INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team SequeL

Sequential Learning

Lille - Nord Europe

Theme : Optimization, Learning and Statistical Methods

Activity Report

2010

Table of contents

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. Team

Research Scientists

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

Faculty Members

Philippe Preux [Team leader, Professor, Université de Lille, secondment at the INRIA until Aug. 31st, 2009, 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]
Jérémie Mary [Assistant professor, Université de Lille 3]

PhD Students

Sébastien Bubeck [ENS Grant, Oct., 2007 - June, 2010]
Alexandra Carpentier [ANR-Région Nord-Pas de Calais Grant, since Oct., 2009]
Pierre-Arnaud Coquelin [École Polytechnique, since Oct., 2005, currently mostly CO of the start-up Vekia that he created in 2007]
Emmanuel Delande [DGA, since Nov., 2008]
Victor Gabillon [MENESR Grant, since Oct., 2009]
Jean-François Hren [MENESR Grant, since Oct., 2007]
Robin Jaulmes [DGA Grant, since Oct., 2006]
Nouha Jaoua [MENESR Grant, since 2009]
Azadeh Khaleghi [CORDIS grant, since Oct., 2010]
Manuel Loth [INRIA-Région Nord-pas-de-calais Grant, Oct., 2006- Sept. 2009; ATER since Sept. 2009]
Odalric-Ambrym Maillard [ENS Grant, since Oct., 2008]
Olivier Nicol [MENESR Grant, since Oct., 2010; from Mar. to Aug. 2010 intern, M2, Univeristé de Lille]
Christophe Salperwyck [CIFRE with France Telecom Grant, since Dec., 2009]
Nicolas Viandier [INRETS, Oct., 2007 - Feb. 2010]

Post-Doctoral Fellows

Sertan Girgin [Région Nord-Pas de Calais]
Alessandro Lazaric [INRIA until Sep. 30th]
Hachem Kadri [Région Nord-Pas de Calais]

Administrative Assistant

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

Others

Antoine Chamot [Master internship, École Centrale de Lille, June to Aug. 2010]
Geoffrey Megardon [Master 2 internship, Université de Lille, May to Aug. 2010]

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 [79]). 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

- This year, Sébastien Bubeck has defended his Ph.D. thesis, entitled “Bandits Games and Clustering Foundations” [11]. Not only is it the first Ph.D. defence in SequeL, but it is also a highly successful one: Sébastien has been awarded a Gilles Kahn 2010 prize, a prize awarded by Specif to the best Ph.D. theses in Computer Science, in France (patronized by the Academy of Science). The thesis supervisor was Rémi Munos.
- The research of Rémi Coulom on artificial Go players has received further recognition, in the form of two important international awards (see Section 9.1). This work has been also highlighted in the popular science magazine “Pour la Science”, featuring on the cover [51].

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 [75].

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

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

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

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

[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.

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

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 [67]) 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) [65], 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 [66]. 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 ([81]):

- 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 [76], 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.* [64] 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 behavior (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 behavior, or than a class of other behaviors.

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 $x^1 = (x^1_1, \cdots, x^1_{n_1}), \cdots, x^N = (x^1_N, \cdots, x^N_{n_N})$, 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: AAAI 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 [71]. 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 [69] 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]
- $Z_{k+1} = \{z_1, ..., z_m\}$ is the current multi-sensor observation, i.e. a collection of measures z_i produced at time $k+1$ by all the sensors;
- $Z^{(k)} = \bigcup_{t \leqslant k} Z_t$ is the collection of observations up to time k;
- $f_{k|k}(W|Z^{(k)})$ is the current multi-target posterior density in state W;
- $f_{k+1|k}(X|W)$ is the current multi-target Markov transition density, from state W to state X;
- $f_{k+1}(Z|X)$ is the current multi-sensor/multi-target likelihood function.

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

Although equations (5) may seem similar to the classical single-sensor/single-target Bayesian equations, they are generally intractable because of the presence of the *set integrals*. For, a RFS Ξ is characterized by the family of its Janossy densities $j_{\Xi,1}(x_1)$, $j_{\Xi,2}(x_1, x_2)$... and not just by one density as it is the case with vectors. Mahler then introduced the PHD, defined on single-target state space. The PHD is the quantity whose integral on any region S is the expected number of targets inside S. Mahler proved that the PHD is the first-moment density of the multi-target probability density. Although defined on single-state space X, the PHD encapsulates information on both target number and states. 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 precisely aims at controlling 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 [74]. 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 [80] 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 [74]
- 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 “ambiant 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 [78], [77] 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 analyze 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. Software

5.1.1. Crazy Stone

Participant: Rémi Coulom.

Crazy Stone, is a top-level Go-playing program that has been developed by Rémi Coulom since 2005. Crazy Stone won several major international Go tournaments in the past. In 2010, a license of Crazy Stone was sold to a Japanese company, Unbalance Corporation. Crazy Stone should be available for sale in Japan in 2011.

6. New Results

6.1. Introduction

New results are organized in the following sections:

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

6.2. Decision under uncertainty

Participants: Sébastien Bubeck, Alexandra Carpentier, Pierre-Arnaud Coquelin, Rémi Coulom, Victor Gabillon, Mohammad Ghavamzadeh, Sertan Girgin, Jean-François Hren, Alessandro Lazaric, Manuel Loth, Odalric-Ambrym Maillard, Rémi Munos, Olivier Nicol, Philippe Preux.

6.2.1. Reinforcement learning and approximate dynamic programming

6.2.1.1. Strengthening the 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 approximate dynamic programming (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. I believe that 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. We derived the first performance bounds for linear function spaces for two widely-used ADP algorithms: least-squares temporaldifference learning [31], [50], [62] and Bellman residual minimization [34]. We also presented the first complete analysis of classification-based policy iteration algorithms, a relatively new and not well-studied class of ADP methods [30], [49], [61]. These algorithms work without explicit value function representation, and define the evaluation of the policy at each iteration and the generation of the next policy together as a classification problem. We have also studied algorithmic methods that can improve sample efficiency in this class of algorithms [54].

6.2.1.1.1. Error Propagation for Approximate Policy and Value Iteration

In this work [22] we address the question of how the approximation error/Bellman residual at each iteration of Approximate Policy / Value Iteration algorithms influence the quality of the resulted policy. We quantify the performance loss as L_p norm of the approximation error / Bellman residual at each iteration. Moreover, we show that the performance loss depends on the expectation of the squared Radon-Nikodym derivative of a certain distribution rather than its supremum as opposed to the previous results. Also our results indicate that the contribution of the Bellman residual / approximation error to the performance loss is more prominent in latter iterations of API/AVI, and the effect of an error term in early iterations decays exponentially fast.

