IN PARTNERSHIP WITH:
CNRS

Université Charles de Gaulle (Lille 3)

Université des sciences et technologies de Lille (Lille 1)

Ecole Centrale de Lille

Activity Report 2011

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, Learning and Statistical Methods

Table of contents

Project-Team SEQUEL

Keywords: Sensor Networks, Machine Learning, Inference, Sequential Learning

SEQUEL *is a joint project with the* LIFL *(UMR 8022 of CNRS, and University of Lille 1, and University of Lille 3) and the* LAGIS *(a joint lab of the École Centrale de Lille and the Lille 1 University).*

1. Members

Research Scientists

Rémi Munos [Co-head, Research Director (DR), INRIA, HdR]
Mohammad Ghavamzadeh [Researcher (CR) INRIA]
Alessandro Lazaric [Researcher (CR) INRIA]
Daniil Ryabko [Researcher (CR) INRIA]

Faculty Members

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

PhD Students

Boris Baldassari [CIFRE with Squoring Technology, since Sep., 2011]
Alexandra Carpentier [ANR-Région Nord-Pas de Calais Grant, since Oct., 2009]
Emmanuel Delande [DGA, since Nov., 2008]
Victor Gabillon [MENESR Grant, since Oct., 2009]
Jean-François Hren [MENESR Grant, since Oct., 2007]
Azadeh Khaleghi [CORDIS grant, since Oct., 2010]
Manuel Loth [INRIA-Région Nord-pas-de-calais Grant, Oct., 2006- Sept. 2009; ATER until July 2011]
Odalric-Ambrym Maillard [ENS Grant, until Oct., 2011]
Sami Naamane [CIFRE with France Telecom Grant, since Nov., 2011]
Olivier Nicol [MENESR Grant, since Oct., 2010]
Christophe Salperwyck [CIFRE with France Telecom Grant, since Dec., 2009]
Amir Sani [CORDIS grant, since Oct., 2011]

Post-Doctoral Fellows

Lucian Busoniu [ANR-Explora, since Apr., 2011]
Sertan Girgin [Région Nord-Pas de Calais until Feb. 28th, then contract with Addressing Business until Aug. 31st]
Hachem Kadri [Région Nord-Pas de Calais until Nov. 2011, then ANR Lampada]
Nathan Korda [COMPLacs, since Oct., 2011]
Michal Valko [COMPLacs, since Sept., 2011]

Administrative Assistant

Sandrine Catillon [Secretary (SAR) INRIA, shared by 3 projects]

Other

Jérôme Daquin [Master 2 internship, Univ. Lille 1, Apr. to Aug. 2011]

2. Overall Objectives

2.1. Introduction

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 [82]). 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

- Renowned for its work on the topic of exploration/exploitation trade-off, SequeL members have also successfully applied their research to the “Exploration and Exploitation Challenge” organized along at the International Conference on Machine Learning (ICML’11). SequeL’s PhD students Olivier Nicol [63] and Christophe Salperwyck [65] ranked first and second respectively.

3. Scientific Foundations

3.1. Introduction

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 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. Markov decision processes

Sequential decision processes occupy the heart of the SEQUEL project; a detailed presentation of this problem may be found in Puterman’s book [78].

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

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

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

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

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

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

where $\mathbb{E}$ is the expectation operator and $\gamma \in (0,1)$ is the discount factor. This value function V^π gives an evaluation of the performance of a given policy π. Other functionals of the sequence of future rewards may be considered, such as the undiscounted reward (see the stochastic shortest path problems [69]) 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) [67], 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 [68]. 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 ([86]):

- 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 challenge 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. Bandits

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 [79], 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.* [66] 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. Sequence prediction

Given a series of observations $x_1, \cdots, x_n$ it is required to predict the probability distribution of the next outcome x_{n+1}, before it 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.

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 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, fund 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. Clustering

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 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. Alternatively, one may assume some specific model on the data, leading to different formalizations of the problem.

3.4. Statistical learning

Before detailing some issues of statistical learning, let us remind the definition of a few terms.

Glossary

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. (source: http://www.aaai.org/AITopics/html/machine.htmlAAAI website)

Statistical learning is an approach to machine intelligence which 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.

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 [73]. 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. Kernel methods for non parametric 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. Non–parametric Bayesian models

Numerous problems in signal processing may be solved efficiently by way of 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 [72] 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 its potential in statistical signal processing still need to be developped.

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 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]

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

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

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. The Probability Hypothesis Density is a well-known method for single-sensor multi-target tracking problems in a Bayesian framework, but the extension to the multi-sensor case seems to remain a challenge.

4. Application Domains

4.1. Outline

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 analysis and processing,
- functional prediction,
- neuroscience.

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 [77]. 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 [83] 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 [77]
- 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 [81], [80] for such applications).

4.3. Signal analysis and 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.

Signals may be processed in several ways. One of the best 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. However, there exists alternatives like belief functions. 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 where once more they can be considered as a serious alternative to probabilities.

4.4. Functional prediction

One of the current trends in machine learning aims at dealing with data that are functions, rather than points or vectors. Generally speaking, functions represent a behavior (of a person, of an apparatus, or of an algorithm, or a response of a system, ...).

One application of functional prediction which is particularly emphasized these days, is the understanding of client behavior, either in material shops, or in virtual shops on the web. This understanding may then be used for different ends, such as the management of stocks according to sales, the proposition of products according to those already bought, the "instantaneous" management of some resource in the shop (advisors, cashiers, instant promotions, personalized advertisement, ...).

4.5. Neuroscience

Machine learning methods may be used for at least two means in neurosciences:

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 has strong ties with various concepts of psychology (Thorndike's law of effect, and the Rescorla-Wagner law to name the two most well-known).

5. Software

5.1. Introduction

In 2011, SEQUEL continued the development of software for computer games (notably Go) and also developed two novel libraries for functional regression and data mining.

5.2. Computer Games

Participant: Rémi Coulom.

We developed three main softwares for computer games:

- ***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 2011, its strength improved to **5 dan** on the KGS Go Server. It is distributed as a commercial product by *Unbalance Corporation* (Japan). 5-month work in 2011. URL: http://remi.coulom.free.fr/CrazyStone/
- ***Crazy Hanafuda*** is a new program to play the Japanese game of Hanafuda. 3 weeks of work in 2011. Discussion are in progress for licensing it.
- ***CLOP*** is a tool for automatic parameter optimization of game-playing programs. Distributed as freeware (GPL). One month of work in 2011. Available at: http://remi.coulom.free.fr/CLOP/

5.3. Functional Regression

Participant: Hachem Kadri.

A software package in C++ of algorithms for nonlinear functional data analysis using our operator-valued kernel framework (see sec. 6.4.1) is under development. A beta-version of the software can be downloaded at: https://gforge.inria.fr/frs/?group_id=982.

The aim of this library is to grow and be shared in our scientific community, and also to be a software resource for our group.

5.4. Data mining library

Participant: Sertan Girgin.

A fully stand-alone library for data mining has been developed, including many classical algorithms for supervised and non supervised learning. This library is available as an internal resource for the group.

6. New Results

6.1. Introduction

The new results are organized in the following sections:

1. decision under uncertainty,
2. foundations of machine learning,
3. supervised learning,
4. signal processing (sensor networks),
5. other results.

6.2. Decision Under Uncertainty

Participants: Lucian Busoniu, Alexandra Carpentier, Rémi Coulom, Victor Gabillon, Mohammad Ghavamzadeh, Sertan Girgin, Jean-François Hren, Alessandro Lazaric, Manuel Loth, Odalric-Ambrym Maillard, Rémi Munos, Olivier Nicol, Philippe Preux, Daniil Ryabko.

6.2.1. Reinforcement learning and approximate dynamic programming

In the domain of reinforcement learning and approximate dynamic programming, we identify two main lines of research.

6.2.1.1. Links between Approximate Dynamic Programming and Statistical Learning Theory

The main objective here is to use tools from *statistical learning theory* to derive finite-sample performance bounds for RL and ADP algorithms. The goal is to derive bounds on the performance of the policies induced by these algorithms in terms of the number of simulation data and the capacity and approximation power of the considered function and policy spaces. The results of this study allow us to have a better understanding of the functionality of these algorithms and help us to design them more efficiently. The main contributions to this research line in 2011 are:

- **Classification-based Policy Iteration with a Critic [25], [51].** In collaboration with Bruno Scherrer (INRIA Nancy - Grand Est, Team MAIA) we extended last year work on classification-based policy iteration by adding a value function approximation component (critic) to rollout classification-based policy iteration (RCPI) algorithms. The idea is to use a critic to approximate the return after we truncate the rollout trajectories. This allows us to control the bias and variance of the rollout estimates of the action-value function. Therefore, the introduction of a critic can improve the accuracy of the rollout estimates, and as a result, enhance the performance of the RCPI algorithm. We presented a new RCPI algorithm, called *direct policy iteration with critic* (DPI-Critic), and provided its finite-sample analysis when the critic is based on the LSTD method. We also empirically evaluated the performance of DPI-Critic and compared it with DPI and LSPI in two benchmark reinforcement learning problems.

