INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team SequeL

Sequential Learning

Lille - Nord Europe

Theme : Optimization, Learning and Statistical Methods

Table of contents

1. Team 1
2. Overall Objectives 2
2.1. Introduction 2
2.2. Highlight of the year 3
3. Scientific Foundations 3
3.1. Introduction 3
3.2. Decision under uncertainty 3
3.2.1. Markov decision processes 3
3.2.2. Bandits 5
3.3. Statistical learning 6
3.3.1. Kernel methods for non parametric function approximation 6
3.3.2. Non parametric Bayesian models 7
4. Application Domains 7
4.1. Outline 7
4.2. Adaptive control 7
4.3. Signal analysis and processing 8
4.4. Functional prediction 9
4.5. Neurosciences 9
5. Software 9
6. New Results 10
6.1. Introduction 10
6.2. Decision under uncertainty 10
6.2.1. Reinforcement learning and approximate dynamic programming 10
6.2.1.1. Approximate Policy Iteration without Value Function Representation 10
6.2.1.2. Natural actor-critic 10
6.2.1.3. Bayesian Multi-Task Reinforcement Learning 10
6.2.1.4. Regularized Fitted Q-iteration for Planning in Continuous-Space MDPs 11
6.2.1.5. Function approximation and representation learning 11
6.2.2. Sensitivity analysis in HMMs 11
6.2.3. Exploration vs. exploitation 11
6.2.3.1. Pure exploration in multi-armed bandits 11
6.2.3.2. Hybrid Stochastic-Adversarial On-line Learning 11
6.2.3.3. Minimax Policies for Adversarial and Stochastic Bandits 12
6.2.4. Applications 12
6.2.4.1. The games of Go and Havannah 12
6.2.4.2. The Ubiquitous Virtual Seller 12
6.2.4.3. Ad selection on web portals 12
6.2.4.4. Games that adapt to player skill 13
6.3. Foundations of machine learning 13
6.3.1. Sequence prediction in the most general form. 13
6.3.2. Statistical inference 13
6.3.3. Steganography 14
6.4. Supervised learning 14
6.4.1. Multi representation 14
6.4.2. New algorithms to induce classifiers, and regressors 14
6.4.2.1. Compressed Least Squares Regression 14
6.4.2.2. Non parametric function approximation: the Equi-Correlation Network algorithm 15
6.4.3. Functional regression 15
6.4.4. Applications 15

6.5. Unsupervised learning 15
6.6. Sensors Networks: Tracking, Localization and Communication 16
6.6.1. The sensor management problem 16
6.6.2. Sequential learning of sensors localization: application to civil engineering 16
6.6.3. Accurate Localization using Satellites in Urban Canyons 16
6.6.4. Internet of Things 17
7. Contracts and Grants with Industry **17**
7.1.1. France Telecom/Orange Labs 17
7.1.2. Inquest 18
7.1.3. ETO 18
7.1.4. Vekia Innovation 18
8. Other Grants and Activities **18**
8.1. Regional activities 18
8.2. National activities 19
8.2.1. DGA / Thalès 19
8.2.2. ANR EXPLORA 19
8.2.3. ANR Kernsig 20
8.2.4. ANR Lampada 20
8.2.5. ANR Co-Adapt 20
8.3. International activities 21
8.3.1. PASCAL2 Network of excellence 21
8.3.2. PASCAL2 Pump-Priming Project 21
8.3.3. University of Alberta, Canada 21
8.3.4. Russia 21
8.3.5. MPI Tübinghen 22
8.3.6. COLT workshop 22
8.3.7. Special session at COGIS'2009 22
8.3.8. Programme Interdisciplinaire de Coopération Scientifique 22
8.4. Visits and invitations 22
9. Dissemination **22**
9.1. Scientific community animation 22
9.2. Teaching 24
10. Bibliography **24**

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 *(UMR 8021 of the École Centrale of Lille and the University of Lille 1).*

1. Team

Research Scientist

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

Faculty Member

Philippe Preux [Team leader, Professor, Université de Lille, secondment at the INRIA until Aug. 31st, 2009, HdR]
Emmanuel Daucé [Assistant Professor, École Centrale de Marseille, partial secondment in SEQUEL until Aug. 31st, 2009]
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]

Technical Staff

Tony Ducrocq [Assistant Engineer, until Sep. 30th, 2009]

PhD Student

Sébastien Bubeck [ENS Grant, since Oct., 2007]
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, 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]
Manuel Loth [INRIA-Région Nord-pas-de-calais Grant, since Oct., 2006]
Odalric-Ambrym Maillard [ENS Grant, since Oct., 2008]
Christophe Salperwyck [CIFRE with France Telecom Grant, since Dec., 2009]
Nicolas Viandier [INRETS, since Oct., 2007]

Post-Doctoral Fellow

Sertan Girgin [Région Nord-Pas de Calais, begins on Sep. 1st, 2009]
Alessandro Lazaric [INRIA until Aug. 31st, then ANR]
Hachem Kadri [CNRS until Aug. 31st, then Région Nord-Pas de Calais]

Administrative Assistant

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

Other

Boris Iolis [Master 1 internship, Université Libre de Bruxelles, Oct. to Dec. 2009]
Victor Gabillon [Master 2 internship, Telecom Sud-Paris and ENS-Cachan, Apr. to Sep. 2009]
Victor Marsault [Lience 3 internship, ENS-Cachan, Jun. to Jul. 2009]

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* in the following[1]. 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 gather information collected so far in some compact representation and combine it to 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 fulfill a certain goal) that impact the environment 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).

We exemplify these various problems:

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 [80]). We do not assume the availability of a model of the system to be controlled.

In all these cases, we 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 than 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 bounding of the performance,
- the implementation of usable algorithms,

all that being understood in a *sequential* framework.

[1] we might also have called them "learning machines", since that's what these agents are here.

2.2. Highlight of the year

In 2009, we would like to highlight the fact that we have obtained several contracts with private societies, either directly or via a "Pôle de compétitivité", as well as academic contracts (ANR, Europe). This is really a strong increase of the contracted part of the activities of SEQUEL, and this increase corresponds to a desire to investigate applications lying at the edge of our research activities, and also to help promote the machine learning technology in solving real problems. Other contracts should be negotiated in 2010 with private societies in particular.

3. Scientific Foundations

3.1. Introduction

SEQUEL is primarily grounded on two domains:

- the problem of decision under uncertainty,
- statistical learning which provides 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 [74].

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

In the MDP $(\mathcal{X}, \mathcal{A}, P, r)$, each initial state x_0 and action sequence $a_0, a_1, ...$ gives rise to a sequence of states $x_1, x_2, ...$, satisfying $\mathbb{P}(x_{t+1} = x'|x_t = x, a_t = a) = p(x'|x, a)$, and rewards[2] $r_1, r_2, ...$ 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.

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

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

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

where $\mathbb{E}$ is the expectation operator and $\gamma \in (0,1)$ is the discount factor. This value function V^π gives an evaluation of the performance of a given policy π. Other functionals of the sequence of future rewards may be considered, such as the undiscounted reward (see the stochastic shortest path problems [64]) and average reward settings. Note also that, here, we considered the problem of maximizing a reward functional, but a formulation in terms of minimizing some cost or risk functional would be equivalent.

In order to maximize a given functional in a sequential framework, one usually applies Dynamic Programming (DP) [62], which introduces the optimal value function $V^*(x)$, defined as the optimal expected sum of rewards when the agent starts from a state x. We have $V^*(x) = \sup_\pi V^\pi(x)$. Now, let us give two definitions about policies:

- We say that a policy π is optimal, if it attains the optimal values $V^*(x)$ for any state $x \in \mathcal{X}$, *i.e.*, if $V^\pi(x) = V^*(x)$ for all $x \in \mathcal{X}$. Under mild conditions, deterministic stationary optimal policies exist [63]. Such an optimal policy is written π^*.
- We say that a (deterministic stationary) policy π is greedy with respect to (w.r.t.) some function V (defined on $\mathcal{X}$) if, for all $x \in \mathcal{X}$,

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

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

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

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

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

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

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

In short, we would like to mention that most of the reinforcement learning methods developed so far are built on one (or both) of the two following approaches ([84]):

- 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.