6.2.1.2. Bayesian Multi-Task Reinforcement Learning:

In this work [29], [48], we consider the multi-task RL (MTRL) scenario in which the learner is provided with a number of MDPs with common state and action spaces. For any given policy, only a small number of samples can be generated in each MDP, which may not be enough to accurately evaluate the policy. In such a MTRL problem, it is necessary to identify classes of tasks with similar structure and to learn them jointly. We considered a particular class of MTRL problems in which the tasks share structure in their value functions. To allow the value functions to share a common structure, it is assumed that they are all sampled from a common prior.We adopted the Gaussian process temporal-difference value function model for each task, modeled the distribution over the value functions using a hierarchical Bayesian model, and developed solutions to the following problems: (i) joint learning of the value functions (multi-task learning), and (ii) efficient transfer of the information acquired in (i) to facilitate learning the value function of a newly observed task (transfer learning).

6.2.1.3. RL in High-dimensional Spaces:

A primary goal here is to devise RL algorithms whose sample and computational complexities do not grow rapidly with the dimension of the state space. We have particularly 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 derived and analyzed a least-squares policy iteration algorithm with random projections [24].

6.2.1.4. Bayesian Policy Gradient and Actor-Critic Algorithms:

Mohammad Ghavamzadeh continued his collaboration with Yaakov Engel (Haifa Israel) on this topic [56], [55]. In this work, we used Bayesian reasoning to develop more sample-efficient policy gradient and actor-critic algorithms. We proposed a Bayesian framework that models the policy gradient as a Gaussian process. This reduces the number of samples needed to obtain accurate gradient estimates, resulting in faster convergence than the conventional Monte-Carlo-based policy gradient and actor-critic algorithms. Moreover, estimates of the natural gradient and a measure of the uncertainty in the gradient estimates, namely, the gradient covariance, are provided at little extra cost.

6.2.1.5. Regularization in RL:

Mohammad Ghavamzadeh continued his 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 [52], [53]. The designed algorithms work in both linear and reproducing kernel Hilbert spaces.

6.2.2. Planning and exploration vs. exploitation trade-off

6.2.2.1. Open Loop Optimistic Planning

In this work [18] we consider the problem of planning in a stochastic and discounted environment with a limited numerical budget. More precisely, we investigate strategies exploring the set of possible sequences of actions, so that, once all available numerical resources (e.g. CPU time, number of calls to a generative model) have been used, one returns a recommendation on the best possible immediate action to follow based on this exploration. The performance of a strategy is assessed in terms of its simple regret, that is the loss in performance resulting from choosing the recommended action instead of an optimal one. We first provide a minimax lower bound for this problem, and show that a uniform planning strategy matches this minimax rate (up to a logarithmic factor). Then we propose a UCB (Upper Confidence Bounds)-based planning algorithm, called OLOP (Open-Loop Optimistic Planning), which is also minimax optimal, and prove that it enjoys much faster rates when there is a small proportion of near-optimal sequences of actions. Finally, we compare our results with the regret bounds one can derive for our setting with bandits algorithms designed for an infinite number of arms.

6.2.2.2. Best Arm Identification in Multi-Armed Bandits

In this work [17] we consider the problem of finding the best arm in a stochastic multi-armed bandit game. The regret of a forecaster is here defined by the gap between the mean reward of the optimal arm and the mean reward of the ultimately chosen arm. We propose a highly exploring UCB policy and a new algorithm based on successive rejects. We show that these algorithms are essentially optimal since their regret decreases exponentially at a rate which is, up to a logarithmic factor, the best possible. However, while the UCB policy needs the tuning of a parameter depending on the unobservable hardness of the task, the successive rejects policy benefits from being parameter-free, and also independent of the scaling of the rewards.

6.2.3. Applications

6.2.3.1. The Ubiquitous Virtual Seller

The first months of work on this project during the Fall of 2009 led us to the conclusion that an important work on the recommendation of products was required to help such a virtual seller. So, we focused on this issue, considering that the problem of recommendation systems is a very general setting that can be tailored to solve many problems (the problem of ad selection on the web for instance, see below), and that our group has to get more acquainted with this domain of research.

In 2010, this work has been mostly of technological nature. We designed a recommendation system to recommend products on a commercial website. This recommendation system comes as a plug-in for the Firefox web browser than can be enabled at will, and that automatically embed recommended products on product pages.

See also the contract section (Sec. 7.1.6) of the report for specific details about the contract itself.

6.2.3.2. Ad selection on web portals

We continued the work initiated in 2009 work on the selection of displayed ads on web pages. We have been able to propose algorithms that significantly improve the resolution of this problem [25], [58]. In particular, we have shown that optimizing advertisement display, handling finite budgets and finite lifetimes in a dynamic and non-stationary setting, is feasible within realistic computational time constraints (such as serving several dozens of ads per second). We have also given some insights in what can be gained by handling these constraints, depending on the properties of the advertisements to display.

Furthermore, Orange provided us some web log files related to the ad service. We have begun to mine these web logs to get more accurate figures about real data, and the real behavior of human beings facing ads on web pages. However, due to the enormous size of these logs, the work has not gone as far as we wished. We have actually acquired a computer with a very large main memory (256 Gbytes) to handle such large datasets. In the coming years, we wish to work on very large datasets (tera-bytes and more), so that this work is our first step towards working with very large datasets.

6.3. Foundations of machine learning

Participant: Daniil Ryabko.

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 $\mathcal{C}$.

This general formulation is motivated by the diversity of sequential prediction problems: they include analysis of biological, financial, textual or web-generated data, to mention a few. Naturally, one has to have different models for these problems, and therefore one is interested in finding a general procedure for constructing a predictor, given only some weak probabilistic constraints on the data; this is formalized by saying that the data-generating process comes from a known but arbitrary family $\mathcal{C}$.