- **Finite-Sample Analysis of Least-Squares Policy Iteration [10], [45].** We extended last year work on the finite-sample analysis of least-squares temporal-difference (LSTD) to the least-squares policy iteration (LSPI) algorithm. In particular, we analyzed how the error at each policy evaluation step is propagated through the iterations of a policy iteration method, and derive a performance bound for the LSPI algorithm.
- **Speedy Q-Learning [16], [48].** We introduce a new convergent variant of Q-learning, called speedy Q-learning, to address the problem of slow convergence in the standard form of the Q-learning algorithm. We prove a PAC bound on the performance of SQL, which shows that for an MDP with n state-action pairs and the discount factor γ only $T = O(log(n)/(\epsilon^2(1-\gamma)^4))$ steps are required for the SQL algorithm to converge to an ϵ-optimal action-value function with high probability. This bound has a better dependency on $1/\epsilon$ and $1/(1-\gamma)$, and thus, is tighter than the best available result for Q-learning. Our bound is also superior to the existing results for both model-free and model-based instances of batch Q-value iteration that are considered to be more efficient than the incremental methods like Q-learning.
- **Selecting the State-Representation in Reinforcement Learning [34].** The problem of selecting the right state-representation in a reinforcement learning problem is considered. Several models (functions mapping past observations to a finite set) of the observations are given, and it is known that for at least one of these models the resulting state dynamics are indeed Markovian. Without knowing neither which of the models is the correct one, nor what are the probabilistic characteristics of the resulting MDP, it is required to obtain as much reward as the optimal policy for the correct model (or for the best of the correct models, if there are several). We propose an algorithm that achieves that, with a regret of order $T^{2/3}$ where T is the horizon time.
- **Transfer from Multiple MDPs [32].** Transfer reinforcement learning (RL) methods leverage on the experience collected on a set of source tasks to speed-up RL algorithms. A simple and effective approach is to transfer samples from source tasks and include them in the training set used to solve a target task. In this paper, we investigate the theoretical properties of this transfer method and we introduce novel algorithms adapting the transfer process on the basis of the similarity between source and target tasks. Finally, we report illustrative experimental results in a continuous chain problem.

6.2.1.2. RL in High-dimensional Spaces

The main objective here is to devise, analyze, implement, and experiment with RL algorithms whose sample and computational complexities do not grow rapidly with the dimension of the state space. We have tackled this problem from two different angles:

- **Exploiting the Regularities of the Problem [57], [8], [27].** In order to solve RL in high dimensions, we should exploit all the regularities of the problem in hand. *Smoothness* is the most common regularity. We continued our collaboration with Amir massoud Farahmand and Csaba Szepesvári at the university of Alberta, Canada, and Shie Mannor at Technion, Israel, on using regularization methods for automatic model selection for value function approximation in RL. We have devised and analyzed the first ℓ_2-regularized RL algorithms by adding ℓ_2-regularization to three well-known ADP algorithms: fitted Q-iteration, modified Bellman residual minimization, and least-squares temporal-difference learning [57], [8]. The designed algorithms work in both linear and reproducing kernel Hilbert spaces. *Sparsity* is another form of regularity that clearly plays a central role in the emerging theory of learning in high dimensions. We have worked on using ℓ_1-regularization in approximate dynamic programming and RL, which may also serve as a method for feature selection in value function approximation. We have derived finite-sample performance bounds for an algorithm resulting from adding ℓ_1-penalty to the widely-used *least-squares temporal-difference learning* (LSTD) algorithm [27].
- **Random Projections [28], [52].** We have looked into recent directions popularized in compressive sensing concerning the preservation of properties, such as norm or inner-product, of high dimensional objects when projected on possibly much lower dimensional random subspaces. We

have studied the popular LSTD algorithm when a space of low dimension is generated with a random projection from the high-dimensional space, and derived performance bounds for the resulting algorithm [28], [52].

6.2.2. Planning and exploration vs. exploitation trade-off

In the domain of planning and exploration-exploitation algorithms, we identify two main lines of research.

6.2.2.1. Multi-arm Bandit, Online Learning and Optimization

- **Active Learning in Multi-Armed Bandit Problems [18], [49], [24], [50].** This can be seen as an online allocation problem with several options and is closely related to the problem of *optimal experimental design* in statistics. The objective here is to allocate a fixed budget to a finite (or possibly infinite) number of options (arms) in order to achieve the best accuracy in estimating the quality of each option. In addition to having application in a number of different fields such as *online advertisement* and *personalizing treatment*, this problem is of specific importance in RL in which generating training data is usually expensive. In this framework, we have studied the following two problems: **1)** estimating the mean values of all the arms uniformly well in a multi-armed bandit setting [18], [49], and **2)** identifying the best arm in each of the bandits in a multi-bandit multi-armed setting [24], [50]. For each problem, we have developed algorithms with theoretical guarantees.
- **Finite Time Analysis of Stratified Sampling for Monte Carlo [20].** We consider the problem of stratified sampling for Monte-Carlo integration. We model this problem in a multi-armed bandit setting, where the arms represent the strata (an interval in the input domain), and the goal is to estimate a weighted average of the mean values of the arms. We propose a strategy that samples the arms according to an upper bound on their standard deviations and compare its estimation quality to an ideal allocation that would know the standard deviations of the strata. We provide two regret analyses: a distribution-dependent bound $O(n^{-3/2})$ that depends on a measure of the disparity of the strata, and a distribution-free bound $O(n^{-4/3})$ that does not.
- **Optimistic Optimization of a Deterministic Function without the Knowledge of its Smoothness [36].** We consider a global optimization problem of a deterministic function f in a semi-metric space, given a finite budget of n evaluations. The function f is assumed to be locally smooth (around one of its global maxima) with respect to a semi-metric. We describe two algorithms based on optimistic exploration that use a hierarchical partitioning of the space at all scales. A first contribution is an algorithm, DOO, that requires the knowledge of . We report a finite-sample performance bound in terms of a measure of the quantity of near-optimal states. We then define a second algorithm, SOO, which does not require the knowledge of the semi-metric under which f is smooth, and whose performance is almost as good as DOO optimally-fitted.
- **Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences [35].** We consider a Kullback-Leibler-based algorithm for the stochastic multi-armed bandit problem in the case of distributions with finite supports (not necessarily known beforehand), whose asymptotic regret matches the lower bound of Burnetas and Katehakis (1996). Our contribution is to provide a finite-time analysis of this algorithm; we get bounds whose main terms are smaller than the ones of previously known algorithms with finite-time analyses (like UCB-type algorithms).
- **Adaptive bandits: Towards the best history-dependent strategy [33].** We consider multi-armed bandit games with possibly adaptive opponents. We introduce models Θ of constraints based on equivalence classes on the common history (information shared by the player and the opponent) which define two learning scenarios: (1) The opponent is constrained, i.e. he provides rewards that are stochastic functions of equivalence classes defined by some model. The regret is measured with respect to (w.r.t.) the best history-dependent strategy. (2) The opponent is arbitrary and we measure the regret w.r.t. the best strategy among all mappings from classes to actions (i.e. the best history-class-based strategy) for the best model. This allows to model opponents (case 1) or strategies (case 2) which handles finite memory, periodicity, standard stochastic bandits and other situations. When only one model is considered, we derive tractable algorithms achieving a tight regret (at time T)

bounded by $O(\sqrt{TAC})$, where C is the number of classes. Now, when many models are available, all known algorithms achieving a nice regret $O(\sqrt{T})$ are unfortunately not tractable and scale poorly with the number of models. Our contribution here is to provide tractable algorithms with regret bounded by $T^{2/3}C^{1/3}\log(|\Theta|)^{1/2}$.

- **Pure Exploration in Finitely-Armed and Continuous-Armed Bandits [5].** We consider the framework of stochastic multi-armed bandit problems and study the possibilities and limitations of forecasters that perform an on-line exploration of the arms. These forecasters are assessed in terms of their simple regret, a regret notion that captures the fact that exploration is only constrained by the number of available rounds (not necessarily known in advance), in contrast to the case when the cumulative regret is considered and when exploitation needs to be performed at the same time. We believe that this performance criterion is suited to situations when the cost of pulling an arm is expressed in terms of resources rather than rewards. We discuss the links between the simple and the cumulative regret. One of the main results in the case of a finite number of arms is a general lower bound on the simple regret of a forecaster in terms of its cumulative regret: the smaller the latter, the larger the former. Keeping this result in mind, we then exhibit upper bounds on the simple regret of some forecasters. The paper ends with a study devoted to continuous-armed bandit problems; we show that the simple regret can be minimized with respect to a family of probability distributions if and only if the cumulative regret can be minimized for it. Based on this equivalence, we are able to prove that the separable metric spaces are exactly the metric spaces on which these regrets can be minimized with respect to the family of all probability distributions with continuous mean-payoff functions.
- **X-Armed Bandits [6].** We consider a generalization of stochastic bandits where the set of arms, X, is allowed to be a generic measurable space and the mean-payoff function is locally Lipschitz with respect to a dissimilarity function that is known to the decision maker. Under this condition we construct an arm selection policy, called HOO (hierarchical optimistic optimization), with improved regret bounds compared to previous results for a large class of problems. In particular, our results imply that if X is the unit hypercube in a Euclidean space and the mean-payoff function has a finite number of global maxima around which the behavior of the function is locally continuous with a known smoothness degree, then the expected regret of HOO is bounded up to a logarithmic factor by $\sqrt{n}$, that is, the rate of growth of the regret is independent of the dimension of the space. We also prove the minimax optimality of our algorithm when the dissimilarity is a metric. Our basic strategy has quadratic computational complexity as a function of the number of time steps and does not rely on the doubling trick. We also introduce a modified strategy, which relies on the doubling trick but runs in linearithmic time. Both results are improvements with respect to previous approaches.
- **Learning with Stochastic Inputs and Adversarial Outputs [11].** Most of the research in online learning is focused either on the problem of adversarial classification (i.e., both inputs and labels are arbitrarily chosen by an adversary) or on the traditional supervised learning problem in which samples are independent and identically distributed according to a stationary probability distribution. Nonetheless, in a number of domains the relationship between inputs and outputs may be adversarial, whereas input instances are i.i.d. from a stationary distribution (e.g., user preferences). This scenario can be formalized as a learning problem with stochastic inputs and adversarial outputs. In this paper, we introduce this novel stochastic-adversarial learning setting and we analyze its learnability. In particular, we show that in binary classification, given a hypothesis space H with finite VC-dimension, it is possible to design an algorithm which incrementally builds a suitable finite set of hypotheses from H used as input for an exponentially weighted forecaster and achieves a cumulative regret of order $\sqrt{nVC(H)logn}$ with overwhelming probability. This result shows that whenever inputs are i.i.d., it is possible to solve any binary classification problem using a finite VC-dimension hypothesis space with a sub-linear regret independently from the way labels are generated (either stochastic or adversarial). We also discuss extensions to multi-label classification, regression, learning from experts and bandit settings with stochastic side information, and application to games.