Recently, Auer *et al.* [60] 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 learning

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

Machine learning refers to a system capable of the autonomous acquisition and integration of knowledge. This capacity to learn from experience, analytical observation, and other means, results in a system that can continuously self-improve and thereby offer increased efficiency and effectiveness. (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. (source: http://www.cs.wisc.edu/~hzhang/glossary.html)

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 [68]. We briefly present the most important for us to date, namely, kernel-based non parametric function approximation, and non parametric Bayesian models.

3.3.1. 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.

3.3.2. 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 lets us handle non linear, as well as non Gaussian problems. In their standard form, they require the formulation of densities of probability in their 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.

To overcome this shortcoming, non parametric methods provide another approach to this problem. In particular, mixtures of Dirichlet processes [66] provide a very powerful formalism.

Mixtures of Dirichlet Processes are an extension of finite mixture models. Given a mixture density $f(\mathbf{x}|\theta)$, and $\mathsf{G}(d\theta) = \sum_{k=1}^{\infty} \omega_k \delta_{U_k}(d\theta)$, a Dirichlet process [3], then we define a mixture of Dirichlet processes as:

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

A mixture of Dirichlet processes is fully parameterized by the mixture density, as well as the parameters of G, that is G_0 and α.

The class of densities that may be written as a mixture of Dirichlet processes is very wide, so that these are really fit to very large amount 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.

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,
- neurosciences.

4.2. Adaptive control

Adaptive control is an important application of the research being done in SEQUEL. Reinforcement learning precisely aims at controling the behavior of systems and may be used in situations with more or less information available. Of course, the more information, the better, in which case methods of (approximate) dynamic programming may be used [73]. 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.

[3] A Dirichlet process is a random distribution almost surely discrete, where the centroids U_k are distributed along a *base distribution* $\mathsf{G}_0(\cdot)$, and where weights follow a certain *stick breaking* law with parameter α [79].

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

- game softwares: in the 1990's, RL has been the basis of a very successful Backgammon program, TD-Gammon [83] that learned to play at an expert level by basically playing a very large amount of games against itself;

  Today, various games are studied with RL techniques.
- many optimization problems that are closely related to operation research, but taking into account the uncertainty, and the stochasticity of the environment: see the job-shop scheduling, or the cellular phone frequency allocation problems, resource allocation in general [73]
- 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 orignal 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. Neurosciences

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

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

5. Software

5.1. Software

5.1.1. Crazy Stone

Participant: Rémi Coulom [correspondent].

Crazy Stone, is a top-level Go-playing program that has been developped by Rémi Coulom since 2005. Crazy Stone won several major international Go tournaments in the past. Because of the media impact of those victories, some software companies showed interest in buying licences of Crazy Stone. So, in 2009, Crazy Stone was registered with the APP (Agence pour la Protection des Programmes). No licence has been sold so far. Crazy Stone is not available publicly.

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. clustering,
5. signal processing.

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, Philippe Preux, Daniil Ryabko.

6.2.1. Reinforcement learning and approximate dynamic programming

6.2.1.1. Approximate Policy Iteration without Value Function Representation

There is a recent interest on approximate policy iteration algorithms in which the action-value function is not approximated over the entire state-action space [71], [67]. The main idea is to remove the policy evaluation and cast the policy improvement as a classification problem. The training set of this classification problem is generated by rollout estimates of the action-value function on a finite number of states. In [58], we present a novel loss function by weighting the number of classification errors with the actual regret associated to each error, *i.e.*, the difference between the action-values of the greedy action and the action chosen by the rollout policy, and provide convergence bounds for the resulting approximate policy iteration algorithm.

6.2.1.2. Natural actor-critic

In [13], [49], we present four new reinforcement learning algorithms based on actor–critic, function approximation, and natural gradient ideas, and we provide their convergence proofs. Actor–critic reinforcement learning methods [61], [81] are online approximations to policy iteration in which the value-function parameters are estimated using temporal difference learning [82] and the policy parameters are updated by stochastic gradient descent. Methods based on policy gradients in this way are of special interest because of their compatibility with function approximation methods, which are needed to handle large or infinite state spaces. The use of temporal difference learning in this way is of special interest because in many applications it dramatically reduces the variance of the gradient estimates. The use of the natural gradient is of interest because it can produce better conditioned parameterizations and has been shown to further reduce variance in some cases. Our results extend prior two-timescale convergence results for actor–critic methods by [69] (also [70]) by using temporal difference learning in the actor and by incorporating natural gradients. Our results extend prior empirical studies of natural actor–critic methods by [72] by providing the first convergence proofs and the first fully incremental algorithms. We present empirical results verifying the convergence of our algorithms.

6.2.1.3. Bayesian Multi-Task Reinforcement Learning

In [54], we consider the problem of multi-task reinforcement learning, where a learner is provided with a set of tasks, for which only a small number of samples can be generated for any given policy. As the number of samples may not be enough to learn an accurate evaluation of the policy, it would be necessary to identify classes of tasks with similar structure and to learn them jointly. We consider the case where the tasks share structure in their value functions, and model this by assuming that the value functions are all sampled from a common prior. We adopt the Gaussian process temporal-difference [65] value function model and use a hierarchical Bayesian approach to model the distribution over the value functions. In this paper, we study two cases, where all the value functions belong to the same class and where they belong to an undefined number of classes. For each case, we present a hierarchical Bayesian model, and derive inference algorithms for:

1. joint learning of the value functions, and
2. efficient transfer of the information gained in (i) to assist learning the value function of a newly observed task.

6.2.1.4. Regularized Fitted Q-iteration for Planning in Continuous-Space MDPs

Reinforcement learning with linear and non-linear function approximation has been studied extensively in the last decade. However, as opposed to other fields of machine learning such as supervised learning, the effect of finite sample has not been thoroughly addressed within the reinforcement learning framework. In this work [27], we propose to use L^2 regularization to control the complexity of the value function in reinforcement learning and planning problems. We consider the regularized fitted Q-iteration algorithm and provide generalization bounds that account for small sample sizes. We use a realistic visual-servoing problem to illustrate the benefits of using the regularization procedure.

6.2.1.5. Function approximation and representation learning

As a follow-up to the 2008 work on the issue of the representation of states, we have worked further on feature discovery in the context of sequential decision problems. Based on our 2008 work on feature discovery in the context of reinforcement learning to discover a good (if not the best) representation of states, we have studied the use of non parametric function approximation in the context of approximate dynamic programming. The striking difference with the usual approach is that we use a non parametric function approximator to represent the value function, instead of a parametric one. See [33], [59].

6.2.2. Sensitivity analysis in HMMs

We considered a sensitivity analysis in Hidden Markov Models with continuous state and observation spaces. We proposed an Infinitesimal Perturbation Analysis (IPA) on the filtering distribution with respect to some parameters of the model. We described a methodology for using any algorithm that estimates the filtering density, such as Sequential Monte Carlo methods, to design an algorithm that estimates its gradient. The resulting IPA estimator is proven to be asymptotically unbiased, consistent and has computational complexity linear in the number of particles. We considered an application of this analysis to the problem of identifying unknown parameters of the model given a sequence of observations. We derived an IPA estimator for the gradient of the log-likelihood, which may be used in a gradient method for the purpose of likelihood maximization. See [23].

6.2.3. Exploration vs. exploitation

6.2.3.1. Pure exploration in multi-armed bandits

We considered the framework of stochastic multi-armed bandit problems where a forecaster is assessed in terms of its simple regret, a regret notion that captures the fact that exploration is only constrained by the number of available rounds (not necessarily known in advance), in contrast to the case when the cumulative regret is considered and when exploitation needs to be performed at the same time. This performance criterion is suited to situations when the cost of pulling an arm is expressed in terms of resources rather than rewards. We discussed the links between the simple and the cumulative regret. Our main result is that the required exploration–exploitation trade-offs are qualitatively different, in view of a general lower bound on the simple regret in terms of the cumulative regret. See [22].

