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

Ecole Centrale de Lille

Activity Report 2013

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, Sequential Decision Making, Inference

Creation of the Project-Team: 2007 July 01.

1. Members

Research Scientists

Mohammad Ghavamzadeh [Inria, Researcher, on leave from Inria since October 2013, working in Adobe Research, San Jose, CA]
Alessandro Lazaric [Inria, Researcher]
Rémi Munos [Inria, Senior Researcher, full secondment with MSR (Boston) since July 2013 (until June 2014), HdR]
Daniil Ryabko [Inria, Researcher, HdR]
Michal Valko [Inria, Researcher]

Faculty Members

Philippe Preux [Team leader, Université Lille 3, Professor, HdR]
Pierre Chainais [Ecole Centrale Lille, Associate Professor, HdR]
Rémi Coulom [Université Lille 3, Associate Professor]
Emmanuel Duflos [Ecole Centrale Lille, Professor, HdR]
Romaric Gaudel [Université Lille 3, Associate Professor]
Jérémie Mary [Université Lille 3, Associate Professor]
Philippe Vanheeghe [Ecole Centrale Lille, Professor, HdR]

External Collaborator

Olivier Pietquin [Université Lille 1, Professor, since Oct 2013, HdR]

Engineers

Romain Laby [Inria, granted by Technologies Broadcasting System, since Mar 2013 until Nov 2013]
Eoin Thomas [Inria, until Oct 2013]

PhD Students

Boris Baldassari [Squoring Technologies]
Victor Gabillon [Université Lille 1]
Frédéric Guillou [Inria, granted by Inria and Région Nord Pas de Calais, since Oct 2013]
Adrien Hoarau [Inria, granted by STREP/Complacs]
Azadeh Khaleghi [Inria, until Oct 2013]
Tomáš Kocák [Inria, granted by Inria, since Oct 2013]
Vincenzo Musco [Université Lille 1, granted by Université Lille 1 and Université Lille 3, since Oct 2013, also member of Adam team-project]
Sami Naamane [Orange Labs, until May 2013]
Olivier Nicol [Université Lille 1]
Amir Sani [Inria, granted by Inria and Région Nord Pas de Calais]
Marta Soare [Inria, granted by Inria and Région Nord Pas de Calais]

Post-Doctoral Fellows

Raphael Fonteneau [FNRS, until Aug 2013]
Nathaniel Korda [Inria, granted by STREP/Complacs, until Sep 2013]
Prashanth Lakshmanrao Anantha Padmanabha [Inria, granted by STREP/Complacs]
Gergely Neu [Inria, granted by ERCIM and Inria, since Sep 2013]
Thanh Hai Nguyen [Inria, from Jan 2013 to Oct 2013]

Balázs Szörényi [Inria, granted by STREP/Complacs]

Visiting Scientists

Gabriel Dulac Arnold [Université Pierre & Marie Curie, PhD student at LIP'6, from Mar 2013 to May 2013]
Gunnar Kedenburg [Berlin Institute of Technology, PhD student at idalab, from Jun 2013 to Nov 2013]

Administrative Assistant

Amélie Supervielle [Inria]

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 [45]). 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.

2.2. Highlights of the Year

- In 2013, Crazy Stone won the 6th edition of the UEC Cup and the first edition of the Denseisen. Crazy Stone is a Go-playing program developed by Rémi Coulom since 2005, based on the Monte Carlo Tree Search method. The UEC Cup is the most important international computer-Go competition, organized yearly by the University of Electro-Communications in Tokyo, Japan. The Denseisen is a match between the winner of the UEC Cup and a top Japanese profesionnal Go player. This year Crazy Stone won a game with 4 stones of handicap against 9-dan profesionnal player Yoshio Ishida.
- The International Machine Learning Society selects SEQUEL to organize the 32$^{\text{nd}}$ International Conference on Machine Learning in 2015 at Lille. ICML is the most important conference in the field of machine learning.

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 [41].

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 $r_t = r(x_t, a_t, x_{t+1})$.

[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.

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

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

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

where $\mathbb{E}$ is the expectation operator and $\gamma \in (0, 1)$ is the discount factor. This value function V^π gives an evaluation of the performance of a given policy π. Other functionals of the sequence of future rewards may be considered, such as the undiscounted reward (see the stochastic shortest path problems [37]) 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) [35], 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 [36]. 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 ([47]):

- 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 [42], 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.* [34] introduced the algorithm UCB (Upper Confidence Bounds) that follows what is now called the "optimism in the face of uncertainty principle". Their algorithm works by computing upper confidence bounds for all the arms and then choosing the arm with the highest such bound. They proved that the expected regret of their algorithm increases at most at a logarithmic rate with the number of trials, and that the algorithm achieves the smallest possible regret up to some sub-logarithmic factor (for the considered family of distributions).

3.3. Statistical analysis of time series

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

3.3.1. Prediction of Sequences of Structured and Unstructured Data

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

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

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

3.3.2. Hypothesis testing

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

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

3.3.3. Change Point Analysis

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

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

3.3.4. Clustering Time Series, Online and Offline

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

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

3.3.5. Online Semi-Supervised Learning

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

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

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

3.4. 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 depend 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 [39]. 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 [38] 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;

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

- $f_{k|k}(W|Z^{(k)})$ 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.

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,
- medical applications,
- 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 [40]. 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 softwares: in the 1990's, RL has been the basis of a very successful Backgammon program, TD-Gammon [46] 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 [40]
- 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 [44], [43] 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. Medical Applications

One of the initial motivations of the multi-arm bandit theory stems from clinical trials when one researches the effects of different treatments while maximizing the improvement of the patients' health states.

Medical health-care and in particular patient-management is up today one of the most important applications of the sequential decision making. This is because the treatment of the more complex health problems is typically sequential: A physician repeatedly observes the current state of the patient and makes the decision in order to improve the health condition as measured for example by *qualys* (quality adjusted life years).

Moreover, machine learning methods may be used for at least two means in neuroscience:

1. as in any other (experimental) scientific domain, the machine learning methods relying heavily on statistics, they may be used to analyse experimental data,
2. dealing with induction learning, that is the ability to generalize from facts which is an ability that is considered to be one of the basic components of "intelligence", machine learning may be considered as a model of learning in living beings. In particular, the temporal difference methods for reinforcement learning have strong ties with various concepts of psychology (Thorndike's law of effect, and the Rescorla-Wagner law to name the two most well-known).

4.5. 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.

4.6. 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. 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/

6. New Results

6.1. Decision-making Under Uncertainty

6.1.1. Reinforcement Learning

Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model [2]

We consider the problem of learning the optimal action-value function in discounted-reward Markov decision processes (MDPs). We prove new PAC bounds on the sample-complexity of two well-known model-based reinforcement learning (RL) algorithms in the presence of a generative model of the MDP: value iteration and policy iteration. The first result indicates that for an MDP with N state-action pairs and the discount factor γin[0, 1) only O(N log(N/δ)/ [(1 - γ)3 ϵ2]) state-transition samples are required to find an ϵ-optimal estimation of the action-value function with the probability (w.p.) 1-δ. Further, we prove that, for small values of ϵ, an order of O(N log(N/δ)/ [(1 - γ)3 ϵ2]) samples is required to find an ϵ-optimal policy w.p. 1-δ. We also prove a matching lower bound of Ω(N log(N/δ)/ [(1 - γ)3 ϵ2]) on the sample complexity of estimating the optimal action-value function. To the best of our knowledge, this is the first minimax result on the sample complexity of RL: The upper bound matches the lower bound interms of N , ϵ, δ and 1/(1 -γ) up to a constant factor. Also, both our lower bound and upper bound improve on the state-of-the-art in terms of their dependence on 1/(1-γ).

Regret Bounds for Reinforcement Learning with Policy Advice [13]

In some reinforcement learning problems an agent may be provided with a set of input policies, perhaps learned from prior experience or provided by advisors. We present a reinforcement learning with policy advice (RLPA) algorithm which leverages this input set and learns to use the best policy in the set for the reinforcement learning task at hand. We prove that RLPA has a sub-linear regret of $\widetilde{O}(\sqrt{T})$ relative to the best input policy, and that both this regret and its computational complexity are independent of the size of the state and action space. Our empirical simulations support our theoretical analysis. This suggests RLPA may offer significant advantages in large domains where some prior good policies are provided.