- **ICML Exploration-Exploitation Challenge [65], [63].** Olivier Nicol and Jérémie Mary won the ICML challenge on Exploration and Exploitation 2 organized by Cambridge on dataset provided by Adobe. The winning approach is based on ideas close to bayesian networks and Thomson sampling as Ad Predictor from Microsoft. These kind of succes emphases the need for better theoretical analysis of theses frameworks. The challenge was also a good occasion to think about the best way to evaluate online politics (this part also attracts interest from Orange Labs). A publication to JLMR is submitted.

6.2.2.2. Planning

- **Optimistic Planning for Sparsely Stochastic Systems [17].** We propose an online planning algorithm for finite action, sparsely stochastic Markov decision processes, in which the random state transitions can only end up in a small number of possible next states. The algorithm builds a planning tree by iteratively expanding states, where each expansion exploits sparsity to add all possible successor states. Each state to expand is actively chosen to improve the knowledge about action quality, and this allows the algorithm to return a good action after a strictly limited number of expansions. More specifically, the active selection method is optimistic in that it chooses the most promising states first, so the novel algorithm is called optimistic planning for sparsely stochastic systems. We note that the new algorithm can also be seen as model-predictive (receding-horizon) control. The algorithm obtains promising numerical results, including the successful online control of a simulated HIV infection with stochastic drug effectiveness.
- **Optimistic Planning in Markov decision processes [46].** 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.

6.2.3. Applications

6.2.3.1. Management of ad campaigns on the web

More work has been dedicated to the topic aiming at optimizing ad campaigns on the web under real-time constraints, in a dynamic environment [9].

6.3. Foundations of Machine Learning

Participants: Daniil Ryabko, Azadeh Khaleghi, Romaric Gaudel.

6.3.1. Sequence prediction in the most general form

The problem of sequence prediction consists in forecasting, on each step of time n, the probabilities of the next outcome of the observed sequence of data $x_1, x_2, \cdots, x_n, \cdots$. In the most general formulation of the problem, we assume that we are given a set $\mathcal{C}$ of probability measures (on the space of infinite sequences). We can then assume that the sequence is generated by an unknown measure μ that belongs to $\mathcal{C}$, or that the measure μ is arbitrary, but we compare the performance of our predictor to that of the best predictor in $\mathcal{C}$.

6.3.1.1. Relation between the realizable and non-realizable cases of the sequence prediction problem

The realizable case of the sequence prediction problem is when the measure μ belongs to an arbitrary but known class $\mathcal{C}$ of process measures. The non-realizable case is when μ is completely arbitrary, but the prediction performance is measured with respect to a given set $\mathcal{C}$ of process measures. We are interested in the relations between these problems and between their solutions, as well as in characterizing the cases when a solution exists, and finding these solutions. In this work [13] we show that if the quality of prediction is measured by total variation distance, then these problems coincide, while if it is measured by expected average KL-divergence, then they are different. For some of the formalizations we also show that when a solution exists, it can be obtained as a Bayes mixture over a countable subset of $\mathcal{C}$. As an illustration to the general results obtained, we show that a solution to the non-realizable case of the sequence prediction problem exists for the set of all finite-memory processes, but does not exist for the set of all stationary processes.

6.3.2. Statistical inference

We continue to obtain new results using the theoretical framework developed recently for studying time series generated by stationary ergodic time series. This year, new results obtained include a topological characterizing of composite hypotheses for which consistent tests exist, as well as new results on clustering.

6.3.2.1. A criterion for the existence of consistent tests

The most general result that we have obtained [14] on hypothesis testing provides a complete characterization (necessary and sufficient conditions) for the existence of a consistent test for membership to an arbitrary family H_0 of stationary ergodic discrete-valued processes, against H_1 which is the complement of H_0 to this class of processes. The criterion is that H_0 has to be closed in the topology of distributional distance, and closed under taking ergodic decompositions of its elements.

6.3.3. Clustering

6.3.3.1. Online clustering of time series

An asymptotically consistent algorithm has been proposed for the problem of online clustering of time series. There is a growing body of time series samples, each of which grows with time. On each time step, it is required to group these time series into k clusters. It is known that each of the time series is generated by one out of k *unknown* stationary ergodic distributions. An algorithm is proposed that, for each fixed portion of samples, eventually (with probability 1) puts into the same group those and only those samples that were generated by the same distribution. Empirical performance of the algorithm is evaluated on synthetic and real data.

6.3.3.2. Clustering of ranked data

We introduced [47] a novel approach to clustering rank data on a set of possibly large cardinality $n \in \mathbb{N}^*$, relying upon Fourier representation of functions defined on the symmetric group $\mathfrak{S}_n$. In the proposed setup, covering a wide variety of practical situations, rank data are viewed as distributions on $\mathfrak{S}_n$. Cluster analysis aims at segmenting data into homogeneous subgroups, hopefully very dissimilar in a certain sense. Whereas considering dissimilarity measures/distances between distributions on the non commutative group $\mathfrak{S}_n$, in a coordinate manner by viewing it as embedded in the set $[0,1]^{n!}$ for instance, hardly yields interpretable results and leads to face obvious computational issues, evaluating the closeness of groups of permutations in the Fourier domain may be much easier in contrast. Indeed, in a wide variety of situations, a few well-chosen Fourier (matrix) coefficients may permit to approximate efficiently two distributions on $\mathfrak{S}_n$ as well as their degree of dissimilarity, while describing global properties in an interpretable fashion. Following in the footsteps of recent advances in automatic feature selection in the context of unsupervised learning, we propose to cast the task of clustering rankings in terms of optimization of a criterion that can be expressed in the Fourier domain in a simple manner.

6.4. Supervised learning

Participants: Alexandra Carpentier, Emmanuel Duflos, Hachem Kadri, Manuel Loth, Odalric-Ambrym Maillard, Rémi Munos, Philippe Preux, Christophe Salperwyck.

6.4.1. Regression and classification

- **Sparse Recovery with Brownian Sensing [19].**

  We consider the problem of recovering the parameter α of a sparse function f (i.e. the number of non-zero entries of α is small compared to the number K of features) given noisy evaluations of f at a set of well-chosen sampling points. We introduce an additional randomization process, called Brownian sensing, based on the computation of stochastic integrals, which produces a Gaussian sensing matrix, for which good recovery properties are proven, independently on the number of sampling points N, even when the features are arbitrarily non-orthogonal. Under the assumption that f is Hölder continuous with exponent at least $1/2$ we provide an estimate of the parameter with quadratic error $O(||\eta||/N)$, where η is the observation noise. The method uses a set of sampling points uniformly distributed along a one-dimensional curve selected according to the features. We report numerical experiments illustrating our method.

- **Operator-valued Kernels for Nonlinear FDA [31], [38], [30], [53]** Following the extension of RKHS to functional setting [74], we further developed this work in [38] for functional supervised classification.

  We introduced a set of rigorously defined operator-valued kernels that can be valuably applied to nonparametric operator learning when input and output data are continuous smooth functions, and we have showed their use for solving the problem of minimizing a L^2-regularized functional in the case of functional outputs without the need to discretize covariate and target functions [53].

  The framework developed can also be applied when the input data are both discrete and continuous [30].

  Our fully functional approach has been successfully applied to the problems of speech inversion [31] and sound recognition [38], showing that the proposed framework is particularly relevant for audio signal processing applications where attributes are really functions and dependent of each other.

  This work is done in collaboration with Francis Bach (INRIA, Sierra), Alain Rakotomamonjy and Stéphane Canu (LITIS, Rouen).

- **Datum-wise representation [44], [54].** We consider supervised classification. We introduce the concept of datum-wise representation for supervised classification [44]. While traditional approaches yield a “best” representation at the data space level, that is, the same representation is used for all the data, we proposed the idea, as well as an algorithm, that yields the “best” representation for each data. Among other appealing properties, this leads to sparse representation of each data, and an averaged sparser representation of each data in the data space. Along a classifier, the learning algorithm produces a “representer”, that is a function that yields a representation given a data.

  We further improved this approach to encompass various settings which are traditionally kept as different (cost-sensitive classification and different structured sparsity) [54].

- **Iso-regularization descent [1].** Manuel Loth has defended his PhD dissertation [1] where he has provided a detailed presentation and analysis of his algorithm to solve the LASSO. This algorithm is very efficient. It is an active set algorithm that solves the LASSO by considering it a convex problem with linear constraints.

- **Learning with few examples [41], [64].** Christophe Salperwyck has studied the performance of various classifiers when few examples are available. This is an important point in incremental learning, and few studies have been devoted to this particular setting. Performance we are accustomed to when the examples are quite numerous are severely disturbed in this setting. For more details, please see [41], [64].

