, promotion, and evaluation juries of Inria.

9.1.5. Other Scientific Activities

- *D. Ryabko* is a member of COST-GTRI committee at Inria.
- *D. Ryabko* is a general advisor at Inria Lille.
- *E. Duflos* is Director of Research of Ecole Centrale de Lille since September 2011.
- *E. Duflos* is the Head of the Signal and Image Team of LAGIS (UMR CNRS 8219).
- *R. Gaudel* is board member of LIFL.
- *A. Lazaric* is a member of the committee for research evaluation (CER) at Inria Lille.
- *R. Gaudel* manages the proml mailing list. This mailing list gathers French-speaking researchers from Machine Learning community.
- *P. Chainais* is a member of the administration council of GRETSI, the French association of researchers in signal and image processing.
- *P. Chainais* is co-responsible for the action "Machine Learning" of the GDR ISIS which gathers french researchers in signal and image processing at the national level.
- *Ph. Preux* is Head of the LIFL/CRIStAL lab at the Université de Lille 3; he is head of the data intelligence (DatInG) thematic group of CRIStAL; he is on the scientific committee of CRIStAL. He is local organization chair for ICML 2015.

9.2. Teaching - Supervision - Juries

9.2.1. Awards

- D. Calandriello won the best master thesis award from the Italian Association for Artificial Intelligence for his thesis "Sparse Multi-Task Reinforcement Learning". The association awards the prize to the best master thesis focused on AI in Italy in 2014. The thesis was written under the co-supervision of A. Lazaric during a year spent in SequeL
- F. Guillou won the ACM RecSys challenge (on recommendation systems)
- P. Chainais won an IBM Faculty Award for the creation of the option DAD (Data Analysis and Decision making) at Ecole Centrale Lille (10000$ have been given to EC Lille). The partnership with IBM about Big Data is getting stronger and new perspectives are coming.

9.2.2. Teaching

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

Licence: R. Gaudel, projet informatique de traitement des données en SHS, 20h eqTD, L2, université Lille 3, France

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

Master: R. Gaudel, fouille du web, 32h eqTD, M2, université Lille 3, France

Master: R. Gaudel, fouille de donées, 30h eqTD, M2, université 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, "Modeling, Computer Science, Mathematics", 72h eqTD, M1 pshychology/cognitive science, université Lille 3, France

Master: Ph. Preux, "Formal neural netwokrs", 30h eqTD, M1 cognitive science, université Lille 3, France

Licence: Ph. Preux, "Supervised Learning", 30h eqTD, L3 MIASHS, université Lille 3, France

EC Lille (3rd y.): P. Chainais, "Machine learning", 34h eqTD, Ecole Centrale Lille, France

EC Lille (3rd y.): P. Chainais, "Matlab", 16h eqTD, Ecole Centrale Lille, France

EC Lille (3rd y.): P. Chainais, "Image processing", 16h eqTD, Ecole Centrale Lille, France

EC Lille (3rd y.): P. Chainais, "Representation and data compression", 8h eqTD, Ecole Centrale Lille, France

EC Lille (1st y.): P. Chainais, "Signal processing", 22h eqTD, Ecole Centrale Lille, France

EC Lille (2nd y.): P. Chainais, "Wavelets and applications", 24h eqTD, Ecole Centrale Lille, France

EC Lille: J. Mary, Machine Learning with R , 20h eqTD

Master: J. Mary, M2 ID - Univ Lille, Programmation web avancée et design pattern, 64h eqTD

Master: J. Mary, M1 ID - Univ Lille, Programmation web , 32h eqTD

Master: J. Mary, M1 ID - Univ Lille, Algorithmique avancée , 32h eqTD

Master: J. Mary, M1 IIES - Univ Lille, Analyse de données avec R, 32h eqTD

Master: J. Mary, C2i - Univ Lille, 24h eqTD

E-learning

Mooc, SPOC, etc. : Enseignant ou auteur, titre du cours, durée en nombre de semaine, plate-forme, établissement porteur du cours, public ciblé, formation initiale ou continue, nombre d'inscrits

Pedagogical resources : enseignant, titre, type (video, pdf, exercice, ou autre), niveau, url

SPOC : R. Gaudel, Marc Tommasi and Alain Preux, culture numérique S2, 8 semaines, Moodle, université Lille 3, licence (L1), formation initiale, tous les étudiants (> 7 000).

9.2.3. Supervision

HDR defended: *Mohammad Ghavamzadeh* defended his "Habilitation à diriger les recherches" on June 12th.