6.3.1.1. Finding predictors for arbitrary families of processes: realizable case

The contribution of this work [15] is in characterizing the families $\mathcal{C}$ of measures for which asymptotically accurate predictors exist, and in providing a specific and simple form in which to look for a solution. We show that if any predictor works, then there exists a Bayesian predictor, whose prior is discrete, and which works too. We also find several sufficient and necessary conditions for the existence of a predictor, in terms of topological characterizations of the family $\mathcal{C}$, as well as in terms of local behavior of the measures in $\mathcal{C}$, which in some cases lead to procedures for constructing such predictors.

It should be emphasized that the framework is completely general: the stochastic processes considered are not required to be i.i.d., stationary, or to belong to any parametric or countable family.

6.3.1.2. 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 [40] 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 have developed a new theoretical framework that has allowed us to solve some classical problems of mathematical statistics in a radically more general setting. Namely, the setting is that the data is generated by a stationary ergodic process (or processes, depending on the problem), and no assumptions of independence, mixing rates, etc., as well as no parametric assumptions, are made. The obtained results include a general hypothesis testing procedure, a consistent change point estimator, and a consistent classification procedure [16]. Previous results on these problems concerned only much more restricted settings (e.g. i.i.d. data).

6.3.2.1. An impossibility result on process discrimination

We have shown [14] that consistent homogeneity testing is impossible in this setting, which means that given two growing samples of data which are only known to be generated by stationary ergodic processes, one cannot in general tell whether they are generated by the same or by different process distributions, even in the weakest asymptotic setting, and even if the processes are binary-valued. This is particularly remarkable in view of our result that establishes a consistent change point estimator. This also solves an open problem about discrimination between ergodic processes [72].

6.3.2.2. A criterion for the existence of consistent tests

The most general result that we have obtained [41] 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.

6.4. Supervised learning

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

6.4.1. Regression and classification

6.4.1.1. Scrambled objects for Least Squares Regression

In this work [36] we consider least-squares regression using a randomly generated subspace $\mathcal{G}_P \subset \mathcal{F}$ of finite dimension P, where $\mathcal{F}$ is a function space of infinite dimension, e.g. $L_2([0,1]^d)$. $\mathcal{G}_P$ is defined as the span of P random features that are linear combinations of the basis functions of $\mathcal{F}$ weighted by random Gaussian i.i.d. coefficients. In particular, we consider multi-resolution random combinations at all scales of a given mother function, such as a hat function or a wavelet. In this latter case, the resulting Gaussian objects are called *scrambled wavelets* and we show that they enable to approximate functions in Sobolev spaces $H^s([0,1]^d)$. As a result, given N data, the least-squares estimate $\widehat{g}$ built from P scrambled wavelets has excess risk $||f^* - \widehat{g}||^2_{\mathcal{P}} = O(||f^*||^2_{H^s([0,1]^d)}(\log P)/P + P(\log N)/N)$ for target functions $f^* \in H^s([0,1]^d)$ of smoothness order $s > d/2$. An interesting aspect of the resulting bounds is that they do not depend on the distribution $\mathcal{P}$ from which the data are generated, which is important in a statistical regression setting considered here. Randomization enables to adapt to any possible distribution. We describe an efficient numerical implementation using lazy expansions with numerical complexity $\widetilde{O}(2^d N^{3/2} \log N + N^2)$, where d is the dimension of the input space.

6.4.1.2. Iso-regularization descent

Revisiting Osborne's papers on the resolution of the LASSO problem [73], M. Loth proposed a new algorithm named the Iso-Regularization descent, to solve this problem [33]. This algorithm is currently the most efficient to be known; in particular, it is more efficient than the cyclic coordinate descent [70], and computes the regularization path of the LASSO as efficiently as the LARS. This algorithm is also able to solve other regularized problems, such as the grouped LASSO, and the elastic net problem. A complete presentation of this algorithm will appear in Manuel Loth's PhD dissertation, in the early 2011, and will be submitted to a journal.

6.4.2. Online Learning

6.4.2.1. Online Learning in Adversarial Lipschitz Environments

We consider [35] the problem of online learning in an adversarial environment when the reward functions chosen by the adversary are assumed to be Lipschitz. This setting extends previous works on linear and convex online learning. We provide a class of algorithms with cumulative regret upper bounded by $O(\sqrt{dT\log(\lambda)})$ where d is the dimension of the search space, T the time horizon, and λ the Lipschitz constant. Efficient numerical implementations using particle methods are discussed. Applications include online supervised learning problems for both full and partial (bandit) information settings, for a large class of non-linear regressors/classifiers, such as neural networks.

6.4.3. Functional regression

The work led by Hachem Kadri on functional regression has made much progress in 2010. In our work, “functional” regression means that we consider regression problems, and more generally supervised learning problems, in which observations, and response(s) are functions. The usual approach to this problem is to consider vectors instead of functions. A kernel approach for this purpose means functions acting on functions, that is, operators; moreover, to be valid, these operators should respect some properties. Exhibiting such operators that respect those properties is difficult, but a basic requirement if we want to use such an approach in any application. Different functional kernels have been exhibited, and different algorithms to solve the minimization problem have been proposed [28]. The multi-task setting has also been investigated, that is the setting is which more than one functional response has to be made. An application to speech inversion has been tackled. To compare our functional approach with the more traditional vector-based approach, we have recently written describing the resolution of such a speech inversion task, that exhibit state of the art performance [60].

6.4.4. Applications

6.4.4.1. Computer graphics

As a follow-up to the successful work performed in 2009 related to computer graphics, an experimental work has been performed to investigate the potential of the ECON algorithm at representing photometric solids [20], [45] with regards to a neural network approach.

6.4.4.2. Medical data analysis

Jérémie Mary made an informal collaboration with an INSERM lab (ERI-12) of the University of Amiens about picture analysis of some cells in order to detect the effect in the cellular mobility of the muscles (based on the vinculine and actine observation). Another work was conducted with the Lab of psychology of the University of Lille 3 on the analysis of human gesture by Geoffrey Megardon under supervision of J. Mary.

6.4.4.3. Thermal model optimisation

Some software has been developed by Jérémie Mary and Antoine Chamot to optimize thermal models which can be used with the software Energy Plus. It achieved a drop in error rate of 40% versus full human modelling, on the data collected by Effigenie (see also Section 7.1.4).

6.5. Unsupervised learning

Participants: Jérémie Mary, Daniil Ryabko.