Optimistic planning for belief-augmented Markov decision processes [11]

This paper presents the Bayesian Optimistic Planning (BOP) algorithm, a novel model-based Bayesian reinforcement learning approach. BOP extends the planning approach of the Optimistic Planning for Markov Decision Processes (OP-MDP) algorithm [10], [9] to contexts where the transition model of the MDP is initially unknown and progressively learned through interactions within the environment. The knowledge about the unknown MDP is represented with a probability distribution over all possible transition models using Dirichlet distributions, and the BOP algorithm plans in the belief-augmented state space constructed by concatenating the original state vector with the current posterior distribution over transition models. We show that BOP becomes Bayesian optimal when the budget parameter increases to infinity. Preliminary empirical validations show promising performance.

Aggregating optimistic planning trees for solving markov decision processes [16]

This paper addresses the problem of online planning in Markov decision processes using a generative model and under a budget constraint. We propose a new algorithm, ASOP, which is based on the construction of a forest of single successor state planning trees, where each tree corresponds to a random realization of the stochastic environment. The trees are explored using a "safe" optimistic planning strategy which combines the optimistic principle (in order to explore the most promising part of the search space first) and a safety principle (which guarantees a certain amount of uniform exploration). In the decision-making step of the algorithm, the individual trees are aggregated and an immediate action is recommended. We provide a finite-sample analysis and discuss the trade-off between the principles of optimism and safety. We report numerical results on a benchmark problem showing that ASOP performs as well as state-of-the-art optimistic planning algorithms.

Optimal Regret Bounds for Selecting the State Representation in Reinforcement Learning [20]

We consider an agent interacting with an environment in a single stream of actions, observations, and rewards, with no reset. This process is not assumed to be a Markov Decision Process (MDP). Rather, the agent has several representations (mapping histories of past interactions to a discrete state space) of the environment with unknown dynamics, only some of which result in an MDP. The goal is to minimize the average regret criterion against an agent who knows an MDP representation giving the highest optimal reward, and acts optimally in it. Recent regret bounds for this setting are of order $O(T^{2/3})$ with an additive term constant yet exponential in some characteristics of the optimal MDP. We propose an algorithm whose regret after T time steps is $O(\sqrt{T})$, with all constants reasonably small. This is optimal in T since $O(\sqrt{T})$ is the optimal regret in the setting of learning in a (single discrete) MDP.

Competing with an Infinite Set of Models in Reinforcement Learning [21]

We consider a reinforcement learning setting where the learner also has to deal with the problem of finding a suitable state-representation function from a given set of models. This has to be done while interacting with the environment in an online fashion (no resets), and the goal is to have small regret with respect to any Markov model in the set. For this setting, recently the BLBãlgorithm has been proposed, which achieves regret of order $T^{2/3}$, provided that the given set of models is finite. Our first contribution is to extend this result to a countably infinite set of models. Moreover, the BLBr˜egret bound suffers from an additive term that can be exponential in the diameter of the MDP involved, since the diameter has to be guessed. The algorithm we propose avoids guessing the diameter, thus improving the regret bound.

A review of optimistic planning in Markov decision processes [30]

We review a class of online planning algorithms for deterministic and stochastic optimal control problems, modeled as Markov decision processes. At each discrete time step, these algorithms maximize the predicted value of planning policies from the current state, and apply the first action of the best policy found. An overall receding-horizon algorithm results, which can also be seen as a type of model-predictive control. The space of planning policies is explored optimistically, focusing on areas with largest upper bounds on the value - or upper confidence bounds, in the stochastic case. The resulting optimistic planning framework integrates several types of optimism previously used in planning, optimization, and reinforcement learning, in order to obtain several intuitive algorithms with good performance guarantees. We describe in detail three recent such

algorithms, outline the theoretical guarantees on their performance, and illustrate their behavior in a numerical example.

6.1.2. Multi-arm Bandit Theory

Automatic motor task selection via a bandit algorithm for a brain-controlled button [4]

Objective. Brain-computer interfaces (BCIs) based on sensorimotor rhythms use a variety of motor tasks, such as imagining moving the right or left hand, the feet or the tongue. Finding the tasks that yield best performance, specifically to each user, is a time-consuming preliminary phase to a BCI experiment. This study presents a new adaptive procedure to automatically select (online) the most promising motor task for an asynchronous brain-controlled button. Approach. We develop for this purpose an adaptive algorithm UCB-classif based on the stochastic bandit theory and design an EEG experiment to test our method. We compare (offline) the adaptive algorithm to a naïve selection strategy which uses uniformly distributed samples from each task. We also run the adaptive algorithm online to fully validate the approach. Main results. By not wasting time on inefficient tasks, and focusing on the most promising ones, this algorithm results in a faster task selection and a more efficient use of the BCI training session. More precisely, the offline analysis reveals that the use of this algorithm can reduce the time needed to select the most appropriate task by almost half without loss in precision, or alternatively, allow us to investigate twice the number of tasks within a similar time span. Online tests confirm that the method leads to an optimal task selection. Significance. This study is the first one to optimize the task selection phase by an adaptive procedure. By increasing the number of tasks that can be tested in a given time span, the proposed method could contribute to reducing 'BCI illiteracy'.

Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation [3]

We consider optimal sequential allocation in the context of the so-called stochastic multi-armed bandit model. We describe a generic index policy, in the sense of Gittins (1979), based on upper confidence bounds of the arm payoffs computed using the Kullback-Leibler divergence. We consider two classes of distributions for which instances of this general idea are analyzed: The kl-UCB algorithm is designed for one-parameter exponential families and the empirical KL-UCB algorithm for bounded and finitely supported distributions. Our main contribution is a unified finite-time analysis of the regret of these algorithms that asymptotically matches the lower bounds of Lai and Robbins (1985) and Burnetas and Katehakis (1996), respectively. We also investigate the behavior of these algorithms when used with general bounded rewards, showing in particular that they provide significant improvements over the state-of-the-art.

Sequential Transfer in Multi-armed Bandit with Finite Set of Models [14]

Learning from prior tasks and transferring that experience to improve future performance is critical for building lifelong learning agents. Although results in supervised and reinforcement learning show that transfer may significantly improve the learning performance, most of the literature on transfer is focused on batch learning tasks. In this paper we study the problem of *sequential transfer in online learning*, notably in the multi–armed bandit framework, where the objective is to minimize the total regret over a sequence of tasks by transferring knowledge from prior tasks. Under the assumption that the tasks are drawn from a stationary distribution over a finite set of models, we define a novel bandit algorithm based on a method-of-moments approach for the estimation of the possible tasks and derive regret bounds for it. We introduce a novel bandit algorithm based on a method-of-moments approach for estimating the possible tasks and derive regret bounds for it. Finally, we report preliminary empirical results confirming the theoretical findings.

Optimizing P300-speller sequences by RIP-ping groups apart [25]

So far P300-speller design has put very little emphasis on the design of optimized flash patterns, a surprising fact given the importance of the sequence of flashes on the selection outcome. Previous work in this domain has consisted in studying consecutive flashes, to prevent the same letter or its neighbors from flashing consecutively. To this effect, the flashing letters form more random groups than the original row-column sequences for the P300 paradigm, but the groups remain fixed across repetitions. This has several important consequences, among which a lack of discrepancy between the scores of the different letters. The new approach

proposed in this paper accumulates evidence for individual elements, and optimizes the sequences by relaxing the constraint that letters should belong to fixed groups across repetitions. The method is inspired by the theory of Restricted Isometry Property matrices in Compressed Sensing, and it can be applied to any display grid size, and for any target flash frequency. This leads to P300 sequences which are shown here to perform significantly better than the state of the art, in simulations and online tests.

Stochastic Simultaneous Optimistic Optimization [26]

We study the problem of global maximization of a function f given a finite number of evaluations perturbed by noise. We consider a very weak assumption on the function, namely that it is locally smooth (in some precise sense) with respect to some semi-metric, around one of its global maxima. Compared to previous works on bandits in general spaces (Kleinberg et al., 2008; Bubeck et al., 2011a) our algorithm does not require the knowledge of this semi-metric. Our algorithm, StoSOO, follows an optimistic strategy to iteratively construct upper confidence bounds over the hierarchical partitions of the function domain to decide which point to sample next. A finite-time analysis of StoSOO shows that it performs almost as well as the best specifically-tuned algorithms even though the local smoothness of the function is not known.