6.2.3.2. Hybrid Stochastic-Adversarial On-line Learning

Most of the research in online learning focused either on the problem of adversarial classification (*i.e.*, both inputs and labels are arbitrarily chosen by an adversary) or on the traditional supervised learning problem in which samples are independently generated from a fixed probability distribution. Nonetheless, in a number of domains the relationship between inputs and labels may be adversarial, whereas input instances are generated according to a constant distribution. We introduced a hybrid stochastic-adversarial classification problem, in which inputs are stochastic, while labels are adversarial. We proposed an online learning algorithm for its solution, and analyzed its performance. In particular, we showed that, given a hypothesis space $\mathcal{H}$ with finite VC dimension, it is possible to incrementally build a suitable finite set of hypotheses that can be used as input

for an exponentially weighted forecaster achieving a cumulative regret of order $O(\sqrt{nVC(\mathcal{H})\log n})$ with overwhelming probability. We also discussed extensions to multi-label classification, learning from experts and bandit settings with stochastic side information, and application to games. See [29].

6.2.3.3. Minimax Policies for Adversarial and Stochastic Bandits

This work deals with four classical prediction games, namely full information, bandit and label efficient (full information or bandit) games as well as three different notions of regret: pseudo-regret, expected regret and tracking the best expert regret. We introduced a new forecaster, INF (Implicitly Normalized Forecaster), for which we proposed a unified analysis of its pseudo-regret in the four games. With well-chosen parameters INF defines a new forecaster, for which we were able to remove the extraneous logarithmic factor in the pseudo-regret bounds for bandit games, and thus fill in a long open gap in the characterization of the minimax rate for the pseudo-regret in the bandit game. We also consider the stochastic bandit game, and prove that an appropriate modification of the upper confidence bound policy UCB achieves the distribution-free optimal rate while still having a distribution-dependent rate logarithmic in the number of plays. See [21].

6.2.4. Applications

6.2.4.1. The games of Go and Havannah

After the 2006 major breakthrough in go realized by Rémi Coulom's Crazy Stone program, the latter has evolved further.

Rémi Coulom's main research topic in 2009 was automatic parameter optimization from noisy observations, applied to his Go-playing program Crazy Stone. The performance of most game-playing programs depends on several parameters. In order to get optimal performance, it is necessary to tune these parameters carefully. This is a very challenging problem, because the number of parameters is very high, and the effect of parameters is measured with very noisy observations. Crazy Stone has thousands of parameters, and observations are binary outcomes of games (win or loss). Early results of using local quadratic regression were presented at the University of Electro-Communications (Japan) in January [53].

From June 15th to July 31st, Rémi Coulom supervised Victor Marsault, a first-year student from ENS Cachan. The topic of this internship was the application of Monte-Carlo tree search to the game of Havannah. Like the game of Go, the game of Havannah is a challenging application domain, where the strongest human players still easily outperform the best computer algorithms. Although they did not manage to reach top human level, they investigated original Monte-Carlo tree search ideas and produced a decent artificial player [57].

6.2.4.2. The Ubiquitous Virtual Seller

This 18 months project aims at studying the design, and implementation, of virtual agents on selling Internet portals. The goal is that this agent will be able to recognize the visitors of the portal, either as regular visitors, or new visitors, and help them, provide advices, develop a selling strategy, ...

Having begun in Sep. 2009, for the moment, the work has mostly been a research of relevant work in the literature, as well as getting acquainted with the other members of the project, in particular the marketing aspects of the project, as well as the private companies expectations.

See also the contract section (Sec. 8.1.1) of the report for specific details about the contract itself.

6.2.4.3. Ad selection on web portals

In 2009, we have begun a work on the selection of displayed ads on web pages, under contract with France Telecom/Orange Labs.

Of course, this problem has already received a lot of attention by major actors of the Internet. However, publicly available works on this problem have never tackled the real problem, with the specific real constraints. In particular, the finiteness of resources (in time, and in the number of ads to display) is not tackled, and asymptotically optimal algorithms are studied. But asymptotic results are not those that are sought, and the performance of these asymptotically optimal algorithms used under finite constraints of time and resource are typically bad. Indeed, our work has shown that handling this finiteness is necessary to obtain good strategies of ad display. We have modeled the problem as the resolution of a linear program, in which some crucial

quantities have to be learned from data. So, we end-up proposing an approach which mixes bandits to estimate these data on which linear programming is applied. Furthermore, this process has to be iterated to handle the fact that ad campaigns have a limited extent in time, new ad campaigns are created, and the discrepancy between the actual visitors of the website, and those that were planned. This work has been accepted, and will be published in 2010.

See also the contract section (Sec. 7.1.1) of the report for specific details about the contract itself.

6.2.4.4. Games that adapt to player skill

It has always been a challenge for computer scientists to try to defeat human experts at any game; among many other games, draughts, Othello, chess, and currently Go have challenged the community. However, for "standard" humans, some programs are desperately too strong; we have been arguing for years that methods of adaptive control may be useful to design new games which, instead of aiming at defeating any human being, at the cost of boredom, adapts to the strength of the human player.

We have had the opportunity to work concretely on this idea in collaboration with the InQuest company located in Villeneuve d'Ascq. We tackled the problem of asking questions to people, according to their skill: the difficulty of a question depends on people, on their age, their culture, ... Jérémie Mary designed a Bayesian approach to assess the difficulty of questions related to a given human being, and ask his/her questions of appropriate difficulty that he/she has a reasonable probability to answer correctly. We have also worked on the inclusion of new, non rated, questions to the catalog of available questions (approx 10^4 different questions, among which a dozen is asked to a given human being: so, the skill of a given player has to be assessed very quickly with the first of these 12 questions).

See also the contract section (Sec. 7.1.2) of the report for specific details about the contract itself.

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, the probabilities of the next outcome of the observed sequence of data. In the most general formulation of the problem, we assume that the data is generated by a stochastic process that belongs to a certain known class of processes $\mathcal{C}$, and the problem is to construct a predictor that works for any (a priory unknown) process coming from $\mathcal{C}$.

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 $\mathcal{C}$.

Our recent breakthrough [38] in solving this general problem is in showing that, when such a predictor can be constructed, in can be constructed as a Bayesian predictor whose prior is concentrated on a countable subset of $\mathcal{C}$.

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 [17]. Previous results on these problems concerned only much more restricted settings (e.g. i.i.d. data). In addition, we have shown [37] 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.

Our most recent results [39], [47] in this direction provide 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.

In addition, the paper on rank tests that was mentioned in the previous report as accepted, has now been published [18].

6.3.3. Steganography

The goal of steganography is to transfer hidden information in seemingly innocuous messages (called "covertexts"), in the presence of an observer who is trying to find out whether hidden information is being transmitted. The innocuous messages may be, for example, photographic images, or human-written notes. They are assumed to be generated by an oracle, whose exact probabilistic characteristics are unknown to the communicating parties. For the case when this probabilistic process is i.i.d. or has a finite memory (which is a natural and a standard assumption) we have constructed [16] a universal (any distribution conforming to the above assumption) perfectly secure (no detection is possible) asymptotically optimal (in terms of the amount of transmitted secret information) and simple (in terms of computation) steganographic system. On the other hand, we have shown [40] that there exist such complicated sources of covertexts, that any stegosystem that meets the perfect security condition must itself have an exponential (in the size of the message) Kolmogorov complexity.