- **Incremental discretization [40].** In incremental learning, discretization should be adaptive in order to cope with the values of the attributes that are observed. This issue is currently under study by Christophe Salperwyck [40].

6.5. Sensors Networks: Tracking, Localization and Communication

Participants: Emmanuel Delande, Emmanuel Duflos, Pierre Chainais, Philippe Vanheeghe.

6.5.1. The sensor management problem

The aim of this work is to manage a set of sensors to track vehicles or groups of people in land applications. Our work focuses on sensor management in the frame of the random finite sets where the Probability Hypothesis Density (PHD) is a well-known method for single-sensor multi-target tracking problems in a Bayesian framework, but the extension to the multi-sensor case seems to remain a challenge. We have proposed an extension of Mahler's work to the multi-sensor case by providing an expression of the true PHD multi-sensor data update equation. Then, based on the configuration of the sensors fields of view (FOVs), a joint partitioning of both the sensors and the state space provides an equivalent yet more practical expression of the data update equation, allowing a more effective implementation in specific FOV configurations ([70]). This work is done in collaboration with Thales Communications. The multi-sensor / multi-target filtering problem by using PHD filtering methods are topics developed in the The PhD thesis of Emmanuel Delande. This PhD thesis entitled "Multi-sensor PHD filtering with application to sensor management" will be defended in December 2011. In addition to the different questions described above, see also [22] and [23]. Then, a new approach using operational objectives, related to the type of application, for sensor manager is proposed.

6.5.2. Statistical signal processing: application to civil engineering

We have obtained a PICS (International Project for Scientific Cooperation) from the CNRS in 2008 for 3 years to work in cooperation with the Department of Civil and Environmental Engineering of the University of Waterloo (Canada). During this cooperation we have developed a belief functions based method to track the building materials on a construction site. ([71]). Based on this cooperation, during 2011 a new common research project with the same department of the University of Waterloo has been built, and is actually submitted for funding. The topic of this project is the using of nonparametric Bayesian models in the area of Non-destructive Testing.

6.5.3. Accurate Localization using Satellites in Urban Canyons

Today, Global Navigation Satellite Systems (GNSS) have penetrated the transport field through applications such as monitoring of containers. These applications do not necessarily request a high availability, integrity and accuracy of the positioning system. For safety applications (as complete guidance of autonomous vehicles), performances require to be more stringent. For, sensors may deliver very erroneous measurements because of such hard external conditions which reduce significantly the possibilities to receive direct signals. The consequences of environmental obstructions are unavailability of the service and reception of reflected signals that degrades in particular the accuracy of the positioning. Indeed, NLOS (Non Line Of Sight) signals, i.e. signals received after reflections on the surrounding obstacles, frequently occur in dense environments and degrade localization accuracy because of the delays observed on the propagation time measurement creating additional error on pseudorange estimation. In the previous years we have proposed new algorithms to improve the localization precision. This algorithm are based on two principles : a jump multimodel approach and a joint state - noise density estimation. We have focused this year on an approach using Dirichlet Process Mixture to track the noise density in urban canyon while estimating the position of the vehicle. Algorithm have been validated on real data collected in a French town : Belfort. Nicolas Viandier has defended his PhD on this subject on June 2011 . ([76], [75], [84], [85] [4]). These results will be presented to the Workshop Non Parametric Bayes at the NIPS Conference en Decembre 2011 ([62]) and to the ICASSP 2011 Conference ([37]).

6.5.4. Internet of Things : Mitigation of Impulsive Noise Effects

The term "Internet of Things" has come to describe a number of technologies and research disciplines that enable the Internet to reach out into the real world of physical objects. Technologies like RFID, short-range wireless communications, real-time localization and sensor networks are now becoming increasingly common, bringing the Internet of Things into commercial use. In such applications the data sent by a *thing* to another may generate an impulse noise in the reception channel of objects in the neighbourhood. The noise appearing in such applications can be considered as α-stable. In this context, we've tackled the problem of interference mitigation in ad hoc networks. In such context, the multiple access interference (MAI) is known to be of an impulsive nature. Therefore, the conventional Gaussian assumption can not be considered to model this type of interference. Contrariwise, it can be accurately modeled by stable distributions. Here, this issue is addressed within an Orthogonal Frequency Division Multiplexing (OFDM) transmission link assuming a symmetric α-stable model for the signal distortion due to MAI. We have proposed a method for the joint estimation of the transmitted multicarrier signal and the noise parameters.Based on sequential Monte Carlo (SMC) methods, the proposed scheme allows the online estimation using a Raoblackwellized particle filter. These results have been presented to the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2011) [29]. We are now focusing on bayesian non linear filtering with non stationnary alpha stable measurement noise. We have shown that a Dirichlet Process Mixture can improve the estimation by modelling the noise by both a infinite Cauchy mixture or a infinite alpha stable mixture. These first results will be presented to the Workshop Non Parametric Bayes at the NIPS Conference en Decembre 2011 [62].

6.5.5. *Image processing and statistical image modeling*

Pierre Chainais arrived in SequeL in september 2010 with the purpose of a thematic evolution toward non parametric Bayesian approaches. This represents an important investment in very new directions on an emerging topic at the interface between machine learning and signal/image processing. Discussions have begun with Emmanuel Duflos and Philippe Vanheeghe on the use of non parametric Bayesian approaches to blind deconvolution of noisy natural images. The main objective is to use together the typical structure and sparsity of space-scale representations of images.

Pierre Chainais has continued working on several older projects. One of them deals with the segmentation of nanotubes in microscopic imaging [12], [43]. B. Lebental at IFFSTAR works on the conception of new nano-sensors based on the use of carbon nanotubes to build a nano-membrane. P. Chainais has developed an image processing pipeline to analyse images of these nanomembranes so as to characterize their properties in a precise and objective manner. Among other properties, the histograms of orientations of the nanotubes is provided. This tool will be very useful since such nanosensors are becoming more and more common.

In solar astronomy [7], [21], we have proposed a tool for the virtual super-resolution of scale invariant textured images. The aim of this project was to provide astronomers with plausible high-resolution images to calibrate next generation spatial telescopes. In particular, our images can be used to optimize the compression algorithm to be embedded in a spatial telescope. In collaboration with M. Chevaldonné and J-M. Favreau (Université Clermont-Ferrand I), we work a software for texture synthesis on 3D surfaces [42] based on multifractal processes. A first version of the software is under current development. More marginal is our work on the use of stochastic processes for the simulation of turbulent pressure fields in collaboration with M. Pachebat (Laboratoire de Mécanique et d'Acoustique de Marseille) and Nicolas Totaro (LVA, INSA Lyon).

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

7.1.1. *Addressing Business*

Participants: Sertan Girgin, Philippe Preux.

Addressing Business develops a software to help their clients (companies) find new clients. Currently, this software is an information system, that helps the human decision maker.

The goal of this contract was to realize a first exploratory step towards using data mining techniques to handle, and possibly improve, this process. Confidentiality issues restrict the communication on this contract [60]. However, this study has been very successful, and we look forward further collaboration with this company on this topic.

7.1.2. Orange Labs

Participants: Jérémie Mary, Olivier Nicol, Philippe Preux, Christophe Salperwyck.

There has been various activities between SEQUEL and Orange Labs.

First, the collaboration around the PhD of Christophe Salperwyck has continued. Second, a CRE has been signed in 2011 to continue our work on web advertising, and more generally, collaborative filtering. On this topic, Sami Naamane has been hired in Fall 2011 as PhD student.

7.1.3. Effigenie

Participant: Jérémie Mary.

We worked on the next steps of the optimisation of thermal control of building with Effigenie. We presented a common project June and this start up won in summer the OSEO innovation prize (and the LMI one).

7.1.4. Squoring Technology

Participants: Boris Baldassari, Philippe Preux.

Boris Baldassari has been hired by Squoring Technology (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.

7.1.5. Through the pôle de compétitivité "Industries du commerce"

Participants: Sertan Girgin, Philippe Preux.

2011 is the last year of the "Ubiquitous Virtual Seller" project of the "Pôle de compétitivité Industries du Commerce" (PICOM). We have completed our contribution related to recommendation systems [59]. This work was done mostly in collaboration with Becquet and Oxylane.

7.1.6. Qualisteo

Participants: Pierre Chainais, Emmanuel Duflos.

In collaboration with Emmanuel Duflos, Pierre Chainais has been involved in the supervising of a Master 2 student within a collaboration with the company Qualisteo on the pattern recognition in electric power consumption signals. A PhD grant is under study. The purpose is to learn how to identify the origin of the electric consumption of a house from the power consumption alone. This problem combines signal processing as well as machine learning questions. This project is still under discussion.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. PICOM

see 7.1.5.

8.2. National Initiatives

8.2.1. DGA/Thales

The work on sensor management went on this year, focusing on the extension to the multisensor case of the PHD filter. This work is realized in the frame of the thesis of Emmanuel Delande (Grant DGA/CNRS) in collaboration with Thales Communication. The defense of this PhD thesis will be held in December 2011.

8.2.2. ANR-Lampada

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

- Title: Learning Algorithms, Models an 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 ; Saint Etienne, Laboratoire d'Informatique de Paris 6.
- See also: http://lampada.gforge.inria.fr/
- Activity Report: Philippe Preux has continued his collaboration with Ludovic Denoyer (assistant professor, Université de Paris 6), Gabriel Arnold-Dulac (PhD student), and Patrick Gallinari (professor, Université de Paris 6). This led to the work on datum-wise representation [44], [54].

8.2.3. ANR EXPLO-RA

Participants: Lucian Busoniu, Alexandra Carpentier, Mohammad Ghavamzadeh, Jean-François Hren, Alessandro Lazaric, Odalric-Ambrym Maillard, Rémi Munos, Daniil Ryabko.