Toward optimal stratification for stratified monte-carlo integration [9]

We consider the problem of adaptive stratified sampling for Monte Carlo integration of a noisy function, given a finite budget n of noisy evaluations to the function. We tackle in this paper the problem of adapting to the function at the same time the number of samples into each stratum and the partition itself. More precisely, it is interesting to refine the partition of the domain in area where the noise to the function, or where the variations of the function, are very heterogeneous. On the other hand, having a (too) refined stratification is not optimal. Indeed, the more refined the stratification, the more difficult it is to adjust the allocation of the samples to the stratification, i.e. sample more points where the noise or variations of the function are larger. We provide in this paper an algorithm that selects online, among a large class of partitions, the partition that provides the optimal trade-off, and allocates the samples almost optimally on this partition

Thompson sampling for one-dimensional exponential family bandits [18]

Thompson Sampling has been demonstrated in many complex bandit models, however the theoretical guarantees available for the parametric multi-armed bandit are still limited to the Bernoulli case. Here we extend them by proving asymptotic optimality of the algorithm using the Jeffreys prior for 1-dimensional exponential family bandits. Our proof builds on previous work, but also makes extensive use of closed forms for Kullback-Leibler divergence and Fisher information (and thus Jeffreys prior) available in an exponential family. This allow us to give a finite time exponential concentration inequality for posterior distributions on exponential families that may be of interest in its own right. Moreover our analysis covers some distributions for which no optimistic algorithm has yet been proposed, including heavy-tailed exponential families.

Finite-Time Analysis of Kernelised Contextual Bandits [27]

We tackle the problem of online reward maximisation over a large finite set of actions described by their contexts. We focus on the case when the number of actions is too big to sample all of them even once. However we assume that we have access to the similarities between actions' contexts and that the expected reward is an arbitrary linear function of the contexts' images in the related reproducing kernel Hilbert space (RKHS). We propose KernelUCB, a kernelised UCB algorithm, and give a cumulative regret bound through a frequentist analysis. For contextual bandits, the related algorithm GP-UCB turns out to be a special case of our algorithm, and our finite-time analysis improves the regret bound of GP-UCB for the agnostic case, both in the terms of the kernel-dependent quantity and the RKHS norm of the reward function. Moreover, for the linear kernel, our regret bound matches the lower bound for contextual linear bandits.

From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning [33]

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.

6.2. Statistical analysis of time series

6.2.1. Change Point Analysis

Nonparametric multiple change point estimation in highly dependent time series [17]

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

6.2.2. Clustering Time Series, Online and Offline

A Binary-Classification-Based Metric between Time-Series Distributions and Its Use in Statistical and Learning Problems [6]

A metric between time-series distributions is proposed that can be evaluated using binary classification methods, which were originally developed to work on i.i.d. data. It is shown how this metric can be used for solving statistical problems that are seemingly unrelated to classification and concern highly dependent time series. Specifically, the problems of time-series clustering, homogeneity testing and the three-sample problem are addressed. Universal consistency of the resulting algorithms is proven under most general assumptions. The theoretical results are illustrated with experiments on synthetic and real-world data.

6.2.3. Semi-Supervised and Unsupervised Learning

Learning from a Single Labeled Face and a Stream of Unlabeled Data [19]

Face recognition from a single image per person is a challenging problem because the training sample is extremely small. We consider a variation of this problem. In our problem, we recognize only one person, and there are no labeled data for any other person. This setting naturally arises in authentication on personal computers and mobile devices, and poses additional challenges because it lacks negative examples. We formalize our problem as one-class classification, and propose and analyze an algorithm that learns a non-parametric model of the face from a single labeled image and a stream of unlabeled data. In many domains, for instance when a person interacts with a computer with a camera, unlabeled data are abundant and easy to utilize. This is the first paper that investigates how these data can help in learning better models in the single-image-per-person setting. Our method is evaluated on a dataset of 43 people and we show that these people can be recognized 90% of time at nearly zero false positives. This recall is 25+% higher than the recall of our best performing baseline. Finally, we conduct a comprehensive sensitivity analysis of our algorithm and provide a guideline for setting its parameters in practice.

Unsupervised model-free representation learning [23]

Numerous control and learning problems face the situation where sequences of high-dimensional highly dependent data are available, but no or little feedback is provided to the learner. In such situations it may be useful to find a concise representation of the input signal, that would preserve as much as possible of the relevant information. In this work we are interested in the problems where the relevant information is in the time-series dependence. Thus, the problem can be formalized as follows. Given a series of observations $X\_0, \cdots, X\_n$ coming from a large (high-dimensional) space $\mathcal{X}$, find a representation function f mapping $\mathcal{X}$ to a finite space $\mathcal{Y}$ such that the series $f(X\_0), \cdots, f(X\_n)$ preserve as much information as possible about the original time-series dependence in $X\_0, \cdots, X\_n$. For stationary time series, the function f can be selected as the one maximizing the time-series information $I\_\infty(f) = h\_0(f(X)) - h\_\infty(f(X))$ where $h\_0(f(X))$ is the Shannon entropy of $f(X\_0)$ and $h\_\infty(f(X))$ is the entropy rate of the time series $f(X\_0), \cdots, f(X\_n), \cdots$. In this paper we study the functional $I\_\infty(f)$ from the learning-theoretic point of view. Specifically, we provide some uniform approximation results, and study the behaviour of $I\_\infty(f)$ in the problem of optimal control.

Time-series information and learning [22]

Given a time series $X\_1, \cdots, X\_n, \cdots$ taking values in a large (high-dimensional) space $\mathcal{X}$, we would like to find a function f from $\mathcal{X}$ to a small (low-dimensional or finite) space $\mathcal{Y}$ such that the time series $f(X\_1), \cdots, f(X\_n), \cdots$ retains all the information about the time-series dependence in the original sequence, or as much as possible thereof. This goal is formalized in this work, and it is shown that the target function f can be found as the one that maximizes a certain quantity that can be expressed in terms of entropies of the series $(f(X\_i))\_i \in \mathcal{N}$. This quantity can be estimated empirically, and does not involve estimating the distribution on the original time series $(X\_i)\_i \in \mathcal{N}$.

6.3. Statistical Learning and Bayesian Analysis

6.3.1. Dictionary learning

Learning a common dictionary over a sensor network [10]

We consider the problem of distributed dictionary learning, where a set of nodes is required to collectively learn a common dictionary from noisy measurements. This approach may be useful in several contexts including sensor networks. Diffusion cooperation schemes have been proposed to solve the distributed linear regression problem. In this work we focus on a diffusion-based adaptive dictionary learning strategy: each node records independent observations and cooperates with its neighbors by sharing its local dictionary. The resulting algorithm corresponds to a distributed alternate optimization. Beyond dictionary learning, this strategy could be adapted to many matrix factorization problems in various settings. We illustrate its efficiency on some numerical experiments.

Distributed dictionary learning over a sensor network [29]

We consider the problem of distributed dictionary learning, where a set of nodes is required to collec- tively learn a common dictionary from noisy measure- ments. This approach may be useful in several con- texts including sensor networks. Diffusion cooperation schemes have been proposed to solve the distributed linear regression problem. In this work we focus on a diffusion-based adaptive dictionary learning strategy: each node records observations and cooperates with its neighbors by sharing its local dictionary. The resulting algorithm corresponds to a distributed block coordi- nate descent (alternate optimization). Beyond dictio- nary learning, this strategy could be adapted to many matrix factorization problems and generalized to var- ious settings. This article presents our approach and illustrates its efficiency on some numerical examples.

6.4. Applications

6.4.1. Medical Applications

Outlier detection for patient monitoring and alerting [5]

We develop and evaluate a data-driven approach for detecting unusual (anomalous) patient-management decisions using past patient cases stored in electronic health records (EHRs). Our hypothesis is that a patient-management decision that is unusual with respect to past patient care may be due to an error and that it is worthwhile to generate an alert if such a decision is encountered. We evaluate this hypothesis using data obtained from EHRs of 4486 post-cardiac surgical patients and a subset of 222 alerts generated from the data. We base the evaluation on the opinions of a panel of experts. The results of the study support our hypothesis that the outlier-based alerting can lead to promising true alert rates. We observed true alert rates that ranged from 25% to 66% for a variety of patient-management actions, with 66% corresponding to the strongest outliers.