PhD defended: *Boris Baldassari* defended his PhD thesis *Apprentissage automatique et développement logiciel*, on July 1st, advisor: Ph. Preux.

PhD defended: *Gabriel Dulac-Arnold* defended his PhD thesis *A General Sequential Model for Constrained Classification*, on Feb. 7th, advisor: Ph. Preux, L. Denoyer (Paris 6), P. Gallinari (Paris 6).

PhD defended: *Victor Gabillon* defended his PhD thesis "Active Learning in Classification-based Policy Iteration", on June 12th, advisor: M. Ghavamzadeh.

PhD defended: *Olivier Nicol* defended his PhD thesis "Data-driven evaluation of Contextual Bandit algorithms and applications to Dynamic Recommendation", on Dec. 18th, advisor: Ph. Preux, J. Mary.

PhD defended: *Emilie Kaufmann* defended her PhD thesis, “Bayesian Bandits”, advisor: R. Munos, O. Cappé, A. Garivier.

PhD in progress: *Frédéric Guillou*, “Sequential Recommender System”, since Oct. 2013, advisor: Ph. Preux, J. Mary, R. Gaudel.

PhD in progress: *Vicenzo Musco*, “Topology and evolution of software graphs”, since Oct. 2013, advisor: P. Preux, M. Monperrus

PhD in progress: *Adrien Hoarau*, “Multi-arm Bandit Theory”, since Oct. 2012, advisor: R. Munos.

PhD in progress: *Tomáš Kocák*, “Sequential Learning with Similarities”, since Oct. 2013, advisor: R. Munos, M. Valko

PhD in progress: *Amir Sani*, “Learning under uncertainty”, Oct. 2011, since advisor: R. Munos, A. Lazaric.

PhD in progress: *Marta Soare*, “Pure Exploration in Multi-arm Bandit”, since Oct. 2012, advisor: R. Munos, A. Lazaric.

PhD in progress: *Hong Phuong Dang*, *Bayesian non parametric methods for dictionary learning and inverse problems*, since Oct. 2013, advisor: P. Chainais.

PhD in progress: *Linh Van Nguyen*, *High resolution reconstruction from low resolution measurements of velocity fields in turbulent flows*, since Oct. 2013, advisor: P. Chainais & J.P. Laval (Laboratoire de Mécanique de Lille).

PhD in progress: *Clément Elvira*, “Bayesian non parametric approaches for blind hyperspectral images unmixing.", since Oct. 2014, advisor: P. Chainais & N. Dobigeon (IRIT, Toulouse).

PhD started: *Daniele Calandriello*, *Efficient Sequential Learning in Structured and Constrained Environments*, since Oct. 2014, advisor: M. Valko & A. Lazaric & P. Preux.

PhD started: *Jean-Bastien Grill*, *Développement et analyse de mé thodes numé riques efficaces pour de l’optimisation lorsque la régularité de la fonction sous-jacente n’est pas connue à priori.*, since Oct. 2014, advisor: M. Valko & R. Munos

PhD started: *Pratik Gajane*, “Sequential Learning and Decision Making under Partial Monitoring”, since Oct. 2014, advisor: Philippe Preux, Tanguy Urvoy (Orange Labs)

9.2.4. Juries

A. Lazaric was part of the jury of the PhD of Mahdi Milani Fard at McGill Univiersity (supervised by J. Pineau).

Ph. Preux was part of the PhD defense jury of W. Wang (Université Paris-Sud, M. Martinez (Université de Lille), G. Dulac-Arnold (Université Paris 6), V. Gabillon, Boris Baldassari, and O. Nicol (all 3 from Université de Lille).

Ph. Preux was part of the HdR defense jury of M. Ghavamzadeh.

P. Chainais was part of the PhD defense jury of Raja Suleiman (supervied by David Mary) at University of Nice, dec. 2014.

9.3. Popularization

- M. Valko gave an Interview on "Face Recognition" at Sciences et Avenir (July 2014)
- M. Valko gave an Interview on "Biometric applications will soon be part of our daily life" at ARTE Future (November 2014)
- Article on research of M. Valko’s collaboration with INTEL - Ford and Intel Mobii project using Face Recognition, at engadget.com (June 2014) http://www.engadget.com/2014/06/25/ford-and-intel-project-mobii-connected-car-cameras/

- Article on research of M. Valko's collaboration with INTEL - Ford prototype using Face Recognition at intel.com (June 2014) http://www.intel.com/content/www/us/en/automotive/ford-mobii-prototype-video.html
- as part of the Inria mediation program, Ph. Preux met high schools pupils to explain what research is.