- Title: EXPLOration - EXPLOitation for efficient Resource Allocation with Applications to optimization, control, learning, and games
- Type: National Research Agency
- Coordinator: INRIA Lille - Nord Europe (SequeL, Rémi Munos)
- Others partners: INRIA Saclay - Ile de France (TAO), HEC Paris (GREGHEC), Ecole Nationale des Ponts et Chaussées (CERTIS), Université Paris 5 (CRIP5), Université Paris Dauphine (LAMSADE).
- See also: https://sites.google.com/site/anrexplora/
- Activity Report: We developed bandit algorithm for planning in Markov Decision Processes based on the optimism in the face of uncertainty principle.

8.2.4. ANR CO-ADAPT

Participants: Alexandra Carpentier, Rémi Munos.

- Title: Brain computer co-adaptation for better interfaces
- Type: National Research Agency
- Duration: 2009-2013
- 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).
- Activity Report: In collaboration with Maureen Clerc and here student Joan Fruitet, we proposed a new Brain Computer interface procedure to select online a discriminative motor task based on a bandit algorithm. The efficient trading off between exploration (getting information about each motor tasks) and exploitation (selecting those that have highest classification rates) enables to reduce the time of the training session.

8.2.5. ANR AMATIS

Participant: Pierre Chainais.

- Title: Multifractal Analysis and Applications to Signal and Image Processing
- Type: National Research Agency
- Duration: 2011-2015
- 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.
- Coordinator: Univ. Paris-Est-Créteil (S. Jaffard)
- Activity Report: Ideas from the multifractal framework are the basis of our current work on the development of a new Bayesian approach to the blind deconvolution of noisy images.

8.2.6. National Partners

- INRIA Nancy - Grand Est, Team MAIA, France.
  - Bruno Scherrer *Collaborator*
    We have had collaboration on the topic of *approximate dynamic programming and statistical learning* and published a conference paper [25] and a technical report [51] this year.
- LITIS : Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes.
  - Stéphane Canu *Collaborator*
    Emmanuel Duflos and Hachem Kadri are collaborating with Pr. Stéphane Canu on Functional RKHS.

8.3. European Initiatives

8.3.1. FP7 Projects

8.3.1.1. PASCAL-2

- Participants: the whole SEQUEL is involved
- Title: Pattern Analysis, Statistical Modeling, and Computational Learning
- Type: Cooperation (ICT), Network of Excellence (NoE)
- Duration: March 2008 - February 2013
- Coordinator: Univ. Southampton
- Others partners: Many european organizations, universities, and research centers.
- See also: http://www.pascal-network.org/

8.3.1.2. PASCAL-2 Pump Priming Programme

Participants: Mohammad Ghavamzadeh, Rémi Munos.

- Title: Sparse Reinforcement Learning in High Dimensions
- Type: PASCAL-2 Pump Priming Programme
- Duration: November 2009 - March 2012
- Partners: INRIA Lille - Nord Europe, Shie Mannor (Technion, Israel)
- See also: http://sites.google.com/site/sparserl/home

8.3.1.3. CompLACS

Participants: Mohammad Ghavamzadeh, Alessandro Lazaric, Rémi Munos, Philippe Preux, Daniil Ryabko.

- Title: Composing Learning for Artificial Cognitive Systems
- Type: Cooperation (ICT), Specific Targeted Research Project (STREP)
- Duration: March 2011 - February 2015
- Coordinator: University College of London
- Partners: University College London, United Kingdom (John Shawe-Taylor, Stephen Hailes, David Silver, Yee Whye Teh), University of Bristol, United Kingdom (Nello Cristianini), Royal Holloway, United Kingdom (Chris Watkins), Radboud Universiteit Nijmegen, The Netherlands (Bert Kappen), Technische Universitat Berlin, Germany (Manfred Opper), Montanuniversitat Leoben, Austria (Peter Auer), Max-Planck Institute of Biological Cybernetics, Germany (Jan Peters).
- See also: http://www.complacs.org/

8.3.1.4. PIPER

Participant: Alessandro Lazaric.

- Title: New Paradigms for Preventing uncontrolled social influence in the future web
- Type: FET-Open Young Explorer Scheme
- Duration: *Submitted*
- Coordinator: Politecnico di Milano (Nicola Gatti)
- Partners: University of Southampton, United Kingdom (Valentin Robu, Enrico Gerdin, Nick Jennings).

8.4. International Initiatives

8.4.1. INRIA Associate Teams: SEQ-RL

- *Title*: Decision-making under Uncertainty with Applications to Reinforcement Learning, Control, and Games
- *INRIA principal investigator*: Rémi Munos
- *International Partner*:
  - *Institution*: University of Alberta (Canada)
  - *Laboratory*: Department of Computer Science
  - *Principal investigator*: Csaba Szepesvári
- *Duration*: January 2010 - January 2013
- *Website*: http://sites.google.com/site/associateteamualberta/home
- This associate team aims at bridging researchers from the SequeL team-project at INRIA Lille with the Department of Computing Science of the University of Alberta in Canada. Our common interest lies in machine learning, especially reinforcement learning, bandit algorithms and statistical learning with applications to control and computer games. The department of Computing Science at the University of Alberta is internationally renown as a leading research institute on these topics. The research work spans from theory to applications. Grounded on an already existing scientific collaboration, this associate team will make it easier to collaborate further between the two institutes, and thus strengthen this relationship. We foresee that the associate team will boost our collaboration, create new opportunities for financial support, and open-up a long-term fruitful collaboration between the two institutes. The collaboration will be through organizing workshops and exchanging researchers, postdoctoral fellows, and Ph.D. students between the two institutes.

8.4.2. INRIA International Partners

- University of Alberta, Edmonton, Alberta, Canada.
  - Prof. Csaba Szepesvari *Collaborator*
    We have been working on the topic of *regularized reinforcement learning* over the last four years. This year, we have one journal paper submitted [57] and one that will be submitted soon [8] on this topic. We are also coordinators of an *INRIA associate team program* with the university of Alberta.
  - Amir massoud Farahmand *Collaborator*
    We have been working on the topic of *regularized reinforcement learning* over the last five years. This year, we have one journal paper submitted [57] and one that will be submitted soon [8] on this topic.
- Technion - Israel Institute of Technology, Haifa, Israel.
  - Prof. Shie Mannor *Collaborator*
    We have been collaborating on the topic of *Bayesian reinforcement learning* for the last six years, on the topic of *regularized reinforcement learning* for the last four years, and on the topic of *reinforcement learning in high dimensions* in the last two year. On the first topic, we have a journal paper (survey) in preparation [58] this year. On the second topic, we have one journal paper under review [57] and one in preparation [8] this year. Finally, on the third topic, we were Co-PI's of a *PASCAL2 pump-priming program* that ended in June 2011.
- University of Waterloo, Waterloo, Ontario, Canada.
  - Prof. Pascal Poupart *Collaborator*
    We have been collaborating on the topic of *Bayesian reinforcement learning* in the last five years. This year, we have a journal paper in preparation [58] on this topic.
- Politecnico di Milano, Italy.
  - Prof. Marcello Restelli *Collaborator*
    We have been working on the topic of *transfer in reinforcement learning* over the last year. In particular, we have one conference paper [32] and a journal paper in preparation.
  - Prof. Nicola Gatti *Collaborator*
    We have started a collaboration on the topic of *bandit mechanisms for sponsored-search auction*. This year, we have submitted a paper to AAMAS [26] and we have collaborated on a proposal for a Marie Curie ITN and a Fet-Open Young Researcher proposal.
- University of Southampton, United Kingdom.
  - Prof. Enrico Gerding *Collaborator*
    We have been working on the topic of *learning and mechanism design* over the last year. In particular, we have collaborated on a proposal for a Marie Curie ITN and a Fet-Open Young Researcher proposal.

8.4.3. Visits of International Scientists

8.4.3.1. International Scientists

- Brahim Chaib-Draa, from Université Laval, Québec.
  His visit has been funded by Université de Lille 3 where he also taught.

- Mohammad G. Azar, Ph.D. student at University of Nijmegen, The Netherlands.
  Period: April 2011 - July 2011
  He worked with Rémi Munos and Mohammad Ghavamzadeh on performance analysis of reinforcement learning algorithms. The outcome of this collaboration has been a conference paper [16] and a technical report [48] so far.

8.4.3.2. Internship

- Matthew Hoffman, Ph.D. student at University of British Columbia, Canada.
  Period: October 2010 - April 2011.
  He worked with Alessandro Lazaric, Rémi Munos, and Mohammad Ghavamzadeh on our PASCAL2 Pump-Priming project on *sparse reinforcement learning in high dimensions*. The outcome of this collaboration has been a conference paper [61] so far.

9. Dissemination

9.1. Animation of the scientific community

9.1.1. Awards

Sébastien Bubeck received the second prize for the best French Ph.D in Artificial Intelligence (AI prize 2011).

9.1.2. Tutorials

- *R. Munos* co-organized with J.-Y. Audibert a tutorial on “Introduction to Bandits: Algorithms and Theory” (https://sites.google.com/site/banditstutorial/) at the International Conference of Machine Learning (ICML’11).

9.1.3. Workshops and Schools

- *R. Munos* co-organized the *Machine Learning Summer School 2011* (MLSS’11) in Bordeaux (2 weeks of lectures for about 80 international students), with François Caron, Manuel Davy, Pierre Del Moral, Pierrick Legrand, Manuel Lopes.
- *R. Munos* co-organized (with Florence Forbes, Bernard Espiau et Monique Thonnat) the *Journées INRIA autour de l’apprentissage statistique*, Décembre 2011.