6.5. Miscellaneous

6.5.1. Miscellaneous

A confidence-set approach to signal denoising [7]

The problem of filtering of finite-alphabet stationary ergodic time series is considered. A method for constructing a confidence set for the (unknown) signal is proposed, such that the resulting set has the following properties. First, it includes the unknown signal with probability γ, where γ is a parameter supplied to the filter. Second, the size of the confidence sets grows exponentially with a rate that is asymptotically equal to the conditional entropy of the signal given the data. Moreover, it is shown that this rate is optimal. We also show that the described construction of the confidence set can be applied to the case where the signal is corrupted by an erasure channel with unknown statistics.

Quantification adaptative pour la stéganalyse d'images texturées [28]

Nous cherchons à améliorer les performances d'un schéma de stéganalyse (i.e. la détection de messages cachées) pour des images texturées. Le schéma de stéganographie étudié consiste à modifier certains pixels de l'image par une perturbation +/-1, et le schéma de stéganalyse utilise les caractéristiques construites à partir de la probabilité conditionnelle empirique de différences de 4 pixels voisins. Dans sa version originale, la stéganalyse n'est pas trés efficace sur des images texturées et ce travail vise àâ€ explorer plusieurs techniques de quantification en utilisant d'abord un pas de quantification plus important puis une quantification adaptative scalaire ou vectorielle. Les cellules de la quantification adaptative sont générées en utilisant un K-means ou un K-means "équilibré" de manière à ce chaque cellule quantifie approximativement le même nombre d'échantillon. Nous obtenons un gain maximal de classification de 3% pour un pas de quantification uniforme de 3. En utilisant l'algorithme K-means équilibré sur [-18,18], le gain par rapport à la version de base est de 4.7%.

Cost-sensitive Multiclass Classification Risk Bounds [8]

A commonly used approach to multiclass classification is to replace the 0-1 loss with a convex surrogate so as to make empirical risk minimization computationally tractable. Previous work has uncovered sufficient and necessary conditions for the consistency of the resulting procedures. In this paper, we strengthen these results by showing how the 0-1 excess loss of a predictor can be upper bounded as a function of the excess loss of the predictor measured using the convex surrogate. The bound is developed for the case of cost-sensitive multiclass classification and a convex surrogate loss that goes back to the work of Lee, Lin and Wahba. The bounds are as easy to calculate as in binary classification. Furthermore, we also show that our analysis extends to the analysis of the recently introduced "Simplex Coding" scheme.

Approximate Dynamic Programming Finally Performs Well in the Game of Tetris [12]

Tetris is a video game that has been widely used as a benchmark for various optimization techniques including approximate dynamic programming (ADP) algorithms. A look at the literature of this game shows that while ADP algorithms that have been (almost) entirely based on approximating the value function (value function based) have performed poorly in Tetris, the methods that search directly in the space of policies by learning the policy parameters using an optimization black box, such as the cross entropy (CE) method, have achieved the best reported results. This makes us conjecture that Tetris is a game in which good policies are easier to represent, and thus, learn than their corresponding value functions. So, in order to obtain a good performance with ADP, we should use ADP algorithms that search in a policy space, instead of the more traditional ones that search in a value function space. In this paper, we put our conjecture to test by applying such an ADP algorithm, called classification-based modified policy iteration (CBMPI), to the game of Tetris. Our experimental results show that for the first time an ADP algorithm, namely CBMPI, obtains the best results reported in the literature for Tetris in both small 10×10 and large 10×20 boards. Although the CBMPI's results are similar to those of the CE method in the large board, CBMPI uses considerably fewer (almost 1/6) samples (calls to the generative model) than CE.

A Generalized Kernel Approach to Structured Output Learning [15]

We study the problem of structured output learning from a regression perspective. We first provide a general formulation of the kernel dependency estimation (KDE) problem using operator-valued kernels. We show that some of the existing formulations of this problem are special cases of our framework. We then propose a covariance-based operator-valued kernel that allows us to take into account the structure of the kernel feature space. This kernel operates on the output space and encodes the interactions between the outputs without any reference to the input space. To address this issue, we introduce a variant of our KDE method based on the conditional covariance operator that in addition to the correlation between the outputs takes into account the effects of the input variables. Finally, we evaluate the performance of our KDE approach using both covariance and conditional covariance kernels on two structured output problems, and compare it to the state-of-the-art kernel-based structured output regression methods.

Gossip-based distributed stochastic bandit algorithms [24]

The multi-armed bandit problem has attracted remarkable attention in the machine learning community and many efficient algorithms have been proposed to handle the so-called exploitation-exploration dilemma in various bandit setups. At the same time, significantly less effort has been devoted to adapting bandit algorithms to particular architectures, such as sensor networks, multi-core machines, or peer-to-peer (P2P) environments, which could potentially speed up their convergence. Our goal is to adapt stochastic bandit algorithms to P2P networks. In our setup, the same set of arms is available in each peer. In every iteration each peer can pull one arm independently of the other peers, and then some limited communication is possible with a few random other peers. As our main result, we show that our adaptation achieves a linear speedup in terms of the number of peers participating in the network. More precisely, we show that the probability of playing a suboptimal arm at a peer in iteration t=Ω(logN) is proportional to 1/(Nt) where N denotes the number of peers. The theoretical results are supported by simulation experiments showing that our algorithm scales gracefully with the size of network.

Sur quelques problèmes non-supervisés impliquant des séries temporelles hautement dèpendantes [1]

Cette thèse est consacrée à l'analyse théorique de problèmes non supervisés impliquant des séries temporelles hautement dépendantes. Plus particulièrement, nous abordons les deux problèmes fondamentaux que sont le problème d'estimation des points de rupture et le partitionnement de séries temporelles. Ces problèmes sont abordés dans un cadre extrêmement général oùles données sont générées par des processus stochastiques ergodiques stationnaires. Il s'agit de l'une des hypothèses les plus faibles en statistiques, comprenant non seulement, les hypothèses de modèles et les hypothèses paramétriques habituelles dans la littérature scientifique, mais aussi des hypothèses classiques d'indépendance, de contraintes sur l'espace mémoire ou encore des hypothèses de mélange. En particulier, aucune restriction n'est faite sur la forme ou la nature des dépendances, de telles sortes que les échantillons peuvent être arbitrairement dépendants. Pour chaque problème abordé, nous proposons de nouvelles méthodes non paramétriques et nous prouvons de plus qu'elles sont, dans ce cadre, asymptotiquement consistantes. Pour l'estimation de points de rupture, la consistance asymptotique se rapporte à la capacité de l'algorithme à produire des estimations des points de rupture qui sont asymptotiquement arbitrairement proches des vrais points de rupture. D'autre part, un algorithme de partitionnement est asymptotiquement consistant si le partitionnement qu'il produit, restreint à chaque lot de séquences, coïncides, à partir d'un certain temps et de manière consistante, avec le partitionnement cible. Nous montrons que les algorithmes proposés sont implémentables efficacement, et nous accompagnons nos résultats théoriques par des évaluations expérimentales. L'analyse statistique dans le cadre stationnaire ergodique est extrêmement difficile. De manière générale, il est prouvé que les vitesses de convergence sont impossibles à obtenir. Dès lors, pour deux échantillons générés indépendamment par des processus ergodiques stationnaires, il est prouvé qu'il est impossible de distinguer le cas où les échantillons sont générés par le même processus de celui où ils sont générés par des processus différents. Ceci implique que des problèmes tels le partitionnement de séries temporelles sans la connaissance du nombre de partitions ou du nombre de points de rupture ne peut admettre de solutions consistantes. En conséquence, une tâche difficile est de découvrir les formulations du problème qui en permettent une résolution dans ce cadre général. La principale contribution de cette thèse est de démontrer (par construction) que malgré ces résultats d'impossibilités théoriques, des formulations naturelles des problèmes considérés existent et admettent des solutions consistantes dans ce cadre général. Ceci inclut la démonstration du fait que le nombre de points de rupture corrects peut être trouvé, sans recourir à des hypothèses plus fortes sur les processus stochastiques. Il en résulte que, dans cette formulation, le problème des points de rupture peut être réduit à du partitionnement de séries temporelles. Les résultats présentés dans ce travail formulent les fondations théoriques pour l'analyse des données séquentielles dans un espace d'applications bien plus large.