10. Bibliography

Publications of the year

Articles in International Peer-Reviewed Journals

[1] R. BUSA-FEKETE, W. CHENG, E. HÜLLERMEIER, B. SZÖRÉNYI, P. WENG. *Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm*, in "Machine Learning", December 2014, vol. 97, n° 3, pp. 327-351 [*DOI :* 10.1007/S10994-014-5458-8], https://hal.inria.fr/hal-01079370

[2] C. DHANJAL, R. GAUDEL, S. CLÉMENÇON. *Efficient Eigen-updating for Spectral Graph Clustering*, in "Neurocomputing", May 2014, vol. 131, pp. 440-452, Correction of several typos [*DOI :* 10.1016/J.NEUCOM.2013.11.015], https://hal.archives-ouvertes.fr/hal-00770889

[3] A. GYÖRGY, G. NEU. *Near-Optimal Rates for Limited-Delay Universal Lossy Source Coding*, in "IEEE Transactions on Information Theory", 2014, pp. 2823-2834 [*DOI :* 10.1109/TIT.2014.2307062], https://hal.archives-ouvertes.fr/hal-01079327

[4] G. NEU, A. GYÖRGY, C. SZEPESVÁRI, A. ANTOS. *Online Markov Decision Processes Under Bandit Feedback*, in "IEEE Transactions on Automatic Control", 2014, vol. 59, pp. 676 - 691 [*DOI :* 10.1109/TAC.2013.2292137], https://hal.archives-ouvertes.fr/hal-01079422

[5] 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

[6] D. RYABKO. *Uniform hypothesis testing for finite-valued stationary processes*, in "Statistics", 2014, vol. 48, n° 1, pp. 121-128 [*DOI :* 10.1080/02331888.2012.719511], https://hal.inria.fr/inria-00610009

[7] B. SCHERRER, M. GHAVAMZADEH, V. GABILLON, B. LESNER, M. GEIST. *Approximate Modified Policy Iteration and its Application to the Game of Tetris*, in "Journal of Machine Learning Research", 2015, 47 p. , A paraître, https://hal.inria.fr/hal-01091341

International Conferences with Proceedings

[8] R. BUSA-FEKETE, E. HÜLLERMEIER, B. SZÖRÉNYI. *Preference-Based Rank Elicitation using Statistical Models: The Case of Mallows*, in "Proceedings of The 31st International Conference on Machine Learning", Beijing, China, JMLR Workshop and Conference Proceedings Volume 32, June 2014, vol. 32, https://hal.inria.fr/hal-01079369

[9] D. CALANDRIELLO, A. LAZARIC, M. RESTELLI. *Sparse Multi-task Reinforcement Learning*, in "NIPS - Advances in Neural Information Processing Systems 26", Montreal, Canada, December 2014, https://hal.inria.fr/hal-01073513

[10] A. CARPENTIER, M. VALKO. *Extreme bandits*, in "Advances in Neural Information Processing Systems 27", Montréal, Canada, December 2014, https://hal.inria.fr/hal-01079354

[11] P. CHAINAIS, P. PFENNIG, A. LERAY. *Quantitative control of the error bounds of a fast super-resolution technique for microscopy and astronomy*, in "Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)", Florence, Italy, May 2014, pp. 2853 - 2857 [*DOI :* 10.1109/ICASSP.2014.6854121], https://hal.archives-ouvertes.fr/hal-01081402

[12] P. CHAINAIS, C. RICHARD. *A diffusion strategy for distributed dictionary learning*, in "2nd "international Traveling Workshop on Interactions between Sparse models and Technology" (iTWIST'14)", Namur, Belgium, Proceedings of the second "international Traveling Workshop on Interactions between Sparse models and Technology" (iTWIST'14), Laurent Jacques, August 2014, https://hal.archives-ouvertes.fr/hal-01104781

[13] E. DAUCÉ, E. THOMAS. *Evidence build-up facilitates on-line adaptivity in dynamic environments: example of the BCI P300-speller*, in "22nd European Symposium on Artificial Neural Networks", Bruges, Belgium, April 2014, https://hal.inria.fr/hal-01104024

[14] C. DHANJAL, R. GAUDEL, S. CLÉMENÇON. *Online Matrix Completion Through Nuclear Norm Regularisation*, in "SDM - SIAM International Conference on Data Mining", Philadelphia, United States, April 2014, Corrected a typo in the affiliation [*DOI :* 10.1137/1.9781611973440.72], https://hal.inria.fr/hal-00926605