6.5.1. Clustering time series data

6.5.1.1. Theory

We have applied [39] the approach to statistical analysis of time series, described in section 6.3.2, to the problem of clustering time series samples. Thus, we have considered the problem of clustering for the case when each data point is a sample generated by a stationary ergodic process. We proposed a very natural asymptotic notion of consistency, and showed that simple consistent algorithms exist, under most general non-parametric assumptions. The notion of consistency is as follows: two samples should be put into the same cluster if and only if they were generated by the same distribution. With this notion of consistency, clustering generalizes such classical statistical problems as homogeneity testing and process classification. We showed that, for the case of a known number of clusters, consistency can be achieved under the only assumption that the joint distribution of the data is stationary ergodic (no parametric or Markovian assumptions, no assumptions of independence, neither between nor within the samples). If the number of clusters is unknown, consistency can be achieved under appropriate assumptions on the mixing rates of the processes. (again, no parametric or independence assumptions). In both cases we give examples of simple (at most quadratic in each argument) algorithms which are consistent.

6.5.1.2. Implementation

An implementation of the “magic distance” of Daniil Ryabko (the empirical estimate of distributional distance) has been made by Jérémie Mary. The clustering process works well on some artificial ergodic data, and we are looking for some real ergodic data (ideally non-Markovian and continuous) to test the developed algorithm on.

6.6. Sensors Networks: Tracking, Localization and Communication

Participants: Emmanuel Delande, Emmanuel Duflos, Philippe Vanheeghe, Nicolas Viandier, Nouha Jaoua.

6.6.1. 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 ([47]). This work is done in collaboration with Thales Communications. Beside this main point we have finalized the optimization of the detection step of a radar ([82]).

6.6.2. Sequential learning of sensors localization: application to civil engineering

This work is done in collaboration with Prof Carl Haas of the University of Waterloo (Canada) and is a continuation of previous research : how can we automatically track the building materials on a construction site? This is a real problem because a lot of time (hence of money) is lost to find these materials that have often been moved away. The ability to detect dislocations automatically for tens of thousands of items can ultimately improve project performance significantly. The proposed solution is to equip each piece with a RFID tag and each people working on the construction site with a RFID receiver, a GPS for the localization, and a transmitter. We have obtained a PICS (International Project for Scientific Cooperation) from the CNRS in 2008 for 3 years to work on this. During the two past years, we have developed a belief functions based method to track the materials. In 2010 we have focused on dislocation detection performances by tuning the basic belief masses. ROC curves obtained on experimental data show a real improvement for the low false alarm rate ([21]).

6.6.3. 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 ([38], [37], [43], [44]).

6.6.4. Internet of Things

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 neighborhood. 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 submitted to the ICASSP Conference at the day of the writing of this report.

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

7.1.1. Addressing Business

Participants: Sertan Girgin, Philippe Preux.

A contracted, and funded, collaboration has begun this Fall between SequeL and a company (SME) named “Addressing Business” located in Roubaix. The aim of this contract is to design and implement a software prototype. For confidentiality and competitiveness reasons, we will not detail this collaboration further than mentioning that it is related to recommendation systems, however in a far from academic setting (data are large and complex, there is a sequential aspect in the data, ...).

7.1.2. Orange Labs

Participants: Sertan Girgin, Jérémie Mary, Philippe Preux, Christophe Salperwyck.

Two contracts are living with Orange Labs.

First, there is an on-going CIFRE contract, funding a PhD on sequential supervised learning (Ch. Salperwyck, 2009-2012).

Second, there is a one-year CRE (externalized research contract) that has been negotiated and signed in the late 2010. This contract deals with the study of sequential machine learning under constraints, with application to the ad selection problem.

7.1.3. Inquest

The work of last year of Jeremie Mary about adaptive quizz has been polished and is now used in production.

7.1.4. Effigenie

Participant: Jérémie Mary.

Effigenie is a future start up (should be created in January 2011), which plans to sell a solution to optimize thermal control of a building with respect to their planned utilization and the weather. Some preliminary tests on real building during winter 2010 allow us to expect around 20% of energetic consumption. The approach used needs a good thermal modelling of the building. This is a problem as having such a modelling is time consuming and needs a human expert. So Jeremie Mary conducted a work based on the optimization day after day of a rough model. Using this kind of approach we were able to optimize models and to reach an error of prediction 40% lower than a hand-made model. The model is quite hard to optimize because there is more than one hundred variables and each evaluation needs several seconds. For next year we plan to have again better optimization making some more local adjustments and to test the new model in winter. Another promising development is to use RL methods in order to control the building without having to build a model. Such a solution would be a fast and very low-cost method to have better efficiency over thermal control.

7.1.5. Unbalance Corporation

Participant: Rémi Coulom.

Unbalance Corporation is a Japanese company who bought a license of Crazy Stone (see section 5.1.1) in 2010. Unbalance is specialized in selling games.

7.1.6. Pôle de Compétitivité "Industries du commerce"

Participants: Sertan Girgin, Jérémie Mary, Philippe Preux.

SEQUEL is taking part in a project named "Ubiquitous Virtual Seller" (VVU) of the Pôle de Compétitivité "Industrie du Commerce" (PICOM). See more details in Section 8.1.1.

8. Other Grants and Activities

8.1. Regional Initiatives

8.1.1. Pôle de Compétitivité "Industries du commerce"

Participants: Sertan Girgin, Jérémie Mary, Philippe Preux.

SEQUEL is taking part in a project named "Ubiquitous Virtual Seller" (VVU) of the Pôle de Compétitivité "Industrie du Commerce" (PICOM). This project has begun on Sep. 1st, 2009 and will last 2 years. The VVU project involves three computer science laboratories (Laboratoire d'Informatique Fondamentale de Lille, INRIA Lille Nord Europe, and Mines de Douai), a marketing school (Skema-Lille), and private companies (Becquet, Oxylane, France Telecom, Artificial Solutions, Nextstage). In this project, we are funded by the Région-Nord Pas de Calais, and the FEDER; funding is mostly for a post-doc over a period of 18 months. The work involves a close collaboration with other computer science teams at the Laboratoire d'Informatique Fondamentale de Lille, and the Mines de Douai. See sec. 6.2.3.1 for more details about 2010 activities on this contract.

8.2. National Initiatives

8.2.1. ANR Lampada

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

This was the first year of this ANR project. In participating to this project, our goal is at least two-fold: getting acquainted with the management of large datasets, both at the fundamental level, and at the technical level; getting acquainted with working or more complex data than mere real vectors, or real functions, such as qualitative data, and data such as trees, or graphs. The underlying assumption is also that data comes as a flow. Noteworthy, our work on ad selection with Orange labs led us to handle very large web log files, which goes along the same line of work of very large streams of data. The aforementioned contract with Addressing Business (see Sec. 7.1.1) is also perfectly compatible with this policy.