Actor-Critic Algorithms for Risk-Sensitive MDPs [32]

In many sequential decision-making problems we may want to manage risk by minimizing some measure of variability in rewards in addition to maximizing a standard criterion. Variance-related risk measures are among the most common risk-sensitive criteria in finance and operations research. However, optimizing many such criteria is known to be a hard problem. In this paper, we consider both discounted and average reward Markov decision processes. For each formulation, we first define a measure of variability for a policy, which in turn gives us a set of risk-sensitive criteria to optimize. For each of these criteria, we derive a formula for computing its gradient. We then devise actor-critic algorithms for estimating the gradient and updating the policy parameters in the ascent direction. We establish the convergence of our algorithms to locally risk-sensitive optimal policies. Finally, we demonstrate the usefulness of our algorithms in a traffic signal control application.

Bayesian Policy Gradient and Actor-Critic Algorithms [31]

Policy gradient methods are reinforcement learning algorithms that adapt a parameterized policy by following a performance gradient estimate. Many conventional policy gradient methods use Monte-Carlo techniques to estimate this gradient. The policy is improved by adjusting the parameters in the direction of the gradient estimate. Since Monte-Carlo methods tend to have high variance, a large number of samples is required to attain accurate estimates, resulting in slow convergence. In this paper, we first propose a Bayesian framework for policy gradient, based on modeling the policy gradient as a Gaussian process. This reduces the number of samples needed to obtain accurate gradient estimates. Moreover, estimates of the natural gradient as well as

a measure of the uncertainty in the gradient estimates, namely, the gradient covariance, are provided at little extra cost. Since the proposed Bayesian framework considers system trajectories as its basic observable unit, it does not require the dynamic within each trajectory to be of any special form, and thus, can be easily extended to partially observable problems. On the downside, it cannot take advantage of the Markov property when the system is Markovian. To address this issue, we then extend our Bayesian policy gradient framework to actor-critic algorithms and present a new actor-critic learning model in which a Bayesian class of non-parametric critics, based on Gaussian process temporal difference learning, is used. Such critics model the action-value function as a Gaussian process, allowing Bayes' rule to be used in computing the posterior distribution over action-value functions, conditioned on the observed data. Appropriate choices of the policy parameterization and of the prior covariance (kernel) between action-values allow us to obtain closed-form expressions for the posterior distribution of the gradient of the expected return with respect to the policy parameters. We perform detailed experimental comparisons of the proposed Bayesian policy gradient and actor-critic algorithms with classic Monte-Carlo based policy gradient methods, as well as with each other, on a number of reinforcement learning problems.

7. 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.
- TBS, 2012-2013
  Participants: Jérémie Mary, Philippe Preux.

  A research project has started in September 2012 in collaboration with the TBS company. The goal is to understand and predict the audience of news related websites. These websites tend to present an ergodic frequentation with respect to a context. The main goal is to separate the effect of the context (big events, elections, ...) and the impact of the policies of the news websites. This work is based on data originating from major French media websites and also involves research of tendencies on the web (as Google Trends and Google Flu do). Used algorithms mix methods from time series prediction (ARIMA and MARSS models) and machine learning methods (L1 penalization, SVM).
- 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.

- 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. National Initiatives

8.1.1. ANR-Lampada

Participants: Mohammad Ghavamzadeh, Jérémie Mary, Olivier Nicol, Philippe Preux, Daniil Ryabko.

- *Title*: Learning Algorithms, Models and sPArse representations for structured DAta
- *Type*: National Research Agency (ANR-09-EMER-007)
- *Coordinator*: Inria Lille – Nord Europe (Mostrare)
- *Others partners*: Laboratoire d'Informatique Fondamentale de Marseille; Laboratoire Hubert Curien à Saint Etienne; Laboratoire d'Informatique de Paris 6.
- *Web site*: http://lampada.gforge.inria.fr/
- *Duration*: ends mid-2014
- *Abstract*: Lampada is a fundamental research project on machine learning and structured data. Lampada focuses on scaling learning algorithms to handle large sets of complex data. The main challenges are 1) high dimension learning problems, 2) large sets of data and 3) dynamics of data. We consider evolving data. The representation of these data involves both structure and content information and are typically large sequences, trees and graphs. The main application domains are web2, social networks and biological data.

  The project proposes to study formal representations of such data together with incremental or sequential machine learning methods and similarity learning methods.

  The representation research topic includes condensed data representation, sampling, prototype selection and representation of streams of data. Machine learning methods include edit distance learning, reinforcement learning and incremental methods, density estimation of structured data and learning on streams.
- *Activity Report*:

Philippe Preux has collaborated with Ludovic Denoyer and Gabriel Dulac-Arnold from LIP'6 to investigate further the idea of datum-wise representation, introduced in 2011.

Mohammad Ghavamzadeh and Philippe Preux have collaborated with Hachem Kadri on an operator-based approach for structured output [15].

Daniil Ryabko has developed a theory for unsupervised learning of time-series dependence, where the time series are either coming from a stationary environment or are a result of interaction with a Markovian environment with a continuous state space. Danil Ryabko and Jeremie Mary have developed methods for using binary classification methods for solving various unsupervised learning problems about time series.

8.1.2. ANR CO-ADAPT

Participant: Rémi Munos.

- *Title*: Brain computer co-adaptation for better interfaces
- *Type*: National Research Agency (ANR-09-EMER-002)
- *Coordinator*: Maureen Clerc
- *Other Partners*: Inria Odyssee project (Maureen Clerc), the INSERM U821 team (Olivier Bertrand), the Laboratory of Neurobiology of Cognition (CNRS) (Boris Burle) and the laboratory of Analysis, topology and probabilities (CNRS and University of Provence) (Bruno Torresani).
- *Web site*: https://twiki-sop.inria.fr/twiki/bin/view/Projets/Athena/CoAdapt/WebHome
- *Duration*: 2009-2014
- *Abstract*: The aim of Co-Adapt is to propose new directions for BCI design, by modeling explicitly the co-adaptation taking place between the user and the system. The goal of CoAdapt is to study the co-adaptation between a user and a BCI system in the course of training and operation. The quality of the interface will be judged according to several criteria (reliability, learning curve, error correction, bit rate). BCI will be considered under a joint perspective: the user's and the system's. From the user's brain activity, features must be extracted, and translated into commands to drive the BCI system. From the point of view of the system, it is important to devise adaptive learning strategies, because the brain activity is not stable in time. How to adapt the features in the course of BCI operation is a difficult and important topic of research. We will investigate Reinforcement Learning (RL) techniques to address the above questions.
- *Activity Report*: The performances of a BCI can vary greatly across users but also depend on the tasks used, making the problem of appropriate task selection an important issue. We develop an adaptive algorithm, UCB-classif, based on the stochastic bandit theory. This shortens the training stage, thereby allowing the exploration of a greater variety of tasks. By not wasting time on inefficient tasks, and focusing on the most promising ones, this algorithm results in a faster task selection and a more efficient use of the BCI training session. See [4] and https://twiki-sop.inria.fr/twiki/bin/view/Projets/Athena/CoAdapt/WebHome

8.1.3. ANR AMATIS

Participant: Pierre Chainais.

- *Title*: Multifractal Analysis and Applications to Signal and Image Processing
- *Type*: National Research Agency
- *Coordinator*: Univ. Paris-Est-Créteil (S. Jaffard)
- *Duration*: 2011-2015
- *Other Partners*: Univ. Paris-Est Créteil, Univ. Sciences et Technologies de Lille and Inria (Lille), ENST (Telechom ParisTech), Univ. Blaise Pascal (Clermont-Ferrand), and Univ. Bretagne Sud (Vannes), Statistical Signal Processing group at the Physics Department at the Ecole Normale Supérieure de Lyon, one researcher from the Math. Department of Institut National des Sciences Appliquees de Lyon and two researchers from the Laboratoire d'Analyse, Topologie et Probabilités (LAPT) of Aix-Marseille University.