9.1.4. Invited Talks

- *M. Ghavamzadeh*, Max Planck Institute for Intelligent Systems, Tübingen, Host: Prof. Jan Peters (June 2011).
- *M. Ghavamzadeh*, University of Liège - Systems & Modeling Research Unit, Host: Prof. Damien Ernst (June 2011).
- *M. Ghavamzadeh*, University of Waterloo - School of Computer Science, Host: Prof. Pascal Poupart (November 2010).
- *M. Ghavamzadeh*, McGill University - School of Computer Science, Host: Prof. Joelle Pineau (November 2010).
- *M. Ghavamzadeh*, University of Alberta - AI Seminar, Host: Prof. Csaba Szepesvári (November 2010).
- *A. Lazaric*, University of Liège - Systems & Modeling Research Unit, Host: Prof. Damien Ernst (October 2011).
- *R. Munos*, University of Liège, Department of Electical Engineering, February 2011.
- *R. Munos*, ICAPS, workshop Monte-Carlo Tree Search, Freiburg, June 2011

- *R. Munos*, Machine learning Summer School in Bordeaux, September 2011.
- *R. Munos*, Oxford, department of Computer Science, November 2011

9.1.5. Review Activities

- ***Participation to the program committees of international conferences***
  - *E. Duflos and P. Vanheeghe* were members of the Fusion'2011 International Program Committee
  - *E. Duflos* is reviewing papers for the following journals : IEEE Transaction on Signal Processing, International Journal of Approximate Reasoning, Information Fusion.
  - *P. Vanheeghe* is reviewing papers for the journal : IEEE Transaction on Signal Processing.
  - *D. Ryabko*: UAI 2011.
  - *M. Ghavamzadeh*: International Joint Conference on Artificial Intelligence (IJCAI 2011), European Workshop on Reinforcement Learning (EWRL 2011), International Conference on Artificial Neural Networks (ICANN 2011), National Conference on Artificial Intelligence (AAAI 2011).
  - *R. Munos*: Area chair for NIPS 2011.
  - *P. Preux*: ADPRL 2011, ICPRAM 2012, EGC 2011.
- ***International journal and conference reviewing activities*** (in addition to the conferences in which we belong to the PC)
  - *M. Ghavamzadeh* is Editorial Board Member of the Machine Learning Journal (MLJ, 2011-2014).
  - *M. Ghavamzadeh*: Annual Conference on Neural Information Processing Systems (NIPS 2011), International Conference on Artificial Intelligence and Statistics (AISTATS 2011), Neurocomputing, Machine Learning Journal (MLJ), Journal of Machine Learning Research (JMLR), Journal of Artificial Intelligence Research (JAIR).
  - *D. Ryabko*: IEEE Trans. Inf. Th., NIPS 2011.
  - *R. Munos*: ADPRL 2011, AISTATS 2011, ALT 2011, CAP 2011, ICML 2011, IJCAI 2011.
  - *P. Preux*: NIPS 2011, CAP 2011,IEEE Trans. on Neural Networks, Revue d'Intelligence Artificielle.
  - *A. Lazaric*: AAAI 2011, AAMAS 2011 & 2012, ACC, ALT, COLT, ICML, IJCAI, Journal of Artificial Intelligence Research (JAIR), Journal of Machine Learning Research (JMLR), IEEE Transactions on Automatic Controls (TAC).
  - *P. Chainais*: IEEE Trans. on Pattern Analysis and Machine Learning, Journal of Statistical Physics, Physica A.

9.1.6. Evaluation activities, expertise

- *Emmanuel Duflos* was appointed Director of Research of the Ecole Centrale in Lille. He has also reviewed proposals for the ANR programs.
- *M. Ghavamzadeh* is a grant proposal reviewer for the Natural Sciences and Engineering Research Council of Canada (NSERC).
- *J. Mary* is expert for the "Ministère de l'Enseignement Supérieur et de la Recherche" on control of"Crédit Impôt Recherche", member of the COS at Lille 3 for one assistant professor in computer science, member of the COS at 3 for one assistant professor un computer science at École Centrale de Lille for one assistant professor in computer science.

- *R. Munos* project evaluation for Research Foundation Flanders (FWO), Belgique, 2011, member of the evaluation committee for the Machine Learning, Université Libre de Bruxelles (ULB). Member of the Comité de sélection Professeur 27ème section for Polytech Paris-Sud, 2011.
- *P. Preux*: reviewer for the CNRS program PEPPI biology-mathematics-computer science, reviewer for the ANR program CONTINT, and the ANR program "blanc", president of the committee of selection (COS) at the University of Lille 3 for one assistant professor in computer science, member of the committee of selection (COS) at the École Centrale de Lille for one assistant professor in computer science.
- *D. Ryabko* is a member of COST-GRI evaluation committee.
- *Philippe Vanheeghe* has reviewed proposals for Discovery Grant applications of the Natural Sciences and Engineering Research Council of Canada (NSERRC - CRSNG), as well as for the ANR.

9.1.7. Participation to PhD and HDR jurys

- *D. Ryabko* is an examiner of the Ph.D. of K. Eltysheva.
- *R. Munos* examiner of the Ph.D. of Louis Dorard (University College of London), Raphael Fonteneau (University of Liège), and member of PhD juries for Lei Yu (University Cergy Pontoise) and Wassim Jouini (Supelec Rennes).
- *R. Munos* is member of HDR Committee for Daniil Ryabko (INRIA Lille - Nord Europe) and Aurélien Garivier (Télécom PariTech), 2011.
- *P. Preux* is member of the Ph.D. juries of Halem Benhabiles and Manuel Loth (Université de Lille 1)
- *E. Duflos* was *rapporteur* for the for PhD thesis of Michele Pace (INRIA Bordeaux), Pierre Neri (ENAC - University of Toulouse), Frédéric Faurie (University of Bordeaux), Sébastien Rougerie (University of Toulouse) and the Habilitation à Diriger des Recherche of Frédéric Dambreville.

9.1.8. Other Scientific Activities

- *R. Munos* is Vice Président du Comité des Projets at INRIA Lille-Nord Europe since September 2011.
- *R. Munos* is member of the Commission d'Evaluation INRIA.
- *R. Munos* is Président du jury d'admissibilité CR1-CR2 INRIA Lille - Nord Europe.
- *R. Munos* is member of the jury d'admission DR2 INRIA en 2011.

9.2. Teaching

9.2.1. Courses

- *R. Munos*, Master: "Introduction to Reinforcement Learning", 30 hours, M2, Master "Mathematiques, Vision, Apprentissage", ENS Cachan.
- *J. Mary*, Master : "Programmation web avancée et design pattern", 32h eq TD, M2, Université de Lille 3, France.
- *J. Mary*, Master : "Introduction à la Programmation R", 32h eq TD, M1, Université de Lille 3, France.
- *J. Mary*, Master : "Programmation R avancée", 32h eq TD, M1, Université de Lille 3, France.
- *P. Chainais*, Master: "Ondelettes et Applications", 24h, niveau M1, Ecole Centrale de Lille, 2ème année.
- *P. Chainais*, Master : "Décision et Apprentissage", 24h, niveau M2, Ecole Centrale de Lille, 3ème année.

- *P. Preux*, Master : “Décision dans l’incertain”, 40h, niveau M2 informatique, Lille 1.
- *P. Preux*, Master : “Mathematiques, Informatique, Modélisation”, 72h, niveau M1 psychologie, Lille 3.
- *E. Duflos*, Master : “Modélisation et Inférence Bayesienne”, 40h, niveau M2, Ecole Centrale de Lille, 3ème année.
- *P. Vanheeghe*, Master : “Estimation, Identification, Observation”, 32h, niveau M2, Ecole Centrale de Lille, 3ème année.

9.2.2. PhD and HdR

- HdR : *Daniil Ryabko*, Learnability in Problems of Sequential Inference, Université de Lille 1, December 19, 2011, [3].
- PhD : *Manuel Loth*, *Active Set Algorithms for the LASSO*, Université de Lille 1, July 8, 2011, Philippe Preux, [1].
- PhD : *Odalric Maillard*, *Active Set Algorithms for the LASSO*, Université de Lille 1 / Université de Toulouse, October 3, 2011, Rémi Munos and Philippe Berthet, [2].
- PhD: Nicolas Viandier, June 2011 (see section 10, [4]) : encadrement Emmanuel Duflos, Juliette Marais (IFSTTAR).
- PhD in progress : *Boris Baldassari*, “Apprentissage automatique et développement logiciel”, Sep. 2011, encadrement : Ph. Preux.
- PhD in progress : *Victor Gabillon*, “Active Learning in Classification-based Policy Iteration”, Sep. 2009, encadrement : M. Ghavamzadeh, Ph. Preux.
- PhD in progress : *Azadeh Khaleghi*, “Unsupervised Learning of Sequential Data”, Sep. 2010, encadrement : D. Ryabko, Ph. Preux.
- PhD in progress : *Sami Naamane*, “Filtrage collaboratif adverse et dynamique”, Nov. 2011, encadrement : J. Mary, Ph. Preux.
- PhD in progress : *Olivier Nicol*, “Apprentissage par renforcement sous contrainte de ressources finies, dans un environnement non stationnaire, face à des flux de données massifs”, Nov. 2010, encadrement : J. Mary, Ph. Preux.
- PhD in progress : *Christophe Salperwyck*, “Apprentissage incrémentale et sur flux de données” , Dec. 2009, encadrement : Ph. Preux.
- PhD in progress : *Amir Sani*, “Learning under uncertainty”, Oct. 2011, encadrement : R. Munos, A. Lazaric.
- PhD in progress : *Jean-François Hren*, “Prise de décision et planification optimiste”, Oct. 2007, encadrement : R. Munos.
- PhD in progress : *Alexandra Carpentier*, “Allocation adaptatives de ressources pour l’apprentissage actif”, Oct. 2007, encadrement : R. Munos.
- PhD in progress : *Emilie Kaufmann*, “Bayesian Bandits”, Oct. 2011, encadrement : R. Munos, O. Cappé, A. Garivier.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] M. LOTH. *Active Set Algorithms for the LASSO*, Université Lille 1 Sciences et Technologies, 2011.