[15] M. GHESHLAGHI AZAR, A. LAZARIC, E. BRUNSKILL. *Online Stochastic Optimization under Correlated Bandit Feedback*, in "31st International Conference on Machine Learning", Beijing, China, June 2014, https://hal.inria.fr/hal-01080138

[16] S. IVÁN, Á. D. LELKES, J. NAGY-GYÖRGY, B. SZÖRÉNYI, G. TURÁN. *Biclique Coverings, Rectifier Networks and the Cost of ε-Removal*, in "16th International Workshop on Descriptional Complexity of Formal Systems, Proceedings", Turku, Finland, August 2014, pp. 174 - 185 [*DOI :* 10.1007/978-3-319-09704-6_16], https://hal.inria.fr/hal-01079368

[17] A. KHALEGHI, D. RYABKO. *Asymptotically consistent estimation of the number of change points in highly dependent time series*, in "International Conference on Machine Learning (ICML)", Beijing, China, June 2014, pp. 539-547, https://hal.inria.fr/hal-01026583

[18] T. KOCÁK, G. NEU, M. VALKO, R. MUNOS. *Efficient learning by implicit exploration in bandit problems with side observations*, in "Advances in Neural Information Processing Systems 27", Montréal, Canada, December 2014, https://hal.inria.fr/hal-01079351

[19] T. KOCÁK, M. VALKO, R. MUNOS, S. AGRAWAL. *Spectral Thompson Sampling*, in "Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence", Québec City, Canada, July 2014, https://hal.inria.fr/hal-00981575

[20] T. KOCÁK, M. VALKO, R. MUNOS, B. KVETON, S. AGRAWAL. *Spectral Bandits for Smooth Graph Functions with Applications in Recommender Systems*, in "AAAI Workshop on Sequential Decision-Making with Big Data", Québec City, Canada, July 2014, https://hal.inria.fr/hal-01045036

[21] G. NEU, M. VALKO. *Online combinatorial optimization with stochastic decision sets and adversarial losses*, in "Advances in Neural Information Processing Systems 27", Montréal, Canada, December 2014, https://hal.inria.fr/hal-01079355

[22] O. NICOL, J. MARY, P. PREUX. *Improving offline evaluation of contextual bandit algorithms via bootstrapping techniques*, in "International Conference on Machine Learning", Beijing, China, E. XING, T. JEBARA (editors), Journal of Machine Learning Research, Workshop and Conference Proceedings; Proceedings of The 31st International Conference on Machine Learning, June 2014, vol. 32, https://hal.inria.fr/hal-00990840

[23] R. ORTNER, O.-A. MAILLARD, D. RYABKO. *Selecting Near-Optimal Approximate State Representations in Reinforcement Learning*, in "International Conference on Algorithmic Learning Theory (ALT)", Bled, Slovenia, LNCS, Springer, October 2014, vol. 8776, pp. 140-154, https://hal.inria.fr/hal-01057562

[24] O. PIETQUIN, H. GLAUDE, C. ENDERLI. *Subspace Identification for Predictive State Representation by Nuclear Norm Minimization*, in "IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning (ADPRL 2014)", Orlando, United States, December 2014, https://hal.inria.fr/hal-01104423

[25] B. PIOT, M. GEIST, O. PIETQUIN. *Difference of Convex Functions Programming for Reinforcement Learning*, in "Advances in Neural Information Processing Systems (NIPS 2014)", Montreal, Canada, December 2014, https://hal.inria.fr/hal-01104419

[26] B. PIOT, O. PIETQUIN, M. GEIST. *Predicting when to laugh with structured classification*, in "InterSpeech 2014", Singapore, Singapore, September 2014, pp. 1786-1790, https://hal-supelec.archives-ouvertes.fr/hal-01104739

[27] P. PREUX, R. MUNOS, M. VALKO. *Bandits attack function optimization*, in "IEEE Congress on Evolutionary Computation", Beijing, China, July 2014, https://hal.inria.fr/hal-00978637

[28] A. SANI, G. NEU, A. LAZARIC. *Exploiting easy data in online optimization*, in "Advances in Neural Information Processing 27", Montreal, Canada, December 2014, https://hal.archives-ouvertes.fr/hal-01079428

[29] M. SOARE, A. LAZARIC, R. MUNOS. *Best-Arm Identification in Linear Bandits*, in "NIPS - Advances in Neural Information Processing Systems 27", Montreal, Canada, December 2014, https://hal.inria.fr/hal-01075701

[30] B. SZÖRÉNYI, G. KEDENBURG, R. MUNOS. *Optimistic planning in Markov decision processes using a generative model*, in "Advances in Neural Information Processing Systems 27", Montréal, Canada, December 2014, https://hal.inria.fr/hal-01079366