Olivier Nicol is beginning his PhD in this context. Furthermore, Ph. Preux is co-advising Gabriel Arnold-Dulac who begins his PhD in the Malire team of the LIP'6, under P. Gallinari and L. Denoyer's supervision (both participate to the Lampada project).

8.2.2. DGA/Thales

Participants: Emmanuel Delande, Emmanuel Duflos.

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.

8.2.3. ARC MaBI

Participants: Emmanuel Duflos, Rémi Munos, Daniil Ryabko, Philippe Vanheeghe.

MaBI stands for "Machine Learning for Brain Computer Interfaces"; the scientific coordinator of this ARC project is Stéphane Canu. Members of SequeL involved: Rémi Munos, Daniil Ryabko, Philippe Vanheeghe, and Emmanuel Duflos. The ARC MaBI started in 2010 for 2 years.

8.2.4. ANR EXPLO-RA

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

EXPLO-RA, acronym for EXPLOration - EXPLOitation for efficient Resource Allocation with Applications to optimization, control, learning, and games, is an ongoing, 3 years ANR-funded project which started in 2009. This is a collaboration between 2 INRIA project teams (SequeL and TAO), HEC Paris (GREGHEC), Les Ponts (CERTIS), Paris 5 (CRIP5), and the Université Paris Dauphine (LAMSADE). Rémi Munos is the coordinator.

8.2.5. ANR CO-ADAPT

Brain computer co-adaptation for better interfaces project, which started in the end of 2009 (for 4 years). This is in collaboration with the 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). Rémi Munos is the SequeL coordinator.

8.2.6. LITIS : Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes

Participants: Emmanuel Duflos, Hachem Kadri.

Emmanuel Duflos and Hachem Kadri are collaborating with Pr. Stéphane Canu on Functional RKHS.

8.3. European Initiatives

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

8.3.1. PASCAL2 Network of excellence

In 2009, SEQUEL has joined the Pascal-2 European network of excellence dedicated to machine learning. SEQUEL has created a new node of this NoE in collaboration with the EPI Mostrare, and Stéphane Canu's group in Rouen. R. Munos is the head of this node.

8.3.2. PASCAL2 Pump Priming Programme

Participants: Mohammad Ghavamzadeh, Rémi Munos.

Sparse Reinforcement Learning in High Dimensions, with Shie Mannor (Technion, Israel), Mohammad Ghavamzadeh and Rémi Munos. This is a 2 year project that started in November 2009.

8.4. International Initiatives

8.4.1. INRIA Associate Team with University of Alberta, Canada

The title of the joint team is Decision-making under Uncertainty with Applications to Reinforcement Learning, Control, and Games. The coordinators from INRIA side are Mohammad Ghavamzadeh and Rémi Munos. The coordinator from University of Alberta side is Csaba Szepesvári. Other collaborators in University of Alberta are Prof. Richard Sutton and Amir-massoud Farahmand.

8.4.2. Programme Interdisciplinaire de Coopération Scientifique

A “Programme Interdisciplinaire de Coopération Scientifique” (PICS) is running over the period 2008–2010 which concerns Ph. Vanheeghe, and E. Duflos, in relation with the Centre for Pavement and Transportation Technology (CPATT), headed by prof. Carl Haas at the University of Waterloo, Canada.

The optimal use of the data provided by the sensors must necessarily lie within a dynamic process suitable to control the acquisition of information. This project proposes to define principles and methods for the management of multisensor systems in the frame of civil engineering. This work, requires the development of specific methodological tools. These tools will be tested on a real civil engineering application, the characterization of new materials for highway pavement. Multisensor management being integrated in this Canadian, very ambitious, civil engineering project. The Canadian team will carry out the instrumentation and the validation, whereas the definition of the tools and method will be carried out in tight partnership and controlled by the French team.

8.4.3. Haifa, Israel

Mohammad Ghavamzadeh collaborates with Yaakov Engel on the topic of *regularized reinforcement learning* over the last four years. This year, we have two journal papers on this topic that will be submitted soon [52], [53].

Mohammad Ghavamzadeh has been also working with Prof. Shie Mannor, on the topic of *Bayesian reinforcement learning* for the last five years, on the topic of *regularized reinforcement learning* for the last three years, and on the topic of *reinforcement learning in high dimensions* in the last year. On the first topic, we have a journal paper (survey) in preparation [57] this year. On the second topic, we have two journal papers in preparation [52], [53] this year. Finally, on the third topic, we are Co-PI’s of a *PASCAL2 pump-priming program*.

8.4.4. University of Waterloo, Waterloo, Ontario, Canada

Prof. Pascal Poupart and Mohammad Ghavamzadeh have been collaborating on the topic of *Bayesian reinforcement learning* in the last four years. This year, we have a journal paper in preparation [57] on this topic.

Emmanuel Duflos and Philippe Vanheeghe have visited twice Prof. Carl Hass at the University of Waterloo (Canada) from September 5th to September 10th and from December 4th to December 10th.

8.4.5. National Taiwan Normal University

Rémi Coulom has been working with Shih-Chieh Huang, a PhD student from the Department of Computer Science and Information Engineering, National Taiwan Normal University. Shih-Chieh Huang’s main advisor is Professor Shun-Shii Lin, and he is co-advised by Rémi Coulom. In 2010 they worked on simulation balancing [26] and time management [27]. Shih-Chieh Huang also won the gold medal of the Computer Olympiad (see award section).

8.4.6. Warsaw University

Rémi Coulom has been working with Łukasz Lew is a PhD student at the Warsaw University, in Poland. Łukasz’s main advisor is Professor Krzysztof Diks. Łukasz visited Sequel for one month in April. During his visit, he wrote a paper with Rémi Coulom about Monte-Carlo search of combinatorial games [32].