[2] O.-A. MAILLARD. *Apprentissage séquentiel: bandits, statistique et renforcement*, Université Lille 1, Lille, France, Octobre 2011.

[3] D. RYABKO. *Learnability in Problems of Sequential Inference*, Université Lille 1 Sciences et Technologies, 2011.

[4] N. VIANDIER. *Modélisation et utilisation des erreurs de pseudodistances GNSS en environnement transport pour l'amélioration des performances de localisation.*, Ecole Centrale de Lille, Juin 2011.

Articles in International Peer-Reviewed Journal

[5] S. BUBECK, R. MUNOS, G. STOLTZ. *Pure Exploration in Finitely-Armed and Continuous-Armed Bandits*, in "Theoretical Computer Science", 2011, vol. 412, p. 1832-1852.

[6] S. BUBECK, R. MUNOS, G. STOLTZ, C. SZEPESVÁRI. *X-Armed Bandits*, in "Journal of Machine Learning Research", 2011, vol. 12, p. 1655-1695.

[7] P. CHAINAIS, E. KŒNIG, V. DELOUILLE, JEAN-FRANÇOIS. HOCHEDEZ. *Virtual Super Resolution of Scale Invariant Textured Images Using Multifractal Stochastic Processes*, in "Journal of Mathematical Imaging and Vision", 2011, vol. 39, n^{o} 1, p. 28-44.

[8] A. M. FARAHMAND, M. GHAVAMZADEH, CS. SZEPESVÁRI, S. MANNOR. *L2-Regularized Policy Iteration*, in "Journal of Machine Learning Research", 2011, submitted.

[9] S. GIRGIN, J. MARY, P. PREUX, O. NICOL. *Managing Advertising Campaigns – an Approximate Planning approach*, in "Frontiers in Computer Science", October 2011.

[10] A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Finite-Sample Analysis of Least-Squares Policy Iteration*, in "Journal of Machine learning Research", 2011, To appear.

[11] A. LAZARIC, R. MUNOS. *Learning with Stochastic Inputs and Adversarial Outputs*, in "Journal of Computer and System Sciences", 2011, To appear.

[12] B. LEBENTAL, P. CHAINAIS, P. CHENEVIER, N. CHEVALIER, E. DELEVOYE, J.-M. FABBRI, S. NICOLETTI, P. RENAUX, A. GHIS. *Aligned carbon nanotube based ultrasonic microtransducers for durability monitoring in civil engineering*, in "Nanotechnology", 2011, vol. 22, n^{o} 39.

[13] D. RYABKO. *On the relation between realizable and non-realizable cases of the sequence prediction problem.*, in "Journal of Machine Learning Research", 2011, vol. 12, p. 2161-2180.

[14] D. RYABKO. *Testing composite hypotheses about discrete ergodic processes*, in "Test", 2011, (to appear).

[15] B. RYABKO, D. RYABKO. *Constructing Perfect Steganographic Systems*, in "Information and Computation", 2011, vol. 209, n^{o} 9, p. 1223-1230.

International Conferences with Proceedings

[16] M. G. AZAR, R. MUNOS, M. GHAVAMZADEH, H. KAPPEN. *Speedy Q-Learning*, in "Proceedings of Advances in Neural Information Processing Systems 24", MIT Press, 2011.

[17] L. BUSONIU, R. MUNOS, B. DE SCHUTTER, R. BABUSKA. *Optimistic Planning for Sparsely Stochastic Systems*, in "IEEE International Symposium on Adaptive Dynamic Programming and Reinforcement Learning", 2011.

[18] A. CARPENTIER, A. LAZARIC, M. GHAVAMZADEH, R. MUNOS, P. AUER. *Upper-Confidence-Bound Algorithms for Active Learning in Multi-Armed Bandits*, in "Proceedings of the Twenty-Second International Conference on Algorithmic Learning Theory", 2011, p. 189-203.

[19] A. CARPENTIER, O.-A. MAILLARD, R. MUNOS. *Sparse Recovery with Brownian Sensing*, in "Advances in Neural Information Processing Systems", 2011.

[20] A. CARPENTIER, R. MUNOS. *Finite Time Analysis of Stratified Sampling for Monte Carlo*, in "Advances in Neural Information Processing Systems", 2011.

[21] P. CHAINAIS, V. DELOUILLE, JEAN-FRANÇOIS. HOCHEDEZ. *Scale invariant images in astronomy through the lens of multifractal modeling*, in "2011 IEEE International Conference on Image Processing (IEEE ICIP2011)", Brussels, Belgium, 9 2011.

[22] E. DELANDE, E. DUFLOS, P. VANHEEGHE, D. HEURGUIER. *Multi-Sensor PHD by Space Partionning: Computation of a True Reference Density Within The PHD Framework*, in "Statistical Signal Processing Workshop (SSP), 2011", Nice, France, IEEE - SIGNAL PROCESSING SOCIETY (editor), IEEE - Signal Processing Society, June 2011, p. 333 - 336 [*DOI :* 10.1109/SSP.2011.5967695], http://hal.inria.fr/hal-00639710/en/.

[23] E. DELANDE, E. DUFLOS, P. VANHEEGHE, D. HEURGUIER. *Multi-Sensor PHD: Construction and Implementation by Space Partitioning*, in "International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011", Prague, Tchèque, République, IEEE - SIGNAL PROCESSING SOCIETY (editor), IEEE - Signal Processing Society, May 2011, p. 3632 - 3635 [*DOI :* 10.1109/ICASSP.2011.5947137], http://hal.inria.fr/hal-00639724/en/.

[24] V. GABILLON, M. GHAVAMZADEH, A. LAZARIC, S. BUBECK. *Multi-Bandit Best Arm Identification*, in "Proceedings of Advances in Neural Information Processing Systems 24", MIT Press, 2011.

[25] V. GABILLON, A. LAZARIC, M. GHAVAMZADEH, B. SCHERRER. *Classification-based Policy Iteration with a Critic*, in "Proceedings of the Twenty-Eighth International Conference on Machine Learning", 2011, p. 1049-1056.

[26] N. GATTI, A. LAZARIC, F. TROVÓ. *A Truthful Learning Mechanism for Contextual Multi-Slot Sponsored Search Auctions with Externalities*, in "AAMAS'12", 2011, submitted.

[27] M. GHAVAMZADEH, A. LAZARIC, R. MUNOS, M. HOFFMAN. *Finite-Sample Analysis of Lasso-TD*, in "Proceedings of the Twenty-Eighth International Conference on Machine Learning", 2011, p. 1177-1184.

[28] M. GHAVAMZADEH, A. LAZARIC, R. MUNOS, O.-A. MAILLARD. *LSTD with Random Projections*, in "Proceedings of the Twenty-Fourth Annual Conference on Advances in Neural Information Processing Systems", 2011.

[29] N. JAOUA, E. DUFLOS, P. VANHEEGHE, L. CLAVIER, F. SEPTIER. *Impulsive interference mitigation in ad hoc networks based on alpha-stable modeling and partile filtering*, in "International Conference

on Acoustics, Speech and Signal Processing (ICASSP), 2011", Prague, Tchèque, République, IEEE - SIGNAL PROCESSING SOCIETY (editor), IEEE - Signal Processing Society, May 2011, p. 3548 - 3551 [*DOI :* 10.1109/ICASSP.2011.5946244], http://hal.inria.fr/hal-00640682/en/.

[30] H. KADRI, E. DUFLOS, P. PREUX, S. CANU. *Multiple functional regression with both discrete and continuous covariates*, in "Proc. of the 2nd International Workshop on Functional and Operatorial Statistics", F. FERRATY (editor), Contributions to Statistics, Physica-Verlag HD, June 2011, p. 189-195, Recent Advances in Functional Data Analysis and Related Topics, Chapter 29.

[31] H. KADRI, E. DUFLOS, P. PREUX. *Learning Vocal Tract Variables With Multi-Task Kernels*, in "Proc. 36th IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP)", IEEE, May 2011.

[32] A. LAZARIC, M. RESTELLI. *Transfer from Multiple MDPs*, in "Advances in Neural Information Processing Systems", August 2011.

[33] O.-A. MAILLARD, R. MUNOS. *Adaptive bandits: Towards the best history-dependent strategy*, in "International conference on Artificial Intelligence and Statistics", 2011.

[34] O.-A. MAILLARD, R. MUNOS, D. RYABKO. *Selecting the State-Representation in Reinforcement Learning*, in "Advances in Neural Information Processing Systems", 2011.

[35] O.-A. MAILLARD, R. MUNOS, G. STOLTZ. *Finite-Time Analysis of Multi-armed Bandits Problems with Kullback-Leibler Divergences*, in "Conference On Learning Theory", 2011.

[36] R. MUNOS. *Optimistic Optimization of Deterministic Functions without the Knokledge of its Smoothness*, in "Advances in Neural Information Processing Systems", 2011.