[31] E. THOMAS, E. DAUCÉ, D. DEVLAMINCK, L. MAHÉ, A. CARPENTIER, R. MUNOS, M. PERRIN, E. MABY, J. MATTOUT, T. PAPADOPOULO, M. CLERC. *CoAdapt P300 speller: optimized flashing sequences and online learning*, in "6th International Brain Computer Interface Conference", Graz, Austria, September 2014, https://hal.inria.fr/hal-01103441

[32] M. VALKO, R. MUNOS, B. KVETON, T. KOCÁK. *Spectral Bandits for Smooth Graph Functions*, in "31th International Conference on Machine Learning", Beijing, China, May 2014, https://hal.inria.fr/hal-00986818

National Conferences with Proceedings

[33] M. PACHEBAT, N. TOTARO, P. CHAINAIS, O. COLLERY. *Synthèse en espace et temps du rayonnement acoustique d'une paroi sous excitation turbulente par synthèse spectrale 2D+T et formulation vibro-acoustique directe*, in "Congrès Français d'acoustique 2014", Poitiers, France, April 2014, 6 pages, p1921, papier N183 p. , 6 Pages, 20 Refs, https://hal.archives-ouvertes.fr/hal-01058151

Conferences without Proceedings

[34] B. PIOT, M. GEIST, O. PIETQUIN. *Méthode de minimisation du résidu de Bellman boostée qui tient compte des démonstrations expertes.*, in "9èmes Journées Francophones de Planification, Décision et Apprentissage (JFPDA'14)", Liège, Belgium, May 2014, https://hal-supelec.archives-ouvertes.fr/hal-01104789

Research Reports

[35] J. MARY, R. GAUDEL, P. PREUX. *Bandits Warm-up Cold Recommender Systems*, Inria Lille, July 2014, n° RR-8563, 18 p. , https://hal.inria.fr/hal-01022628

[36] 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

[37] V. MUSCO, M. MONPERRUS, P. PREUX. *A Generative Model of Software Dependency Graphs to Better Understand Software Evolution*, Inria, 2014, https://hal.archives-ouvertes.fr/hal-01078716

Other Publications

[38] P. CHAINAIS, A. LERAY. *Statistical performance analysis of a fast super-resolution technique using noisy translations*, November 2014, 15 pages, submitted, https://hal.archives-ouvertes.fr/hal-01104759

[39] F. GUILLOU, R. GAUDEL, J. MARY, P. PREUX. *User Engagement as Evaluation: a Ranking or a Regression Problem?*, October 2014, 1. Introduction 2. Recsys Challenge 2014: Data and Protocol 2.1 Data Characteristics and Statistics 2.2 About User Engagement as Evaluation 2.3 Input Features for the Model 3. Method 3.1 LambdaMART Model 3.2 Random Forests 3.3 Description of the Approach 4. Experiments 4.1 Experimental results 4.2 Relevant Features 5. Discussions 6. Conclusions 7. Acknowledgments 8. References [*DOI :* 10.1145/2668067.2668073], https://hal.inria.fr/hal-01077986

References in notes

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

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

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

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

[44] T. FERGUSON. *A Bayesian Analysis of Some Nonparametric Problems*, in "The Annals of Statistics", 1973, vol. 1, n° 2, pp. 209–230

[45] T. HASTIE, R. TIBSHIRANI, J. FRIEDMAN. *The elements of statistical learning — Data Mining, Inference, and Prediction*, Springer, 2001

[46] P. NGUYEN, O.-A. MAILLARD, D. RYABKO, R. ORTNER. *Competing with an Infinite Set of Models in Reinforcement Learning*, in "AISTATS", Arizona, United States, JMLR W&CP, 2013, vol. 31, pp. 463-471, https://hal.inria.fr/hal-00823230

[47] W. POWELL. *Approximate Dynamic Programming*, Wiley, 2007

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

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

[50] J. RUST. *How Social Security and Medicare Affect Retirement Behavior in a World of Incomplete Market*, in "Econometrica", July 1997, vol. 65, n^{o} 4, pp. 781–831, http://gemini.econ.umd.edu/jrust/research/rustphelan.pdf

[51] J. RUST. *On the Optimal Lifetime of Nuclear Power Plants*, in "Journal of Business & Economic Statistics", 1997, vol. 15, n^{o} 2, pp. 195–208

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

[53] G. TESAURO. *Temporal Difference Learning and TD-Gammon*, in "Communications of the ACM", March 1995, vol. 38, n^{o} 3

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