- *Abstract*: Multifractal analysis refers to two concepts of different natures: On the theoretical side, it corresponds to pointwise singularity characterization and fractional dimension determination ; on the applied side, it is associated with scale invariance characterization, involving a family of parameters, the scaling function, used in classification or model selection. Following the seminal ideas of Parisi and Frisch in the mid-80s, these two components are usually related by a Legendre transform, stemming from a heuristic argument relying on large deviation and statistical thermodynamics principles: The multifractal formalism. This led to new theoretical approaches for the study of singularities of functions and measures, as well as efficient tools for classification and models selection, that allowed to settle longstanding issues (*e.g.*, concerning the modeling of fully developed turbulence). Though this formalism has been shown to hold for large classes of functions of widely different origins, the generality of its level of validity remains an open issue. Despite its popularity in applications, the interactions between theoretical developments and applications are unsatisfactory. Its use in image processing for instance is still in its infancy. This is partly due to discrepancy between the theoretical contributions mostly grounded in functional analysis and geometric measure theory, and applications naturally implying a stochastic or statistical framework. The AMATIS project aims at addressing these issues, by proposing a consistent and documented framework combining different theoretical approaches and bridging the gap towards applications. To that end, it will both address a number of challenging theoretical issues and devote significant efforts to elaborating a WEB platform with softwares and documentation. It will combine the efforts of mathematicians with those of physicists and experts in signal and image processing. Dissemination among and interactions between scientific fields are also intended via the organization of summer schools and workshop.
- *Activity Report*: a collaboration with P. Bas (CR CNRS, LAGIS) deals with the steganalysis of textured images. While steganography aims at hiding a message within some support, *e.g.* a numerical image, steganalysis aims at detecting the presence or not of any hidden message in the support. Steganalysis involves two main tasks: first identify relevant features which may be sensitive to the presence of a hidden message, then use supervised classification to build a detector. While the steganalysis of usual images has been well studied, the case of textured images, for which multifractal models may be relevant, is much more difficult. Indeed, textured images have a rich and disordered content which favors hiding information in an unperceptible manner. A student internship of 8 months at Master level in 2012 has led us to consider a very fundamental question. Steganalysis is usually proceeded to a classification based on histograms of features (bag of words). We consider the problem of the optimization of the bins of such histograms with respect to the performance of the classifier. We have shown that a balanced version of K-means which fills each cell equally yields an efficient quantization to this respect [28].

8.1.4. National Partners

- Laboratoire de Mathématiques d'Orsay, 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.
- LIF - CMI - Université de Provence.
  - Julien Audiffren *Collaborator*
    M. Valko, A. Lazaric, and M. Ghavamzadeh work with Julien on Semi-Supervised Apprenticeship Learning. We have recently developed 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*. We have published 2 conference papers [29] and [10].

- 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*.
- Biophotonics team at the Interdisciplinary Research Institute (IRI), Villeneuve d'Ascq, France.
  - Aymeric Leray *Collaborator*
    We have co-supervised an intern student (Pierre Pfennig, 2 months) on the topic of *quantitative guarantees of a super resolution method via concentration inequalities*. A paper is submitted to ICASSP 2014.
- LAGIS, Ecole Centrale Lille - Université de Lille 1, France.
  - Patrick Bas *Collaborator*
    We have a collaboration on the topic of *adaptive quantization to optimize classification from histrograms of features with an applicaiton to the steganalysis of textured images*.

8.2. European Initiatives

8.2.1. FP7 Projects

8.2.1.1. CompLACS

Type: COOPERATION
Defi: Composing Learning for Artificial Cognitive Systems
Instrument: Specific Targeted Research Project
Objectif: Cognitive Systems and Robotics
Duration: March 2011 - February 2015
Coordinator: University College London
Partner:

- Centre for Computational Statistics and Machine Learning, University College London (United Kingdom)
- Department of Computer Science, University of Bristol (United Kingdom)
- Department of Computer Science, Royal Holloway, University of London (United Kingdom)
- SNN Machine Learning, Radboud Universiteit Nijmegen (The Netherlands)
- Institut für Softwaretechnik und Theoretische Informatik, TU Berlin (Germany)
- University of Leoben (Austria)
- Computer Science Department, Technische Universitaet Darmstadt (Germany)

Inria contact: Rémi MUNOS
Website: COMPLACS
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.

8.2.2. Collaborations with Major European Organizations

Alexandra Carpentier: University of Cambridge (UK).

Michal Valko collaborates with Alexandra on extreme event detection (such as network intrusion) with limited allocation capabilities.

Prof. Marcello Restelli and Prof. Nicola Gatti: Politecnico di Milano (Italy).

A. Lazaric continued his collaboration on transfer in reinforcement learning which is leading to an extended version of the last year work on transfer of samples in MDPs. Furthermore, we are going to submit an extended version of an application of multi-arm bandit in a strategic environment such as sponsored search auctions.

8.3. International Initiatives

8.3.1. Inria Associate Teams

- *Inria principal investigator*: Mohammad Ghavamzadeh and Rémi Munos
  - *Institution*: McGill university (Canada)
  - *Laboratory*: Reasoning and Learning Lab
  - *Principal investigator*:
    * Prof. Joelle Pineau *Collaborator*
    * Prof. Doina Precup *Collaborator*
    * Amir massoud Farahmand *Collaborator*
- *Duration*: January 2013 - January 2015

8.3.2. Inria International Partners

8.3.2.1. Declared Inria International Partners

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.

8.3.2.2. Informal International Partners

- eHarmony Research, California.
  - Václav Petříček *Collaborator*
    Michal Valko has started to collaborate with eHarmony on sequential decision making for online dating and offline evaluation.
- University of Alberta, Edmonton, Alberta, Canada.
  - Csaba Szepesvári and Bernardo Avila Pires *Collaborator*
    We have been collaborating on the topic of *risk bounds in cost-sensitive multiclass classification* this year. We have an accepted paper [8] at ICML.
- 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 [21].

- School of Computer Science, Carnegie Mellon University, USA.
  - Prof. Emma Brunskill *Collaborator*
  - Mohammad Gheshlaghi Azar, PhD *Collaborator*
    A. Lazaric started a profitable collaboration on transfer in multi-arm bandit and reinforcement learning which led to two publications at ECML and NIPS. We are currently working on extensions of the previous algorithms and development of novel regret minimisation algorithms in non-iid settings.
- 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 [19]. 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 and two submissions to ICML.

8.4. International Research Visitors

8.4.1. Visits of International Scientists

8.4.1.1. Internships

- Daniele Calandriello, student at Politecnico di Milano, Italy

  Period: since April 2013.

  He is working with A. Lazaric on multi-task reinforcement learning.

8.4.2. Visits to International Teams

- Rémi Munos, since July 2013, Microsoft Research New-England, USA
- Mohammad Ghavamzadeh, since November 2013, Adobe Research, San Jose, CA
- Victor Gabillon visited Technicolor research lab, Palo Alto, from March to September 2013.
- Azadeh Khaleghi visited Walt Disney Animation Studios, Burbank, from March to September 2013.

9. Dissemination

9.1. Scientific Animation

9.1.1. Awards

- ***Crazy Stone*** 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.
- ***Alexandra Carpentier*** obtained an AFIA ex-aequo accessit for her PhD, (french machine learning/artificial intelligence second price).

9.1.2. Tutorials

- Tutorial by Rémi Munos at AAAI 2013: From Bandits to Monte Carlo Tree Search: The optimistic principle applied to Optimization and Planning.

9.1.3. Conferences, Workshops and Schools

- *Philippe Preux* and Marc Tommasi were the main organizers of the Conférence sur l'Apprentissage Automatique (CAP'13).

- *Rémi Munos* was the main organizer of the 8^{ith} Journées Francophones sur la Planification, la Décision et l'Apprentissage (JFPDA'13) along with *Marta Soare*, *Raphael Fonteneau*, *Michal Valko* and *Alessandro Lazaric*.
- *Rémi Munos* was co-chair of the Algorithmic Learning Conference, in Singapore, 2013.