9. Dissemination

9.1. Animation of the scientific community

- Awards:
  - Sébastien Bubeck's Ph.D. thesis, entitled "Bandits Games and Clustering Foundations" [11] has been awarded a Gilles Kahn 2010 prize, ranking second. This is a prize awarded by Specif to the best Ph.D. theses in Computer Science, in France (patronized by the Academy of Science). The thesis supervisor was Rémi Munos.
  - Gold medal at the Computer Olympiad: The Go-program developed by Shih-Chieh Huang under the supervision of Rémi Coulom, Erica, won the gold medal in the 2010 Computer Olympiad in Kanazawa, Japan. The Computer Olympiad is regarded as the most important international computer-Go tournament. All the major commercial and academic programs participated.
  - 2008 ICGA Journal Award. This award is given every year by the ICGA (International Computer Games Association) for the best paper of a first-time author, published in the ICGA Journal. The award for year 2008 was given in 2010 to a paper by Rémi Coulom [68] (that was actually published in the December 2007 issue).
  - Victor Gabillon, Jeremie Mary and Philippe Preux have received a best paper award at the conference "Extraction et Gestion des Connaissances" for their work [23] on ad selection problem on Internet portals.
- Alessandro Lazaric gives a tutorial at AAMAS 2010: Reinforcement Learning and Beyond
- participation to the program committees of international conferences:
  - R. Coulom: CG'2010: International Conference on Computers and Games, Kanazawa, Japan
  - R. Coulom: TAAI'2010: Technologies and Applications of Artificial Intelligence, Taipei, Taiwan
  - E. Duflos and Ph. Vanheeghe: Fusion'2010 workshop on the Theory of Belief Function (Brest, April 1-2, 2010),
  - E. Duflos: workshop on the Theory of Belief Function (Brest, April 1-2, 2010),
  - M. Ghavamzadeh: European Conference on Machine Learning (ECML 2010), International Conference on Machine Learning (ICML 2010)
  - R. Munos: Area chair for NIPS 2010
  - Ph. Preux: ICML 2010, CAP 2010, EGC 2010.
  - D. Ryabko: UAI 2010.
- GDR ISIS : following a request of Jean-Yves Tourneret (in charge of theme A in the GDR ISIS), Emmanuel Duflos organized a one-day workshop in June 2010 : *Advances in Signal Processing and Data Fusion for Localisation*. 68 researchers attended to this workshop. This workshop was co-organized with the GT2 of the GDR Robotics (with Roland Chapuis : lASMEA).
- Invited talks:
  - R. Munos is Invited speaker in (in addition to the conferences) Journées MAS 2010 (Modélisation Aléatoire et Statistique), SMILE 2010 (Statistical Machine Learning in Paris), NIPS 2010 Workshop (Learning and planning from batch time series data).
  - O. Maillard is Invited speaker in GDR ISIS "Apprentissage et parcimonie".

  - M. Ghavamzadeh gives an invited talk at University of Alberta - AI Seminar, Host : Prof. Csaba Szepesvári (2010).
  - D. Ryabko gives an invited talk at GdR ISIS workshop “Journée spéciale Stéganographie et Stéganalyse”.
  - Jeremie Mary has given an invited talk at Bilab (ENST Paris) and SMILE 2010 (Statistical Machine Learning in Paris).

- international journal and conference reviewing activities (in addition to the conferences in which we belong to the PC):
  - E. Duflos: IEEE Transaction on Signal Processing, International Journal of Approximate Reasonning, Information Fusion.
  - M. Ghavamzadeh: Annual Conference on Neural Information Processing Systems (NIPS 2010), Neurocomputing, Machine Learning Journal (MLJ), Journal of Machine Learning Research (JMLR), Journal of Artificial Intelligence Research (JAIR),
  - R. Munos: Machine Learning, IEEE Transactions on Automatic Control, Revue d’Intelligence Artificielle, ALT 2010, ICML 2010.
  - Ph. Preux: NIPS 2010, STACS 2010, ECML 2010
  - D. Ryabko: IEEE Trans. Inf. Th., NIPS 2010.

- Evaluation activities, expertise
  - E. Duflos and Ph. Vanheeghe have reviewed proposals for the ANR programs
  - Ph. Preux has reviewed project proposals for the ECOS-Sud program (France), ANRT (France), and the IWT (Belgium)
  - Ph. Preux is expert for the AERES. He expertized masters in computer science, as well as a laboratory.
  - Ph. Preux is member of the “Gilles Kahn”/Specif jury which elects an outstanding PhD in computer science of the year.
  - Ph. Preux served as president of the committee of an INRIA-Université de Lille 3 chair in computer science/statistical learning (Spring 2010), the committee to recruit an assistant professor in statistical learning (Fall 2010), and member of the committee to recruit 4 assistant professors in computer science at the Université d’Artois (Spring 2010).
  - R. Munos has been a member of the following committees:
    - Membre jury de recrutement DR2 INRIA, 2010
    - Membre jury de recrutement CR2 INRIA Lille, 2010
    - Scientific organizer of the INRIA evaluation seminar, theme Optimisation, apprentissage et méthodes statistiques, March 2010.
    - Membre du Comité d’animation du domaine thématique INRIA, Mathématiques appliquées, calcul et simulation.
    - Délégué (titre Commission d’Evaluation) pour la création des projets INRIA: CLASSIC, SIERRA
    - Recommendation letter for promotion of a Senior Lecturer position (kept anonymous) in Technion, Israel
    - Referee in PASCAL2 Programme Pump Priming programme

- participation to PhD and HDR jurys:

- R. Munos was a Rapporteur for PhD thesis of: Jia Yuan Yu (Mc Gill University), Olga Kozlova (Université Paris 6), Christophe Thiery (INRIA Nancy), Sarah Philippi (Telecom ParisTech).
- R. Munos was a Member of HDR Committee: Jean-Yves Audibert (Ecole Nationale des Ponts et Chaussées).
- Emmanuel Duflos was *rapporteur* for the for PhD thesis of Gregory Mallet (INSA Rouen) and Adrien Chen (ENAC)

9.2. Teaching

We list the classes that are related to the research activities in SEQUEL that were going on in 2010.

- Rémi Munos teaches a class in reinforcement learning in the M2 “Mathematics-Vision-Learning” (MVA) at the ENS-Cachan.
- Philippe Preux teaches:
  - in the Master 2 MIASHS (Maths and computer science for humanities): 2 data mining classes (data mining, and web mining)
  - in a Master 2 of psychology (neuro-cognitive processes), and in a Master 1 of psychology (analysis of behavior): machine learning, reinforcement learning, models of adaptive behavior, models of learning in animal (incl. human beings).
- Jérémie Mary is head of the speciality “Informatique et Documents” of the Master MIASHS.
- Jérémie Mary has followed 4 M2 students and 2 M1 in their internship in external societies.