6.4. Supervised learning

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

6.4.1. Multi representation

This work considers the problem of semi-supervised multi-view classification, where each view corresponds to a Reproducing Kernel Hilbert Space. We propose an algorithm based on co-regularization methods with extra penalty terms reflecting smoothness and general agreement properties. We first provide explicit tight control on the Rademacher (L1) complexity of the corresponding class of learners for arbitrary many views, then give the asymptotic behavior of the bounds when the co-regularization term increases, making explicit the relation between consistency of the views and reduction of the search space. Since many views involve many parameters, we third provide a parameter selection procedure, based on the stability approach with clustering and localization arguments. To this aim, we give an explicit bound on the variance (L2-diameter) of the class of functions. See [32].

6.4.2. New algorithms to induce classifiers, and regressors

6.4.2.1. Compressed Least Squares Regression

We considered the problem of learning, from K input data, a regression function in a function space of high dimension N using projections onto a random subspace of lower dimension M. From any linear approximation algorithm using empirical risk minimization (possibly penalized), we provided bounds on the excess risk of the estimate computed in the projected subspace (compressed domain) in terms of the excess risk of the estimate built in the high-dimensional space (initial domain). We applied the analysis to the ordinary Least-Squares regression and showed that by choosing $M = O(\sqrt{K})$, the estimation error (for the quadratic loss) of the "Compressed Least Squares Regression" is $O(1/\sqrt{K})$ up to logarithmic factors. See [31]

6.4.2.2. Non parametric function approximation: the Equi-Correlation Network algorithm

We have designed a new algorithm, named the Equi-Correlation Network (ECON), to perform supervised classification, and regression. ECON is a kernelized LARS-like algorithm, by which we mean that ECON uses an l_1 regularization to produce sparse estimators. ECON efficiently rides the regularization path to obtain the estimator associated to any value of the constant of regularization, and ECON represents the data by way of features induced by a feature function. The originality of ECON is that it automatically tunes the parameters of the features while riding the regularization path. So, ECON has the unique ability to produce optimally tuned features for each value of the constant of regularization. Experimentally, we have obtained remarkable performance of ECON on standard benchmark datasets in regression and supervised classification.

We have also used ECON to tackle the problem of representing photometric solids in computer graphics. See the application section below, as well as [30], [55], [56].

6.4.3. Functional regression

Functional regression deals with the setting in which the attributes of data, as well as their associated label, are functions. Traditionally, functional regression considers discretized attributes, and apply the classical regression techniques (see [75] for instance).

We have tackled this problem considering functions as functions, whereas the traditional approach consists in dealing with discretized functions, thus vectors. We have developed a RKHS approach for it, kernels being now operators mapping a function to a function. We have demonstrated the basic theorems (basic properties of such functional kernel, existence of such kernel, representer theorem) on which a sound functional RKHS approach can be built. We have also exhibited a functional kernel, and provided preliminary experimental results.

A preliminary version of this work is available as an INRIA research report [50], and a further worked version is under submission for publication.

This work takes place under the ANR Kernsig project (see Sec. 8.2.3).

6.4.4. Applications

To create realistic images, photometric solids are used that represent how the energy of a wave of light of a certain wavelength is reflected in any direction. This data is available for a huge amount of materials. This whole data is traditionally represented in mere tables, which are thus huge, and interpolation is used to estimate the reflected energy for directions which are not available.

In collaboration with a team working in computer graphics, we have studied the use of the machine learning technology to represent these data in a much more compact way. Mere back propagated neural networks have first been used, and then ECON has been used. The expected results have been obtained: having much more compact representation of these photometric solids, while keeping the same quality of rendered images, which is the ultimate goal in computer graphics. See [26], [48], [30].

This collaboration has shown us that the field of computer graphics is a rich field of applications of machine learning technology, yet to be exploited. This collaboration is going on.

6.5. Unsupervised learning

Participant: Sébastien Bubeck.

6.5.1. Nearest Neighbor Clustering

Clustering is often formulated as a discrete optimization problem. The objective is to find, among all partitions of the data set, the best one according to some quality measure. However, in the statistical setting where we assume that the finite data set has been sampled from some underlying space, the goal is not to find the best partition of the given sample, but to approximate the true partition of the underlying space. We argue that the discrete optimization approach usually does not achieve this goal, and instead can lead to inconsistency. We construct examples which provably have this behavior. As in the case of supervised learning, the cure is

to restrict the size of the function classes under consideration. For appropriate "small" function classes we can prove very general consistency theorems for clustering optimization schemes. As one particular algorithm for clustering with a restricted function space we introduce "nearest neighbor clustering". Similar to the k-nearest neighbor classifier in supervised learning, this algorithm can be seen as a general baseline algorithm to minimize arbitrary clustering objective functions. We prove that it is statistically consistent for all commonly used clustering objective functions. See [14].

6.6. Sensors Networks: Tracking, Localization and Communication

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

6.6.1. The sensor management problem

This class of applications took a new turn this year with the thesis of Emmanuel Delande, supervised by Emmanuel Duflos and Philippe Vanheeghe, in collaboration with Thales Communication. The aim of this work is to manage a set of sensors to track vehicles or groups of people in land applications. The dynamic of each target is controlled by a velocity vector field defined over the area of interest. Such a modelling allows the use of particle filters to track the targets. In real application, the high dimension state is however an obstacle to an accurate estimation of the targets parameters since it is well known that the estimation error increase with the number of targets. That is the reason why our work focuses today on random sets based estimation filter and more precisely on the PHD filter. The sensors management modelling work is still under progresses. It is clear today that such an optimization problem is very close to the reinforcement learning problem, and current research focuses on how to model a sensor management problem as a reinforcement learning optimization problem.

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). This collaboration is related to a problem occurring in civil engineering: how can we automatically locate 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 then learn sequentially the position of the pieces using the incoming detection information sent automatically by the transmitter to a central processor when the workforces walk near these pieces and detect them. RFID systems and localization systems as GPS allow to treat such a problem in the more general context of randomly distributed communication nodes localization. We have obtained a PICS (International Project for Scientific Cooperation) from the CNRS in 2008 for 3 years to work on the specific problems arising when huge amount of sensors are used in civil engineering application. This activity deals with both sensor management and signal analysis. The work achived in 2009 [36], is a continuation of previous research, in which we tackled the location estimation problem by fusing the data from a simulation model.

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. The American system GPS (Global Positioning System) is the only fully operational solution for the moment. This monopole reduces the possibilities of measurement redundancy and diversity, thus limits the reachable performances. Unfortunately most all these transport applications are mainly used in dense urban environments, highly constraining for signal propagation. 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. The worst case of reception is the alternate path. In this case the LOS signal from a satellite cannot reach the antenna and receiver tracks only reflected signals. Such phenomena make the pseudorange error distribution becomes a non-white and non-Gaussian distribution ([41]). As a consequence, the classical localization methods like Extended Kalman Filter (EKF), assuming that state and observation noises are white and Gaussian, are not efficient anymore and make positioning error more important. Thus, to enhance the localization accuracy in case of alternate path reception, the filtering part of the receiver (after correlators) must be improved. Furthermore, in order to limit costs, we have chosen to work only with GNSS signals. In a goal of enhanced position accuracy, we propose a new statistical filtering method based on a better definition (and use) of the observation noise for each satellite signal. Moreover, in a very constraint environment (like urban environment or canyon) where reflected signals are frequent, the pseudorange noise density takes an unknown form. Consequently, to estimate such unknown distribution form, a mixture model can be a suitable solution. In previous works, a first approach was studied based on Jump Markov System (JMS) algorithm ([34], [35]). JMS switches between several observation noise models according to the estimated reception state of each satellite. The law parameters which describe the observation noise of each available pseudorange are next use in a particle filter to estimate the position. JMS showed its performances in terms of accuracy and continuity of service. However some drawbacks of JMS show that the density modeling can be improved. Indeed, the proposed JMS version is strongly related to the study of the close propagation environment and consequently a punctual reflection cannot be detected by the Markov Chain. This can create false detection and missed detection and consequently the chosen model by the algorithm should be wrong. Moreover, we need to allocate T seconds for initialization. Another default is that in the context of dynamic models, the assumption of stationary is wrong. And finally the number of Gaussian components is limited in the Gaussian mixture and consequently the estimated model does not represent the true distribution but an approximation of it. That is why we opted for the use of Dirichlet Process Mixture (DPM). We have shown that the DPM, which is an infinite mixture model, is more efficient than a finite mixture model to estimate sequentially an unknown distribution. The first step of this algorithm is the sampling of hyperparameters which are a couple of parameters: the mean and the standard deviation of each Gaussian law which composes the infinite mixture. This sampling is performed by a Gibbs sampler. Then the hyperparameters are used as inputs of a Rao-Blackwellised particle filter (RBPF) to compute the position. This approach outperforms standard models commonly used to represent observation noise distributions, *i.e.* white and Gaussian noise. The efficiency of this approach has been demonstrated by applying a validation step involving real GPS data. These data have been acquired in an urban environment and in a public transport context.