9.1.4. Invited Talks

- Daniil Ryabko gave a talk entitled "Time-series information and unsupervised representation learning" at SMILE seminar in Paris
- Michal Valko gave an talk "Sequential Face Recognition with Minimal Feedback" which was opening talk of the series named 30 minutes of Science, a new format at Inria Lille to support intra-center collaboration.
- Rémi Munos gave a course (6 hours) at the Summer School Netadis in Hillerod, Denmark in September 2013.
- Rémi Munos was invited to give a talk at CMU in November 2013.
- Alessandro Lazaric was invited to give a talk at CMU in March 2013.
- Pierre Chainais gave a talk "Learning a common dictionary over a sensor network" at GDR Phénix - ISIS workshop about "Analysis and inference for networks" in Paris in november 2013.
- Pierre Chainais gave a tutorial talk on "Multifractal analysis of images and applicaitons" at the "Groupe Image of the company TOTAL in Paris La Défense on sept. 11th, 2013.
- Jérémie Mary gave a invited talk "Recommendation system from a bandit perspective" at GDR "Estimation et traitement statistique en grande dimension" on May 16th, 2013 - Télécom Paristech.
- Jérémie Mary gave an invited talk "Bandit point of view on recommenders" at Large-scale Online Learning and Decision Making Workshop Cumberland Lodge, Windsor, UK in September, 2013.
- Jeremie Mary gave an invited talk on recommender systems at "Journées rencontres AFIA/IHM" in may 2013.

9.1.5. Review Activities

- ***Participation to the program committee of international conferences***
  - International Conference on Pattern Recognition Applications and Methods (ICPRAM 2013)
  - Algorithmic Learning Theory (ALT 2013)
  - AAAI Conference on Artificial Intelligence (AAAI 2013)
  - European Workshop on Reinforcement Learning (EWRL 2013)
  - Annual Conference on Neural Information Processing Systems (NIPS 2013)
  - International Conference on Artificial Intelligence and Statistics (AISTATS 2013)
  - European Conference on Machine Learning (ECML 2013)
  - International Conference on Machine Learning (ICML 2013 and 2014)
  - International Conference on Uncertainty in Artificial Intelligence (UAI 2013)
  - French Conference on Planning, Decision-making, and Learning in Control Systems (JFPDA 2013)
  - IEEE FUSION 2013
  - IEEE Approximate Dynamic Programming and Reinforcement Learning (ADPRL 2013)
  - ICML workshop "Prediction with Sequential Models"
- ***International journal and conference reviewing activities*** (in addition to the conferences in which we belong to the PC)
  - 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)

9.1.6. Evaluation activities, expertise

- *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*, *R. Gaudel* and *J. Mary* are experts for *Crédit Impôt Recherche* (CIR).
- *E. Duflos* is a project proposal reviewer for ANR.
- *R. Munos* is a Member of the Belgium Commission Evaluation F.R.S-FNRS, 2013.

9.1.7. Other Scientific Activities

- *R. Munos* was Vice Président du Comité des Projets at Inria Lille-Nord Europe, until July 2013.
- *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.
- *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.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Ecole Centrale de Lille: *P. Chainais*, , “Machine Learning”, 36 hours, 3rd year.

Ecole Centrale de Lille: *P. Chainais*, “Wavelets and Applications”, 24 hours, 2nd year.

Ecole Centrale de Lille: *P. Chainais*, “Introduction to Matlab”, 16 hours, 3rd year.

Ecole Centrale de Lille: *P. Chainais*, “Signal processing”, 22 hours, 1st year.

Ecole Centrale de Lille: *P. Chainais*, “Data Compression”, 16 hours, 2nd year.

Ecole Centrale de Lille: *Ph. Preux*, “Data Data Data Data”, 2 hours, 3rd year.

P. Chainais is Responsible for a new 3rd year program called Decision making & Data analysis.

Master: *O. Pietquin*, “Decision under uncertainty”, 46 hours, M2, Master in Computer Science, Université de Lille 1.

Master: *A. Lazaric*, “Introduction to Reinforcement Learning”, 30h eq. TD, M2, Master “Mathematiques, Vision, Apprentissage”, ENS Cachan.

Master: *R. Gaudel*, “Data Mining”, 30h eq. TD, M2, Université Lille 3.

Master: *R. Gaudel*, “Web Mining”, 32h eq. TD, M2, Université Lille 3.

Master: *R. Gaudel*, “Algorithmic”, 19h eq. TD, M2, Université Lille 3.

Master: *Ph. Preux*, “Mathematics, Computer Science, and Modeling”, M1 of psychology, Université of Lille 3.

Master: *Ph. Preux*, “Algorithms, and programming in Python”, M1 MIASHS, Université of Lille 3.

Licence: *Ph. Preux*, “Algorithms, and programming in Python”, L3 MIASHS, Université of Lille 3.

Licence: *R. Gaudel*, “Programing”, 2×16h eq. TD, L1, Université Lille 3.

Licence: *R. Gaudel*, “Logic”, 31.5h eq. TD, L3, Université Lille 3.

Licence: *R. Gaudel*, “Information and Communication Technologies”, 2×16h eq. TD, L1, Université Lille 3.

Licence: *R. Gaudel*, “Artificial Intelligence”, 31.5h eq. TD, L2, Université Lille 3.

Licence: *R. Gaudel,*, “C2i”, 25h eq. TD, L1-3, Université Lille 3.

Licence: *R. Mary,*, “C2i”, 25h eq. TD, L1-3, Université Lille 3.

Master: *J. Mary*, “Programmation et analyse de donnée en R”, 24h eq TD, M1, Université de Lille 3, France.

Master: *J. Mary*, “Programmation web avancée”, 24h eq TD,M2, Université de Lille 3, France.

Master: *J. Mary*, “Programmation objet et Design Pattern”, 48h eq TD,M2, Université de Lille 3, France.

Master: *J. Mary*, “Algorithmique”, 12h eq TD,M1, Université de Lille 3, France.

Master (3rd year of Engineer School): *J. Mary*, “Machine Learning avec R" , 16 hours, M2, Option "Data Analysis and Decision", Ecole Centrale de Lille, France.

Master (3rd year of Engineer School): *E. Duflos*, “Advanced Estimation" , 20 hours, M2, Option "Data Analysis and Decision", Ecole Centrale de Lille, France.

Master (3rd year of Engineer School): *E. Duflos*, “Multi-Objects Filreting" , 16 hours, M2, Option "Data Analysis and Decision", Ecole Centrale de Lille, France.

9.2.2. Supervision

PhD: *Azadeh Khaleghi*, “Sur Quelques Problèmes non supervisés impliquant des séries temporelles hautement dépendantes”, Nov. 2013, Université de Lille 1, advisor: D. Ryabko.

PhD in progress: *Boris Baldassari*, *Apprentissage automatique et développement logiciel*, since May 2011, advisor: Ph. Preux.

PhD in progress: *Gabriel Dulac-Arnold*, *A General Sequential Model for Constrained Classification*, since Oct. 2011, advisor: Ph. Preux, L. Denoyer, P. Gallinari.

PhD in progress: *Victor Gabillon*, "Active Learning in Classification-based Policy Iteration", since Sep. 2009, advisor: Ph. Preux, M. Ghavamzadeh.

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: *Olivier Nicol*, "Data-driven evaluation of Contextual Bandit algorithms and applications to Dynamic Recommendation", since Nov. 2010, advisor: Ph. Preux, J. Mary.

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: *Emilie Kaufmann*, "Bayesian Bandits", since Oct. 2011, advisor: R. Munos, O. Cappé, A. Garivier.

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).

9.2.3. Juries

- member of the recruitement committee for an assistant professor position at Université de Lille 3: R. Gaudel, Ph. Preux
- member of the recruitement committee for an assistant professor position at Université de Lille 1: P. Chainais
- member of the recruitement committee for a professor position at Université de Paris 6: Ph. Preux
- Member of the jury DR2 Inria 2013: R. Munos
- Member of the jury CR2 Rocquencourt Inria 2013: R. Munos

9.3. Popularization

- "Small or big (data), make it sequentially!", J. Mary, Ph. Preux, invited talk at Euratechnologies, March 2013.
- Inria publishes an article about Face Recognition, Michal Valko, http://www.inria.fr/centre/lille/actualites/intel-collabore-avec-inria, March 2013
- Jérémie Mary highlighted on TV and on Inria website: you are how you browse: http://www.inria.fr/en/centre/lille/news/you-are-how-you-browse, Dec. 2013

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] A. KHALEGHI. , *Sur quelques problèmes non-supervisés impliquant des séries temporelles hautement dépendantes*, Institut national de recherche en informatique et en automatique (Inria), November 2013, http://hal.inria.fr/tel-00920184