Otherwise, each of the 4 professors and assistant professors of the SEQUEL team teaches 192 hours per year. Taught classes include machine learning, data mining, and signal processing classes.

10. Bibliography

Major publications by the team in recent years

[1] J.-Y. AUDIBERT, R. MUNOS, C. SZEPESVÁRI. *Exploration-exploitation trade-off using variance estimates in multi-armed bandits*, in "Theoretical Computer Science", 2009, vol. 410, p. 1876-1902.

[2] S. BUBECK, R. MUNOS, G. STOLTZ, C. SZEPESVÁRI. *Online Optimization of X-armed Bandits*, in "Proceedings of Advances in Neural Information Processing Systems", MIT Press, 2008, vol. 22.

[3] F. CARON, M. DAVY, A. DOUCET, E. DUFLOS, P. VANHEEGHE. *Bayesian Inference for Linear Dynamic Models With Dirichlet Process Mixtures*, in "IEEE Transactions on Signal Processing", January 2008, vol. 56, n^{o} 1, p. 71–84.

[4] F. CARON, M. DAVY, E. DUFLOS, P. VANHEEGHE. *Particle Filtering for Multisensor Data Fusion with Switching Observation Models. Application to Land Vehicle Positioning*, in "IEEE Transactions on Signal Processing", June 2006, vol. 55, n^{o} 6, p. 2703–2719.

[5] R. COULOM. *Computing Elo Ratings of Move Patterns in the Game of Go*, in "International Computer Games Association Journal", 2007.

[6] 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.

[7] A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Finite-Sample Analysis of LSTD*, in "Proceedings of the Twenty-Seventh International Conference on Machine Learning", 2010, p. 615-622.

[8] O.-A. MAILLARD, R. MUNOS. *Compressed Least Squares Regression*, in "Proceedings of Advances in Neural Information Processing Systems", 2009.

[9] D. RYABKO, M. HUTTER. *On the Possibility of Learning in Reactive Environments with Arbitrary Dependence*, in "Theoretical Computer Science", 2008, vol. 405, n^{o} 3, p. 274–284.

[10] D. RYABKO. *On Finding Predictors for Arbitrary Families of Processes.*, in "Journal of Machine Learning Research", 2010, vol. 11, p. 581–602.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] S. BUBECK. *Jeux de Bandits et Fondations du Clustering*, Université Lille 1, Lille, France, June 2010.

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

Articles in International Peer-Reviewed Journal

[13] D. MAZOUNI, J. HARMAND, A. RAPAPORT, H. HAMMOURI. *Optimal time switching control for multi-reaction batch process*, in "Optimal Control Application and Methods", 2010, vol. 31, p. 289-301.

[14] D. RYABKO. *Discrimination between B-processes is impossible*, in "Journal of Theoretical Probability", 2010, vol. 23, n^{o} 2, p. 565–575.

[15] D. RYABKO. *On Finding Predictors for Arbitrary Families of Processes.*, in "Journal of Machine Learning Research", 2010, vol. 11, p. 581–602.

[16] D. RYABKO, B. RYABKO. *Nonparametric Statistical Inference for Ergodic Processes*, in "IEEE Transactions on Information Theory", 2010, vol. 56, n^{o} 3, p. 1430–1435.

International Peer-Reviewed Conference/Proceedings

[17] J.-Y. AUDIBERT, S. BUBECK, R. MUNOS. *Best Arm Identification in Multi-Armed Bandits*, in "Conference on Learning Theory", 2010.

[18] S. BUBECK, R. MUNOS. *Open Loop Optimistic Planning*, in "Conference on Learning Theory", 2010.

[19] R. COULOM, P. ROLET, N. SOKOLOVSKA, O. TEYTAUD. *Handling Expensive Optimization with Large Noise*, in "Foundations of Genetic Algorithms XI", 2010, http://hal.archives-ouvertes.fr/hal-00517157/en/.

[20] S. Delepoulle, F. Rousselle, C. Renaud, P. Preux. *A comparison of two machine learning approaches for photometric solids compression*, in "Proc. of the 13th Int'l Conf. on Computer Graphics and Artificial Intelligence (3IA)", May 2010, p. 132–142.

[21] 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.

[22] A. M. Farahmand, R. Munos, Cs. Szepesvári. *Error Propagation for Approximate Policy and Value Iteration*, in "Advances in Neural Information Processing Systems", 2010.

[23] V. Gabillon, J. Mary, P. Preux. *Affichage de publicités sur des portails web*, in "Proc. Extraction et Gestion des Connaissances (EGC)", January 2010, This paper received a best paper award.

[24] 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", 2010.

[25] S. Girgin, J. Mary, P. Preux, O. Nicol. *Advertising Campaigns Management: Should We Be Greedy?*, in "Proc. IEEE International Conference on Data Mining", 2010.

[26] S.-C. Huang, R. Coulom, S.-S. Lin. *Monte-Carlo Simulation Balancing in Practice*, in "International Conference on Computers and Games", Kanazawa, Japan, September 2010.

[27] S.-C. Huang, R. Coulom, S.-S. Lin. *Time Management for Monte-Carlo Tree Search Applied to the Game of Go*, in "International Workshop on Computer Games", Taiwan, 2010.

[28] 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.

[29] A. Lazaric, M. Ghavamzadeh. *Bayesian Multi-Task Reinforcement Learning*, in "Proceedings of the Twenty-Seventh International Conference on Machine Learning", 2010, p. 599-606.

[30] A. Lazaric, M. Ghavamzadeh, R. Munos. *Analysis of a Classification-based Policy Iteration Algorithm*, in "Proceedings of the Twenty-Seventh International Conference on Machine Learning", 2010, p. 607-614.

[31] A. Lazaric, M. Ghavamzadeh, R. Munos. *Finite-Sample Analysis of LSTD*, in "Proceedings of the Twenty-Seventh International Conference on Machine Learning", 2010, p. 615-622.

[32] L. Lew, R. Coulom. *Simulation-based Search of Combinatorial Games*, in "ICML 2010 Workshop on Machine Learning and Games", Haifa, Israel, 2010.

[33] M. Loth, P. Preux. *The Iso-lambda Descent Algorithm for the LASSO*, in "Proc. of 17th International Conference on Neural Information Processing (ICONIP)", LNCS, Springer, November 2010.