6.6.4. Internet of Things

A new thesis, supervised by Emmanuel Duflos and Philippe Vanheeghe, has started in september within the frame of the 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 neighbourhood. The noise appearing in such applications can be considered as α-stable which means that moment higher than 2 does not exist. New estimation algorithms must therefore be developped to estimate sequentially the parameters of the probability density function which may vary according to time as well as the data received by each node of the network.

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

7.1.1. France Telecom/Orange Labs

We have had a 10 months externalized research contract (CRE) with France Telecom in 2009 on the problem of selecting ads to display on web pages. During his internship in the EPI, V. Gabillon has made his master thesis on this subject; J. Mary and Ph. Preux have dedicated a significant part of their time to work on this contract. Based on the very interesting results that were obtained during this CRE, a new contract is under negotiation for 2010 as a follow-up to this first work.

More technical details are available in section 6.2.4.3 of this document.

7.1.2. Inquest

We have had a collaboration with inQuest[4], a society working on casual games, located in Villeneuve d'Ascq.

These new methods should be used in production very shortly and a contract is under negotiation. See sec. 7.1.2 for more about this contract.

7.1.3. ETO

A collaboration has been initiated with the private society ETO, located in Roubaix. ETO manages large databases of customers, and fidelity programs, for a few dozens very well-known commercial brands (both national, and international brands). ETO also proposes human support in order to follow and exploit these data: identification of high value customers, building of ads campaigns, ...Their software is called X27 and requires a lot of human intervention to tailor it to their new customers. ETO wishes to render automatic a maximum of steps in order to reduce the costs and widespread their solution. That is the so-called A-27 project.

One of the problems is to cluster customers in a sequential framework. The sequence of data is the list of the visits to a shop. In an ideal world we would model the behavior of any customer. This objective is impossible to reach because we do not have enough data on each customer. So, we wish to classify the customers in groups based on their habits. However, customers' habits change over time (they are single, then in couple, then have babies, ... they live in a flat, then in a house, ... they earn more and more, ...). One challenge here is to study and detect the switch of customers from one cluster to an other along time.

An other goal is to evaluate by simulation the impact of a new ad campaign. It would be used to help marketing to optimize its decisions.

Jérémie Mary conducted some preliminary work on their data, showing these objectives may be reached. This led to the project Simul-Market between ETO, Vekia and INRIA (involving Jeremie Mary and Philippe Preux). Then, this project has been proposed, assessed, and labelled by the PICOM and A-27 will be funded by the Région Nord-Pas de Calais and the FEDER (basically, this will fund 2 years post-doc funding, and 1 year of engineer, over 2010 and 2011).

7.1.4. Vekia Innovation

Vekia Innovation is the name of the spin-off two of us (P-A. Coquelin and M. Davy) created in 2007, originally under the name "Predict & Control".

We have done a work on the clustering of temporal series, with an application to the clustering of calls to call centers. A software toolbox has been implemented to demonstrate various algorithms.

This collaboration was funded by OSEO.

8. Other Grants and Activities

8.1. Regional activities

8.1.1. Pôle de Compétitivité "Industries du commerce"

Participants: Sertan Girgin, Jérémie Mary, Philippe Preux.

[4] http://www.inquest.fr.

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 (ESC-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. 6.2.4.2 for more details about 2009 activities on this contract.

8.2. National activities

8.2.1. DGA / Thalès

Participants: Emmanuel Duflos, Philippe Vanheeghe, Emmanuel Delande.

The work on sensor management went on this year, focusing on three main points:

- Modelling the dynamic of the moving object for land applications
- Modelling the tracking problem in the Random Finite Sets framework
- Modelling the optimization problem as it may usually be done in reinforcement learning

8.2.2. ANR EXPLORA

Participants: Sébastien Bubeck, Alexandra Carpentier, Emmanuel Delande, Victor Gabillon, Mohammad Ghavamzadeh, Jean-François Hren, Alessandro Lazaric, Manuel Loth, Jérémie Mary, Odalric-Ambrym Maillard, Rémi Munos, Philippe Preux, Daniil Ryabko.

Rémi Munos is the coordinator of the **ANR EXPLO-RA**[5] (EXPLOration - EXPLOitation for efficient Resource Allocation. Applications to optimization, control, learning, and games) 3 years project which started in 2009. This is a collaboration between 2 INRIA team project (SEQUEL and TAO), HEC Paris (GREGHEC), Les Ponts (CERTIS), Paris 5 (CRIP5), and the Université Paris Dauphine (LAMSADE).

This project deals with the question of how to make the best possible use of available resources in order to optimize the performance of some decision-making task. In the case of simulated scenarios, the term resource refers to a piece of computational effort (for example CPU time, memory) devoted to the realization of some computation. Nonetheless, we will also consider the case of real-world scenarios where the term resource denotes some effort (real-world experiment) that has a real, *e.g.* financial, cost. Making a good use of the available resources means designing an exploration strategy that would allocate the resources in a clever way such as to maximize (among the space of possible exploration strategies) the performance of the resulting task. Potential applications are numerous and may be found in domains where a one-shot decision or a sequence of decisions has to be made, such as in optimization, control, learning, and games.

For that purpose we will consider several ways of combining algorithms which perform a good job in balancing resources between exploitation (making the best decision based on our current, but possibly imperfect, knowledge) and exploration (decisions that may appear sub-optimal but which may yield additional information about the unknown parameters, and, as a result, could improve the relevance of future decisions). These exploration/exploitation algorithms, also called bandit algorithms, or regret-minimization algorithms, will be the building blocks of our methods. They will be combined either in a hierarchical way, or as a population, either in collaborative or adversary working mode.

[5] http://sites.google.com/site/anrexplora/.

A motivating example concerns min-max tree search in large scale games. The goal here is to explore the tree to find the best move for the next play, given a limited amount of simulation resources (*e.g.*, CPU time). Here, resource allocation means an exploration strategy that selects which branch one should explore deeper at each time step; the aim being that at the end of the available resources, the collected information allows making the best decision (or an almost optimal decision). Previous works in efficient tree exploration using hierarchical bandits for the game of go have shown very promising results (such as the MoGo program [Gelly et al., 2006] currently among the world best computer-go programs), which have motivated our research for extending both the theoretical analysis of the underlying ideas and their scope to a wide range of applications.

We expect to develop new simulation techniques based on a clever use of available computational resources, in order to solve large scale optimization and decision making problems previously considered unsolvable. See sec. 6.2.3 for details about 2009 scientific activities.

8.2.3. ANR Kernsig

Participants: Emmanuel Duflos, Hachem Kadri, Philippe Preux.

The ANR Kernsig project began in 2007 and it is headed by Prof. S. Canu with the INSA-Rouen. It deals with the study of kernel methods for signal processing.

See the section 6.4.3 for scientific details of 2009 activities.

8.2.4. ANR Lampada

Participants: Mohammad Ghavamzadeh, Jérémie Mary, Philippe Preux.