Articles in International Peer-Reviewed Journals

[2] M. G. AZAR, R. MUNOS, H. KAPPEN. *Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model*, in "Machine Learning", 2013, vol. 91, n^{o} 3, pp. 325-349, http://hal.inria.fr/hal-00831875

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

[4] J. FRUITET, A. CARPENTIER, R. MUNOS, M. CLERC. *Automatic motor task selection via a bandit algorithm for a brain-controlled button*, in "Journal of Neural Engineering", January 2013, vol. 10, n^{o} 1 [*DOI :* 10.1088/1741-2560/10/1/016012], http://hal.inria.fr/hal-00798561

[5] M. HAUSKRECHT, I. BATAL, M. VALKO, S. VISWESWARAN, G. F. COOPER, G. CLERMONT. *Outlier detection for patient monitoring and alerting*, in "Journal of Biomedical Informatics", February 2013, vol. 46, pp. 47-55 [*DOI :* 10.1016/J.JBI.2012.08.004], http://hal.inria.fr/hal-00742097

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

[7] B. RYABKO, D. RYABKO. *A confidence-set approach to signal denoising*, in "Statistical Methodology", 2013, vol. 15, pp. 115–120, http://hal.inria.fr/hal-00913253

International Conferences with Proceedings

[8] B. AVILA PIRES, M. GHAVAMZADEH, C. SZEPESVARI. *Cost-sensitive Multiclass Classification Risk Bounds*, in "International Conference on Machine Learning", Atlanta, United States, 2013, http://hal.inria.fr/hal-00840485

[9] A. CARPENTIER, R. MUNOS. *Toward optimal stratification for stratified monte-carlo integration*, in "International Conference on Machine Learning", United States, 2013, http://hal.inria.fr/hal-00923685

[10] P. CHAINAIS, C. RICHARD. *Learning a common dictionary over a sensor network*, in "CAMSAP 2013", Saint-Martin, France, December 2013, pp. 1-4, http://hal.inria.fr/hal-00923742

[11] R. FONTENEAU, L. BUSONIU, R. MUNOS. *Optimistic planning for belief-augmented Markov decision processes*, in "IEEE International Symposium on Adaptive Dynamic Programming and reinforcement Learning, ADPRL 2013", Singapore, April 2013, http://hal.inria.fr/hal-00840202

[12] V. GABILLON, M. GHAVAMZADEH, B. SCHERRER. *Approximate Dynamic Programming Finally Performs Well in the Game of Tetris*, in "Neural Information Processing Systems (NIPS) 2013", South Lake Tahoe, United States, 2013, http://hal.inria.fr/hal-00921250

[13] M. GHESHLAGHI AZAR, A. LAZARIC, B. EMMA. *Regret Bounds for Reinforcement Learning with Policy Advice*, in "ECML/PKDD - European conference on machine learning and principles and practice of knowledge discovery in databases - 2013", Prague, Czech Republic, September 2013, http://hal.inria.fr/hal-00924021

[14] M. GHESHLAGHI AZAR, A. LAZARIC, B. EMMA. *Sequential Transfer in Multi-armed Bandit with Finite Set of Models*, in "NIPS - Advances in Neural Information Processing Systems 25 - 2013", Lake Tahoe, United States, December 2013, http://hal.inria.fr/hal-00924025

[15] H. KADRI, M. GHAVAMZADEH, P. PREUX. *A Generalized Kernel Approach to Structured Output Learning*, in "International Conference on Machine Learning (ICML)", Atlanta, United States, 2013, http://hal.inria.fr/hal-00695631

[16] G. KEDENBURG, R. FONTENEAU, R. MUNOS. *Aggregating optimistic planning trees for solving markov decision processes*, in "Advances in Neural Information Processing Systems", United States, 2013, pp. 2382-2390, http://hal.inria.fr/hal-00923681

[17] A. KHALEGHI, D. RYABKO. *Nonparametric multiple change point estimation in highly dependent time series*, in "Proc. 24th International Conf. on Algorithmic Learning Theory (ALT'13)", Singapore, Springer, 2013, pp. 382-396, http://hal.inria.fr/hal-00913250

[18] N. KORDA, E. KAUFMANN, R. MUNOS. *Thompson sampling for one-dimensional exponential family bandits*, in "Advances in Neural Information Processing Systems", United States, 2013, http://hal.inria.fr/hal-00923683

[19] B. KVETON, M. VALKO. *Learning from a Single Labeled Face and a Stream of Unlabeled Data*, in "10th IEEE International Conference on Automatic Face and Gesture Recognition", Shanghai, China, January 2013, http://hal.inria.fr/hal-00749197

[20] O.-A. MAILLARD, P. NGUYEN, R. ORTNER, D. RYABKO. *Optimal Regret Bounds for Selecting the State Representation in Reinforcement Learning*, in "ICML - 30th International Conference on Machine Learning", Atlanta, USA, United States, 2013, vol. 28(1), pp. 543-551, http://hal.inria.fr/hal-00778586

[21] 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, http://hal.inria.fr/hal-00823230

[22] D. RYABKO. *Time-series information and learning*, in "ISIT - International Symposium on Information Theory", Istanbul, Turkey, 2013, pp. 1392-1395, http://hal.inria.fr/hal-00823233

[23] D. RYABKO. *Unsupervised model-free representation learning*, in "Proc. 24th International Conf. on Algorithmic Learning Theory (ALT'13)", Singapore, Springer, 2013, pp. 354-366, http://hal.inria.fr/hal-00913244

[24] B. SZORENYI, R. BUSA-FEKETE, I. HEGEDÜS, R. ORMANDI, M. JELASITY, B. KÉGL. *Gossip-based distributed stochastic bandit algorithms*, in "30th International Conference on Machine Learning (ICML 2013)", Atlanta, United States, S. DASGUPTA, D. MCALLESTER (editors), 2013, vol. 28, pp. 19-27, http://hal.inria.fr/in2p3-00907406

[25] E. M. THOMAS, M. CLERC, A. CARPENTIER, E. DAUCÉ, D. DEVLAMINCK, R. MUNOS. *Optimizing P300-speller sequences by RIP-ping groups apart*, in "IEEE/EMBS 6th international conference on neural engineering (2013)", San Diego, United States, IEEE/EMBS, November 2013, http://hal.inria.fr/hal-00907781

[26] M. VALKO, A. CARPENTIER, R. MUNOS. *Stochastic Simultaneous Optimistic Optimization*, in "30th International Conference on Machine Learning", Atlanta, United States, February 2013, http://hal.inria.fr/hal-00789606

[27] M. VALKO, N. KORDA, R. MUNOS, I. FLAOUNAS, N. CRISTIANINI. *Finite-Time Analysis of Kernelised Contextual Bandits*, in "The 29th Conference on Uncertainty in Artificial Intelligence", Bellevue, United States, 2013, http://hal.inria.fr/hal-00826946

National Conferences with Proceedings

[28] P. BAS, P. CHAINAIS, E. ZIDEL - CAUFFET. *Quantification adaptative pour la stéganalyse d'images texturées*, in "GRETSI 2013", Brest, France, September 2013, http://hal.inria.fr/hal-00868550

[29] P. CHAINAIS, C. RICHARD. *Distributed dictionary learning over a sensor network*, in "CaP 2013", Villeneuve d'Ascq, France, July 2013, pp. 1-4, http://hal.inria.fr/hal-00923741

Scientific Books (or Scientific Book chapters)

[30] L. BUSONIU, R. MUNOS, R. BABUSKA. *A review of optimistic planning in Markov decision processes*, in "Reinforcement Learning and Adaptive Dynamic Programming for Feedback Control", F. LEWIS, D. LIU (editors), IEEE Press Series on Computational Intelligence, Wiley-IEEE Press, January 2013, chap. 22, pp. 494-516, http://hal.inria.fr/hal-00756742

Research Reports

[31] M. GHAVAMZADEH, Y. ENGEL. , *Bayesian Policy Gradient and Actor-Critic Algorithms*, January 2013, http://hal.inria.fr/hal-00776608

[32] P. L.A., M. GHAVAMZADEH. , *Actor-Critic Algorithms for Risk-Sensitive MDPs*, February 2013, http://hal.inria.fr/hal-00794721

[33] R. MUNOS. , *From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning*, 2013, http://hal.inria.fr/hal-00747575

References in notes

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

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

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

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

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

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

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

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

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

[43] 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

[44] 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

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

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

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