[37] A. RABAOUI, E. DUFLOS, N. VIANDIER, J. MARAIS. *Selecting the Hyperparameters of the DPM Models fir the density estimation of observation errors*, in "International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011", Prague, Tchèque, République, IEEE - SIGNAL PROCESSING SOCIETY (editor), IEEE - Signal Processing Society, May 2011, p. 4092 - 4095.

[38] A. RABAOUI, H. KADRI, P. PREUX, E. DUFLOS, A. RAKOTOMAMONJY. *Functional Regularized Least Squares Classification with Operator-Valued Kernels*, in "Proc. 28th International Conference on Machine Learning (ICML)", New York, NY, USA, L. GETOOR, T. SCHEFFER (editors), ACM, June 2011, p. 993–1000.

[39] B. RYABKO, D. RYABKO. *Confidence Sets in Time–Series Filtering*, in "Proc. 2011 IEEE International Symposium on Information Theory (ISIT)", St. Petersburg, Russia, 2011, p. 2436-2438.

[40] C. SALPERWYCK, V. LEMAIRE. *Incremental discretization for supervised learning*, in "CLADAG: CLAssification and Data Analysis Group — 8th International Meeting of the Italian Statistical Society", September 2011.

[41] C. SALPERWYCK, V. LEMAIRE. *Learning with few examples: an empirical study on leading classifiers*, in "International Joint Conference on Neural Networks (IJCNN)", IEEE, August 2011.

National Conferences with Proceeding

[42] P. Chainais, M. Chevaldonné, J.-M. Favreau. *Synthèse de textures multifractales directement sur des surfaces 3D*, in "Proc. of GRETSI", 2011.

[43] P. Chainais, B. Lebental. *Caractérisation statistique d'une assemblée de nanotubes en imagerie microscopique*, in "Proc. of GRETSI", 2011.

Scientific Books (or Scientific Book chapters)

[44] G. Arnold-Dulac, L. Denoyer, P. Preux, P. Gallinari. *Datum-wise classification. A sequential Approach to sparsity*, in "Machine Learning and Knowledge Discovery in Databases", D. Gunopulos, T. Hofmann, D. Malerba, M. Vazirgiannis (editors), Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, vol. 6911, p. 375-390, Proc. European Conference on Machine Learning (ECML), http://dx.doi.org/10.1007/978-3-642-23780-5_34.

[45] L. Busoniu, A. Lazaric, M. Ghavamzadeh, R. Munos, R. Babuska, B. De Schutter. *Least-squares methods for policy iteration*, in "Reinforcement Learning: State of the Art", M. Wiering, M. van Otterlo (editors), Springer, 2011.

[46] L. Busoniu, R. Munos, R. Babuska. *Optimistic Planning in Markov decision processes*, in "Reinforcement Learning and Adaptive Dynamic Programming for feedback control", F. Lewis, D. Liu (editors), Wiley, 2011, To appear.

[47] S. Clémençon, R. Gaudel, J. Jakubowicz. *Clustering Rankings in the Fourier Domain*, in "Machine Learning and Knowledge Discovery in Databases: Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD'11)", D. Gunopulos, T. Hofmann, D. Malerba, M. Vazirgiannis (editors), Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2011, p. 343–358, http://dx.doi.org/10.1007/978-3-642-23780-5_32.

Research Reports

[48] M. G. Azar, R. Munos, M. Ghavamzadeh, H. Kappen. *Reinforcement Learning with a Near Optimal rate of Convergence*, INRIA, 2011, n^{o} inria-00636615, http://hal.inria.fr/inria-00636615/en/.

[49] A. Carpentier, A. Lazaric, M. Ghavamzadeh, R. Munos, P. Auer. *Upper-Confidence-Bound Algorithms for Active Learning in Multi-Armed Bandits*, INRIA, 2011, n^{o} inria-00594131.

[50] V. Gabillon, M. Ghavamzadeh, A. Lazaric, S. Bubeck. *Multi-Bandit Best Arm Identification*, INRIA, 2011, n^{o} inria-00632523, http://hal.inria.fr/hal-00632523_v3/.

[51] V. Gabillon, M. Ghavamzadeh, A. Lazaric, B. Scherrer. *Classification-based Policy Iteration with a Critic*, INRIA, 2011, n^{o} inria-00590972, http://hal.inria.fr/hal-00590972_v1/.

[52] M. Ghavamzadeh, A. Lazaric, O.-A. Maillard, R. Munos. *LSPI with Random Projections*, INRIA, 2011, n^{o} inria-00530762, http://hal.inria.fr/inria-00530762_v1/.

[53] H. Kadri, P. Preux, E. Duflos, S. Canu. *Operator-valued Kernels for Nonparametric Operator Estimation*, INRIA, 2011, n^{o} 7607, http://hal.inria.fr/inria-00587649/en.

Other Publications

[54] G. ARNOLD-DULAC, L. DENOYER, P. PREUX, P. GALLINARI. *Sequential Approaches for Learning Datum-Wise Sparse Representations*, October 2011, (submitted).

[55] J. DAQUIN. *Factorisation non-négative de matrices*, Université de Lille, 2011, master in applied mathematics, Ph. Preux, B. Beckermann advіors.

[56] E. DUFLOS, N. VIANDIER, J. MARAIS, A. RABAOUI, P. VANHEEGHE. *GNSS Urban Localization Enhencement using Dirichlet Process Mixture Modelling*, December 2011, Workshop Non Parametric Bayes at NIPS 2011 (Genada, Spain).

[57] A. M. FARAHMAND, M. GHAVAMZADEH, CS. SZEPESVÁRI, S. MANNOR. *L2-Regularized Fitted-Q Iteration Algorithm*, 2011, in preparation.

[58] M. GHAVAMZADEH, S. MANNOR, P. POUPART. *Bayesian Reinforcement Learning: A Survey*, 2011, in preparation.

[59] S. GIRGIN. *VVU Project report*, July 2011, Deliverable, VVU project, PICOM, France.

[60] S. GIRGIN, P. PREUX. *Identification of prospective clients*, October 2011, Report for the contract with Addressing Business (confidential).

[61] M. HOFFMAN, A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Regularized Least Squares Temporal Difference Learning with Nested ℓ_2 and ℓ_1 Penalization*, in "Ninth European Workshop on Reinforcement Learning", 2011.

[62] N. JAOUA, E. DUFLOS, P. VANHEEGHE. *Nonparametric Bayesian state estimation in nonlinear dynamic systems with alpha-stable measurement noise*, December 2011, Workshop Non Parametric Bayes at NIPS 2011 (Genada, Spain).

[63] O. NICOL. *On-line Trading of Exploration and Exploitation*, June 2011, invited speech Exploration/Exploitation challenge ICML workshop.

[64] C. SALPERWYCK, V. LEMAIRE. *Impact de la taille de l'ensemble d'apprentissage : une étude empirique*, January 2011, Atelier CIDN : Clustering incrémental et méthodes de détection de nouveauté de la conférence Extraction et Gestion des Connaissances (EGC).

[65] C. SALPERWYCK, T. URVOY. *On-line Trading of Exploration and Exploitation*, June 2011, invited speech Exploration/Exploitation challenge ICML workshop.

References in notes

[66] 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, p. 235–256.

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

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

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

[70] E. DELANDE, E. DUFLOS, D. HEURGUIER, P. VANHEEGHE. *Multi-target PHD filtering: proposition of extensions to the multi-sensor case*, INRIA, 2010, n^o 7337.

[71] E. DUFLOS, S. RAZAVI, C. HAAS, P. VANHEEGHE. *Belief Function Based Algorithm for Material Detection and Tracking in Construction*, in "Proceedings of Workshop on the theory of belief functions", April 2010, CDROM - 6 pages.

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

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

[74] H. KADRI, E. DUFLOS, P. PREUX, S. CANU, M. DAVY. *Nonlinear functional regression: a functional RKHS approach*, in "Proc. of the 13th Artificial Intelligence and Statistics (AI & Stats), JMLR: W&CP 9", May 13-15 2010, p. 374–380.

[75] J. MARAIS, E. DUFLOS, N. VIANDIER, D. NAHIMANA, A. RABAOUI. *Advanced signal processing techniques for multipath mitigation in land transportation environment*, in "Proceedings of ITSC 2010", September 2010, Proceedings on CD ROM (6 pages).

[76] J. MARAIS, N. VIANDIER, A. RABAOUI, E. DUFLOS. *GNSS multipath bias models for accurate positioning in urban environments*, in "Proceedings of ITST 2010", November 2010, Proceedings on CD ROM (6 pages).

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

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

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

[80] 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, p. 781–831, http://gemini.econ.umd.edu/jrust/research/rustphelan.pdf.

[81] J. RUST. *On the Optimal Lifetime of Nuclear Power Plants*, in "Journal of Business & Economic Statistics", 1997, vol. 15, n^o 2, p. 195–208, http://129.3.20.41/eprints/io/papers/9512/9512002.abs.

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

[83] G. TESAURO. *Temporal Difference Learning and TD-Gammon*, in "Communications of the ACM", March 1995, vol. 38, n^o 3, http://www.research.ibm.com/massive/tdl.html.

[84] N. VIANDIER, A. RABAOUI, J. MARAIS, E. DUFLOS. *GNSS pseudorange error density tracking using Dirichlet Process Mixture*, in "Proceedings of FUSION 2010", July 2010, Proceedings on CD ROM (7 pages).

[85] N. VIANDIER, A. RABAOUI, J. MARAIS, E. DUFLOS. *Studies on DPM for the density estimation of pseudorange noises and evaluations on real data*, in "Proceedings of IEEE Plans", May 2010, Proceedings on CD ROM (8 pages).

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