The ANR Lampada project has been submitted, and approved in 2009, and will officially begin in 2010. Lampada means "Learning Algorithms, Models an sPArse representations for structured DAta"[6]. This project involves approximately 30 people from Paris (LIP'6, P. Gallinari's group), Marseille (LIF, F. Denis' group), Saint-Étienne (LHC, M. Sebban's group), the Mostrare and SEQUEL EPIs. M. Tommasi from Mostrare is the head of this ANR.

Lampada is a fundamental research project on machine learning and structured data. It focuses on scalling learning algorithms to handle large sets of complex data. The main challenges are:

1. high dimension learning problems,
2. large sets of data and
3. dynamics of data.

Complex data we consider are evolving and composed of parts among which there are some relations. Representations of these data embed both structure and content information and are typically large sequences, trees and graphs. The main application domains are web2, social networks and biological data.

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

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

SEQUEL is particularly concerned with the learning of the representation of data in high dimensional spaces, in particular the work on feature extraction, and non parametric supervised learning algorithms.

8.2.5. ANR Co-Adapt

Participant: Rémi Munos.

[6] project website: http://lampada.gforge.inria.fr/.

This ANR project has been submitted, and approved in 2009. Rémi Munos is the SEQUEL coordinator of the **ANR CO-ADAPT** (Brain computer co-adaptation for better interfaces) project which starts 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).

8.2.5.1. Workshop "Localisation Précise pour les Transports Terrestres"

Emmanuel Duflos was the main organizer, in collaboration with the LEOST, Heudiasyc and LCPC french laboratories, of a workshop on precised localization for land transportations. This workshop was held in Paris on June, $16^{\text{th}}th$. There were more than 30 attendees. A CD-ROM has been edited (INRETS publisher) on which papers are in english to ease the spreading of the presented works.

8.3. International activities

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 Project

Pump-Priming is a program organized by the PASCAL2 network of excellence. The goal of this program is to provide support for collaborative research on novel topics that are not yet sufficiently mature to attract mainstream funding. Rémi Munos and Mohammad Ghavamzadeh, along with Shie Mannor, an associate professor at the department of electrical engineering at Technion, Haifa, Israel, submitted a proposal on "Sparse Reinforcement Learning in High Dimensions" to this program. Our proposal was accepted for funding in September 2009. This is a 2 year project that starts in November 2009.

The main objective of this project is to find appropriate representations for value function approximation in high-dimensional spaces, and to use them to develop efficient reinforcement learning algorithms. By appropriate we mean representations that facilitate fast and robust learning, and by efficient we mean algorithms whose sample and computational complexities do not grow too rapidly with the dimension of the observations. We further intend to provide theoretical analysis for these algorithms as we believe that such results will help us refine the performance of such algorithms. We intend to empirically evaluate the performance of the developed algorithms in real-world applications such as a complex network management domain and a dogfight flight simulator.

This is a fundamental research project that would also help us to establish a collaboration with a very strong research group at Technion in Israel.

8.3.3. University of Alberta, Canada

We have continued our collaboration with the University of Alberta in Canada:

- with Prof. Csaba Szepesvári and Amir massoud Farahmand at the University of Alberta, Canada, on the topic of *regularities in sequential decision making problems*. We have published two conference papers [28], [27] and had two workshop papers accepted on this topic this year.
- with Prof. Richard Sutton from the University of Alberta, Canada, and Prof. Shalabh Bhatnagar from the Indian Institute of Science, Bangalore, India, on the topic of *actor-critic algorithms*, on which we have published a journal paper [13] and a technical report [49] this year.

8.3.4. Russia

D. Ryabko obtained an INRIA grant in the "collaboration avec la Russie" framework, for collaboration on steganography and statistics with Institute of Computational Technologies Siberian Branch of Russian Academy of Science, which funds two mutual visits. As a part of this funding scheme, D. Ryabko is also going to make a visit to Laboratoire J-V. Poncelet, Moscow, and give a talk on sequence prediction and statistics of processes there.

8.3.5. MPI Tübinghen

Sébastien Bubeck collaborates with U. von Luxburg on clustering.

8.3.6. COLT workshop

A 1 day "On-line Learning with Limited Feedback" (PASCAL2 sponsored event) has been organized by Alessandro Lazaric, Rémi Munos, Daniil Ryabko, Sébastien Bubeck, Odalric Maillard, Jean-Yves Audibert, Peter Auer, and Csaba Szepesvári.

8.3.7. Special session at COGIS'2009

Along with François Caron (EPI Alea, Bordeaux), E. Duflos organized a session on multi-target tracking at the conference COGIS'2009, held in Paris, Nov. 16-18th, 2009.

8.3.8. 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. Visits and invitations

- E. Duflos and Ph. Vanheeghe visit Carl Haas, U. Waterloo, Ontario, Canada, to work further in the frame of their joint PICS (November 28th to December 5th)
- Daniil Ryabko visits the J-V. Poncelet laboratory in Moscow.
- Daniil Ryabko visits Petri Myllymaki at the University of Helsinki, Finland.
- Rémi Munos and Rémi Coulom were invited to the Japanese-French conference, Tokyo, Jan. 2009

9. Dissemination

9.1. Scientific community animation

- A. Lazaric presented a tutorial on "Transfer Learning in Reinforcement Learning Domains" at both conferences AAMAS'2009, and ECML'2009.
- participation to the program committees of international conferences:
  - R. Coulom: "Advances in Computer Games 12"
  - E. Duflos: workshop on the Theory of Belief Function (Brest, April 1-2, 2010), Fusion 2009, Gretsi 2009
  - M. Ghavamzadeh: International Conference on Machine Learning (ICML 2009), Annual Conference on Neural Information Processing Systems (NIPS 2009)
  - R. Munos: NIPS 2009, ADPRL 2009, AISTATS 2009, ALT 2009, ICML 2009, JFPDA 2009

  - Ph. Preux: ECML 2009, IJCAI 2009, ADPRL 2009, EGC 2009 and 2010
  - D. Ryabko: "Learning from non-IID data" ECML 2009 workshop
  - E. Duflos: Fusion 2009

- 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: Machine Learning Journal (MLJ), Journal of Artificial Intelligence Research (JAIR), Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS),
  - J. Mary: Journal of Machine Learning Research (JMLR), EGC 2010
  - R. Munos: Annals of Telecommunications, Machine Learning, Mathematics of Operations Research, Revue d'Intelligence Artificielle,
  - Ph. Preux: Machine Learning Journal, IEEE Trans. on SMC-C, Algorithms
  - D. Ryabko: Uncertainty in Artificial Intelligence (UAI) 2009
- R. Munos and Ph. Preux have reviewed proposals in the ANR Blanc program (2009)
- R. Munos has reviewed proposals in the ANR Jeunes Chercheurs program (2009), and ANR COSINUS program
- R. Munos has been a member of the following committees:
  - INRIA Senior Researcher (DR 2) recruitment, 2009
  - INRIA Junior Researcher (CR 2) recruitment in Nancy, 2009
  - Scientific organizer of the INRIA evaluation seminar theme "Optimisation, apprentissage et méthodes statistiques", scheduled in March 2010.
  - animation comity of the INRIA theme "Mathématiques appliquées, calcul et simulation".
  - INRIA Evaluation committee
- participation to PhD jurys:
  - R. Munos was Rapporteur for PhD thesis of Matthieu Geist (Supélec Metz), and member of the PhD defense jury of Lucian Busoniu (Delft University, Nederland), Emmanuel Rachelson (University of Toulouse), Olivier Caelen (ULB, Belgium)
  - Ph. Preux was Rapporteur for the PhD thesis of A. Machado (Lip 6). He also serves as a member of the "Jury Gilles Kahn 2009" which aims at awarding the "best" computer science PhD dissertation of the year.
- expertise:
  - R. Munos was a referee in:
    * ERC starting grants evaluation Panel PE6 (Computer Science and Informatics)
    * Digiteo project in logiciel et systèmes complexes (Ile-de-France region)
    * Review for a CRC book on Reinforcement Learning
- invited talks:
  - R. Munos was invited speaker: seminars given at Delft University (Nederlands), Imperial College of London, Université Libre de Bruxelles, Atelier PIRSTEC (Lyon).
  - R. Coulom was invited speaker at the "Japanese-French Frontiers of Science Symposium JFFoS'2009)"
  - J. Mary was invited to give a Smile seminar at the "École des Mines de Paris"

- J. Mary gives a lecture as the Montebello high-school in Lille in a program aiming at drawing more students towards scientific studies.