[34] O.-A. Maillard, R. Munos, A. Lazaric, M. Ghavamzadeh. *Finite-Sample Analysis of Bellman Residual Minimization*, in "Proceedings of the Second Asian Conference on Machine Learning", 2010.

[35] O.-A. MAILLARD, R. MUNOS. *Online Learning in Adversarial Lipschitz Environments*, in "European Conference on Machine Learning", 2010.

[36] O.-A. MAILLARD, R. MUNOS. *Scrambled Objects for Least-Squares Regression*, in "Advances in Neural Information Processing Systems", 2010.

[37] 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).

[38] 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).

[39] D. RYABKO. *Clustering processes*, in "Proc. the 27th International Conference on Machine Learning (ICML 2010)", Haifa, Israel, 2010, p. 919–926.

[40] D. RYABKO. *Sequence prediction in realizable and non-realizable cases*, in "Proc. the 23rd Conference on Learning Theory (COLT 2010)", Haifa, Israel, 2010, p. 119–131.

[41] D. RYABKO. *Testing composite hypotheses about discrete-valued stationary processes*, in "Proc. IEEE Information Theory Workshop (ITW'10)", Cairo, Egypt, IEEE, 2010, p. 291–295.

[42] D. RYABKO. *Uniform hypothesis testing for ergodic time series distributions*, in "IEEE R8 International Conference on Computational Technologies in Electrical and Electronics Engineering (SIBIRCON 2010)", Irkutsk, Russia, IEEE, 2010, p. 23–27.

[43] 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).

[44] 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).

Scientific Books (or Scientific Book chapters)

[45] S. DELEPOULLE, F. ROUSSELLE, C. RENAUD, P. PREUX. *A Comparison of Two Machine Learning Approaches for Photometric Solids Compression*, in "Intelligent Computer Graphics", D. PLEMENOS, G. MIAOULIS (editors), Studies in Computational Intelligence, Springer, 2010, vol. 321, p. 145–164.

[46] R. MUNOS. *Approximate Dynamic Programming*, in "Markov Decision Processes in Artificial Intelligence", O. SIGAUD, O. BUFFET (editors), ISTE Ltd and John Wiley & Sons Inc, 2010, chap. 3, p. 67–98.

Research Reports

[47] 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, http://hal.inria.fr/inria-00501502/.

[48] A. LAZARIC, M. GHAVAMZADEH. *Bayesian Multi-Task Reinforcement Learning*, INRIA, 2010, n^{o} inria-00475214.

[49] A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Analysis of a Classification-based Policy Iteration Algorithm*, INRIA, 2010, n^o inria-00482065.

[50] A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Finite-Sample Analysis of LSTD*, INRIA, 2010, n^o inria-00482189.

Scientific Popularization

[51] R. COULOM. *Jeux et sports : le problème des classements*, in "Pour la Science", July 2010, n^o 393.

Other Publications

[52] A. M. FARAHMAND, M. GHAVAMZADEH, CS. SZEPESVÁRI, S. MANNOR. *L2-Regularized Fitted-Q Iteration Algorithm*, 2010, in preparation.

[53] A. M. FARAHMAND, M. GHAVAMZADEH, CS. SZEPESVÁRI, S. MANNOR. *L2-Regularized Policy Iteration*, 2010, in preparation.

[54] V. GABILLON, A. LAZARIC, M. GHAVAMZADEH. *Rollout Allocation Strategies for Classification-based Policy Iteration*, in "Workshop on Reinforcement Learning and Search in Very Large Spaces at the Twenty-Seventh International Conference on Machine Learning", 2010.

[55] M. GHAVAMZADEH, Y. ENGEL. *Bayesian Actor-Critic Algorithms*, 2010.

[56] M. GHAVAMZADEH, Y. ENGEL. *Bayesian Policy Gradient Algorithms*, 2010.

[57] M. GHAVAMZADEH, S. MANNOR, P. POUPART. *Bayesian Reinforcement Learning: A Survey*, 2010, in preparation.

[58] S. GIRGIN, J. MARY, P. PREUX, O. NICOL. *Planning-based Approach for Optimizing the Display of Online Advertising Campaigns*, December 2010, submitted to the NIPS workshop “Machine Learning in Online Advertising”.

[59] H. KADRI, E. DUFLOS, P. PREUX, S. CANU, M. DAVY. *Function-Valued Reproducing Kernel Hilbert Spaces and Applications*, December 2010, NIPS workshop “Tensors, Kernels, and Machine Learning”.

[60] H. KADRI, E. DUFLOS, P. PREUX. *Learning Vocal Tract Variables With Multi-Task Kernels*, 2010, submitted.

[61] A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Analysis of a Classification-based Policy Iteration Algorithm*, 2010, submitted to the Journal of Machine Learning Research.

[62] A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Finite-Sample Analysis of Least-Squares Policy Iteration*, 2010, submitted to the Journal of Machine Learning Research.

[63] D. RYABKO. *On sequential prediction for arbitrary classes of discrete-valued processes*, in "European Meeting of Statisticians", Piraeus, Greece, 2010.

References in notes

[64] 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.

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

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

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

[68] R. COULOM. *Computing Elo Ratings of Move Patterns in the Game of Go*, in "ICGA Journal", December 2007, vol. 30, n^{o} 4, p. 198–208.

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

[70] J. FRIEDMAN, T. HASTIE, H. HÖFLING, R. TIBSHIRANI. *Pathwise coordinate optimization*, in "Annals of Applied Statistics", 2007, vol. 1, n^{o} 2, p. 302–332.

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

[72] D. ORNSTEIN, B. WEISS. *How Sampling Reveals a Process*, in "Annals of Probability", 1990, vol. 18, n^{o} 3, p. 905–930.

[73] M. OSBORNE, B. PRESNELL, B. TURLACH. *A new approach to variable selection in least squares problems*, in "Journal of Numerical Analysis", 2000, vol. 20, n^{o} 3, p. 389–403.

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

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

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

[77] 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.

[78] 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.

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

[80] 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.

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

[82] M. G. DE VILMORIN, E. DUFLOS, P. VANHEEGHE. *Radar Optimal Times Detection Allocation in Multitarget Environment*, in "IEEE Journal Systems", 2009, vol. 3, n^{o} 2, p. 210-220, Répertorié dans ISI Web of Knowledge sans facteur d'Impact.