9.2. Teaching

We list the classes that are related to the research activities in SEQUEL that happened in 2008.

- Rémi Munos teaches a class in reinforcement learning in the M2 “Mathematics-Vision-Learning” (MVA) at the ENS-Cachan.
- Philippe Preux teaches:
  - in the M2 MIASHS, 2 data mining classes
  - in the M2 of computer science at the University of Lille a class on reinforcement learning.
- Jérémie Mary is head of the speciality “Informatique et Documents” of the Master MIASHS.
- Jérémie Mary and Rémi Coulom are teaching data mining in master MIASHS at the University of Lille.

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. *Tuning Bandit Algorithms in Stochastic Environments*, in "Theoretical Computer Science", 2008, To appear.

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

[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", vol. 56, n^{o} 1, January 2008, 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", vol. 55, n^{o} 6, June 2006, 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] O.-A. MAILLARD, R. MUNOS. *Compressed Least Squares Regression*, in "Proceedings of Advances in Neural Information Processing Systems", 2009.

[7] R. MUNOS. *Performance Bounds in Lp norm for Approximate Value Iteration*, in "SIAM J. Control and Optimization", vol. 46, n^{o} 2, 2008, p. 541–561.

[8] R. MUNOS, C. SZEPESVÁRI. *Finite time bounds for sampling based fitted value iteration*, in "Journal of Machine Learning Research", 2007.

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

[10] D. RYABKO, M. HUTTER. *Predicting Non-Stationary Processes*, in "Applied Mathematics Letters", vol. 21, n^o 5, 2008, p. 477-482.

Year Publications

Doctoral Dissertations and Habilitation Theses

[11] F. NAHIMANA. *Impact des multitrajets sur les performances des systèmes de navigation par satellite : Contribution à l'amélioration de la précision de localisation par modémisation bayesienne*, Ecole Centrale de Lille, feb 2009, Ph. D. Thesis.

Articles in International Peer-Reviewed Journal

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

[13] S. BHATNAGAR, R. SUTTON, M. GHAVAMZADEH, M. LEE. *Natural Actor-Critic Algorithms*, in "Automatica", vol. 45, n^o 11, 2009, p. 2471-2482 CA IN .

[14] S. BUBECK, U. VON LUXBURG. *Nearest Neighbor Clustering: A Baseline Method for Consistent Clustering with Arbitrary Objective Functions*, in "Journal of Machine Learning Research", vol. 10, 2009, p. 657-698 DE .

[15] D. MAZOUNI, J. HARMAND, A. RAPAPORT, H. HAMMOURI. *Multi Reaction Batch Process and Optimal Time Switching Control*, in "Journal of Optimal Control Application and Methods", 2009.

[16] B. RYABKO, D. RYABKO. *Asymptotically Optimal Perfect Steganographic Systems*, in "Problems of Information Transmission", vol. 45, n^o 2, 2009, p. 184–190 RU .

[17] D. RYABKO, B. RYABKO. *Nonparametric Statistical Inference for Ergodic Processes*, in "IEEE Transactions on Information Theory", 2010 RU .

[18] D. RYABKO, J. SCHMIDHUBER. *Using Data Compressors to Construct Order Tests for Homogeneity and Component Independence*, in "Applied Mathematics Letters", vol. 22, n^o 7, 2009, p. 1029–1032 CH .

[19] M. DE VILMORIN, E. DUFLOS, P. VANHEEGHE. *Radar Optimal Times Detection Allocation in Multitarget Environment*, in "IEEE Systems Journal", vol. 3, n^o 2, Jne 2009, p. 210–220.

Articles in National Peer-Reviewed Journal

[20] R. COULOM. *Le jeu de go et la révolution de Monte Carlo*, in "Interstices", April 2009.

International Peer-Reviewed Conference/Proceedings

[21] J.-Y. AUDIBERT, S. BUBECK. *Minimax Policies for Adversarial and Stochastic Bandits*, in "22th annual conference on learning theory", 2009.

[22] S. BUBECK, R. MUNOS, G. STOLTZ. *Pure Exploration in Multi-Armed Bandits Problems*, in "Proc. of the 20th International Conference on Algorithmic Learning Theory", 2009.

[23] P. COQUELIN, R. DEGUEST, R. MUNOS. *Sensitivity analysis in HMMs with application to likelihood maximization*, in "Proceedings of Advances in Neural Information Processing Systems", 2009.

[24] R. COULOM. *The Monte-Carlo Revolution in Go*, in "Japanese-French Frontiers of Science Symposium (JFFoS'2009), Shonan, Japan", January 2009.

[25] E. DAUCÉ. *A Model of Neuronal Specialization Using Hebbian Policy-Gradient with Slow Noise*, in "Proc. of the Int'l Conf. on Artificial Neural Networks (ICANN)", Lecture Notes in Computer Science (LNCS), vol. 5768, Springer, 2009, p. 218–228.

[26] S. DELEPOULLE, C. RENAUD, P. PREUX. *Photometric compression and interpolation for light source representation*, in "Proc. 3IA, Athens, Greece", May 2009.

[27] A. M. FARAHMAND, M. GHAVAMZADEH, CS. SZEPESVÁRI, S. MANNOR. *Regularized Fitted Q-iteration for Planning in Continuous-Space Markovian Decision Problems*, in "Proceedings of the American Control Conference", 2009 CN IL .

[28] A. M. FARAHMAND, M. GHAVAMZADEH, CS. SZEPESVÁRI, S. MANNOR. *Regularized Policy Iteration*, in "Proceedings of Advances in Neural Information Processing Systems 21", MIT Press, 2009, p. 441-448 CN IL .

[29] A. LAZARIC, R. MUNOS. *Hybrid Stochastic-Adversarial On-line Learning*, in "Proceedings of Computational Learning Theory", 2009.

[30] M. LOTH, P. PREUX, S. DELEPOULLE, C. RENAUD. *ECON: a Kernel Basis Pursuit Algorithm with Automatic Feature Parameter Tuning, and its Application to Photometric Solids Approximation*, in "Proc. International Conference on Machine Learning and Applications (ICML-A)", IEEE Press, December 2009, –.

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

[32] O.-A. MAILLARD, N. VAYATIS. *Complexity versus Agreement for Many Views*, in "Proc. of the 20th International Conference on Algorithmic Learning Theory", 2009, p. 232-246.

[33] P. PREUX, S. GIRGIN, M. LOTH. *Feature Discovery in Approximate Dynamic Programming*, in "Proc. IEEE Approximate Dynamic Programming and Reinforcement Learning (ADPRL)", IEEE Press, Mar–Apr. 2009, p. 109–116 TR .

[34] A. RABAOUI, N. VIANDIER, J. MARAIS, E. DUFLOS. *On the Use of Dirichlet Process Mixtures for the Modelling of Pseudorange Errors in Multi-constellation Based Localisation*, in "Proceedings of the 9th International Conference on ITS Telecommunications", October 2009 TN .

[35] A. RABAOUI, N. VIANDIER, J. MARAIS, E. DUFLOS. *Using Dirichlet Process Mixtures for the Modelling of GNSS Pseudorange Errors in Urban Canyon*, in "Proceedings of ION GNSS 2009", September 2009 TN .

[36] S. N. RAVAZI, C. HAAS, E. DUFLOS, P. VANHEEGHE. *Real world implementation of belief function theory to detect dislocation of materials in construction*, in "Proceedings of the 12th International Conference on Information Fusion", ISIF, july 2009, p. 748–755 CN .

[37] D. RYABKO. *An impossibility result for process discrimination*, in "Proc. 2009 IEEE International Symposium on Information Theory, Seoul, South Korea", IEEE, 2009, p. 1734-1738.

[38] D. RYABKO. *Characterizing predictable classes of processes*, in "Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI'09), Montreal, Canada", J. BILMES, A. NG (editors), 2009.

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

[40] B. RYABKO, D. RYABKO. *Using Kolmogorov Complexity for Understanding Some Limitations on Steganography*, in "Proc. 2009 IEEE International Symposium on Information Theory, Seoul, South Korea", IEEE, 2009, p. 2733-2736 RU .

[41] N. VIANDIER, A. RABAOUI, J. MARAIS, E. DUFLOS. *Enhancement of Galileo and multi-constellation accuracy by modeling pseudorange noises*, in "Proceedings of the 9th International Conference on ITS Telecommunications", October 2009 TN .

National Peer-Reviewed Conference/Proceedings

[42] N. VIANDIER, F. NAHIMANA, J. MARAIS, E. DUFLOS. *GNSS Accuracy enhancement in urban environments based on error modeling and sequential Monte Carlo*, in "Proceedings of Worhshop on Localisation Precise pour les Transports Terrestres", INRETS, June 2009.

Workshops without Proceedings

[43] A. M. FARAHMAND, M. GHAVAMZADEH, CS. SZEPESVÁRI, S. MANNOR. *Regularization in Reinforcement Learning*, in "Multidisciplinary Symposium on Reinforcement Learning (MSRL)", 2009 CN IL .

[44] A. M. FARAHMAND, M. GHAVAMZADEH, CS. SZEPESVÁRI, S. MANNOR. *Robot Learning with Regularized Reinforcement Learning*, in "Workshop on Regression in Robotics: Approaches and Applications at Robotics: Science and Systems Conference (RSS)", 2009 CN IL .

[45] M. GHAVAMZADEH, Y. ENGEL. *Bayesian Actor-Critic: A Bayesian Model for Value Function Approximation and Policy Learning*, in "Workshop on Regression in Robotics: Approaches and Applications at Robotics: Science and Systems Conference (RSS)", 2009 IL .

[46] M. GHAVAMZADEH. *Hierarchical Hybrid Reinforcement Learning Algorithms*, in "Workshop on Bridging the Gap between High-Level Discrete Representations and Low-Level Continuous Behaviors at Robotics: Science and Systems Conference (RSS)", 2009.

[47] D. RYABKO. *Criteria for hypothesis testing for discrete-valued stationary processes*, in "European Meeting of Statisticians, Toulouse, France", 2009.

Scientific Books (or Scientific Book chapters)

[48] S. DELEPOULLE, C. RENAUD, P. PREUX. *Light Source Storage and Interpolation for Global Illumination: a neural solution*, in "Intelligent Computer Graphics", Studies in Computational Intelligence, chap. 5, Springer, 2009, p. 87–104.

Research Reports

[49] S. BHATNAGAR, R. SUTTON, M. GHAVAMZADEH, M. LEE. *Natural Actor-Critic Algorithms*, n^{o} TR09-10, Department of Computing Science, University of Alberta, 2009, Technical report IN CN .

[50] H. KADRI, E. DUFLOS, M. DAVY, P. PREUX, S. CANU. *A General Framework for Nonlinear Functional Regression with Reproducing Kernel Hilbert Spaces*, n^{o} RR-6908, INRIA, April 2009, Research Report.

[51] M. LOTH, P. PREUX. *The Equi-Correlation Network: a New Kernelized-LARS with Automatic Kernel Parameters Tuning*, n^{o} RR-6794, INRIA, January 2009, Research Report.

Other Publications

[52] R. COULOM. *Criticality: a Monte-Carlo Heuristic for Go Programs*, January 2009, Invited presentation at the University of Electro-Communication, Tokyo, Japan.

[53] R. COULOM. *Local Quadratic Logistic Regression for Stochastic Optimization of Parameters*, January 2009, Invited presentation at the University of Electro-Communication, Tokyo, Japan.

[54] A. LAZARIC, M. GHAVAMZADEH. *Bayesian Multi-Task Reinforcement Learning*, 2009, in preparation.

[55] M. LOTH, P. PREUX. l_1 *regularization path for functional features*, April 2009, Sparsity in Machine Learning and Statistics Workshop, Cumberland Lodge, UK (1 page).

[56] M. LOTH, P. PREUX. *Automatic kernel parameter tuning for supervised learning: the ECON approach*, May 2009, Benelearn (2 pages abstract).

[57] V. MARSAULT. *Développement d'algorithmes de planification pour le jeu de Havannah*, ENS Cachan, 2009, Mémoire de Licence.

[58] R. MUNOS, A. LAZARIC, M. GHAVAMZADEH. *Approximate Policy Iteration without Value Function Representation*, 2009, in preparation.

[59] P. PREUX, S. GIRGIN. *Sparsity in Adaptive Control*, April 2009, Sparsity in Machine Learning and Statistics Workshop, Cumberland Lodge, UK (1 page).

References in notes

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

[61] A. BARTO, R. SUTTON, C. ANDERSON. *Neuron-Like Elements that can Solve Difficult Learning Control Problems*, in "IEEE Transaction on Systems, Man and Cybernetics", vol. 13, 1983, p. 835-846.

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

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

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

[65] Y. ENGEL, S. MANNOR, R. MEIR. *Reinforcement Learning with Gaussian Processes*, in "Proceedings of the Twenty Second International Conference on Machine Learning", 2005, p. 201-208.

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

[67] A. FERN, S. YOON, R. GIVAN. *Approximate Policy Iteration with a Policy Language Bias: Solving Relational Markov Decision Processes*, in "Journal of Artificial Intelligence Research", vol. 25, 2006, p. 85-118.

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

[69] V. KONDA, J. TSITSIKLIS. *Actor-Critic Algorithms*, in "Proceedings of Advances in Neural Information Processing Systems 12", 2000, p. 1008-1014.

[70] V. KONDA, J. TSITSIKLIS. *On Actor-Critic Algorithms*, in "SIAM Journal on Control and Optimization", vol. 42, n^{o} 4, 2003, p. 1143-1166.

[71] M. LAGOUDAKIS, R. PARR. *Reinforcement Learning as Classification: Leveraging Modern Classifiers*, in "Proceedings of the Twentieth International Conference on Machine Learning", 2003, p. 424-431.

[72] J. PETERS, S. VIJAYAKUMAR, S. SCHAAL. *Natural Actor-Critic*, in "Proceedings of the Sixteenth European Conference on Machine Learning", 2005, p. 280-291.

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

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

[75] J. RAMSEY, B. SILVERMAN. *Functional Data Analysis*, 2nd edition, Springer, 2005.

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

[77] J. RUST. *How Social Security and Medicare Affect Retirement Behavior in a World of Incomplete Market*, in "Econometrica", vol. 65, n^{o} 4, July 1997, 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", vol. 15, n^{o} 2, 1997, p. 195–208, http://129.3.20.41/eprints/io/papers/9512/9512002.abs.

[79] J. SETHURAMAN. *A constructive definition of Dirichlet priors*, in "Statistica Sinica", vol. 4, 1994, p. 639-650.

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

[81] R. SUTTON. *Temporal credit assignment in reinforcement learning*, University of Massachusetts Amherst, 1984, Ph. D. Thesis.

[82] R. SUTTON. *Learning to Predict by the Methods of Temporal Differences*, in "Machine Learning", vol. 3, 1988, p. 9-44.

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

[84] P. WERBOS. *ADP: Goals, Opportunities and Principles*, in "Handbook of learning and approximate dynamic programming", J. SI, A. BARTO, W. POWELL, D. WUNSCH (editors), IEEE Press, 2004, p. 3–44.