INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team SequeL

Sequential Learning

Futurs

THEME COG

Table of contents

1. Team

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 1 and the University of Lille 1).*

Head of the team

Philippe Preux [Professor, Université de Lille, HdR]

Vice-Head of the team

Rémi Munos [Research Director (DR), Inria, HdR]

Project assistant

Véronique Couvreur [Secretary (SAR) Inria, shared by 3 projects, left on Sep 30th, 2007]

Sandrine Catillon [Secretary (SAR) Inria, shared by 3 projects, arrives on Oct 8th, 2007]

Staff member

Manuel Davy [Researcher (CR) CNRS, HdR]

Daniil Ryabko [Researcher (CR) INRIA, arrives on Dec 1st, 2007]

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]

Post-Doctoral fellow

Thomas Bréhard [INRIA, left on Jun 30th, 2007]

Sertan Girgin [INRIA, begins on Jul 1st, 2007]

Djalel Mazouni [INRIA, begins on Sep 1st, 2007]

Stéphane Rossignol [CNRS & industry]

Ph.D student

Pierre-Arnaud Coquelin [École Polytechnique]

Robin Jaulmes [DGA]

Manuel Loth [INRIA-Région Nord-pas-de-calais Grant]

Jean-François Hren [MENESR Grant]

Raphaël Maîtrepierre [MENESR Grant]

Sébastien Bubeck [ENS Grant]

Amine Chouiha [CORDI Grant]

Technical staff

Antoine Labitte [Assistant Engineer, begins on Oct 1th, 2007]

Student interns

Jean-François Hren [Master 2 internship, Feb to Jun 2007]

Raphaël Maîtrepierre [Master 2 internship, Feb to Jun 2007]

Amine Chouiha [Master 2 internship, Apr to Sep 2007]

Yzao Wang [Master 2 internship, Apr to Jul 2007]

Thomas Huguerre [Master 2 internship, Feb to Jun 2007]

Guillaume Libersat [Master 1 internship, Feb to Jun, and Sep 2007]

Loïc Villanné [Master 1 internship, Feb to Jul, 2007]

Simon Perrault [Master 1 internship, Feb to Jul, 2007]

Michel Moyart [ISEN engineer internship]

Gaël Ladreyt [ENS-Cachan internship]

Stepan Albrecht [internship]

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 [85]). 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

A major activity of the year is the incubation of the Predict & Control spin-off under the responsability of Pierre-Arnaud Coquelin.

Predict & Control is officialy created in december 2007. A significant part of the year has been dedicated to the creation of the spin-off, the definition of its relations with SEQUEL and INRIA, the search for scientific advisors, and the search of private societies to work with.

Predict & Control aims at providing expertize in machine learning, targeting particularly the commerce fields. This is the topic of one of the "pôles de compétitivité" of the region Nord-Pas de Calais. The expertize may span from a feasability study, to the design and realization of a software to solve a particular problem.

3. Scientific Foundations

3.1. Introduction

SEQUEL is primarily grounded on two domains:

- Markov decision problems which provide the general setting of the problem we want to solve,
- statistical learning which provide the general concepts and tools to solve this problem.

We briefly present key ideas below.

3.2. Markov decision problems

Keywords: *approximate dynamic programming*, *dynamic programming*, *policy search*, *reinforcement learning*, *sequential decision problem*.

Sequential decision problems occupy the heart of the SEQUEL project [83].

A Markov Decision Process 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.

The goal of the Markov Decision Problem is to find a policy π that maximizes 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:

[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.

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

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 [74]) 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) [72], 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 [73]. 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 (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:

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

In short, we would like to mention that most of the reinforcement learning methods developed so far are built on one (or both) of the two following approaches ([86]):

- Bellman's dynamic-programming approach, based on the introduction of the value function. It consists in learning a "good" approximation of the optimal value function, and then using it to derive a greedy policy w.r.t. this approximation. The hope (well justified in several cases) is that the performance V^{π} of the policy π greedy w.r.t. an approximation V of V^* will be close to optimality. This approximation issue of the optimal value function is one of the major challenge inherent

to the RL 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 RL 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.

3.3. Statistical learning

Keywords: *Bayesian formalism*, *Monte-Carlo methods*, *kernel methods*.

machine learning 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 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 [79]. We briefly present the most important for us to date, namely, kernel-based non parametric function approximation, sequential Monte-Carlo methods, and non parametric Bayesian models.

3.3.1. Kernel methods for non parametric function approximation

In SEQUEL , the model to be learned is a real-valued function defined in a multi-dimension space.

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).

For the moment, we consider non parametric methods since they do not make any assumptions about the function to learn. Locally weighted regression have yielded efficient methods to learn a policy in reinforcement learning, as well as good performance in regression settings. The kernelized version gives us a wide ability to handle sample points and combine them to obtain the approximation. To keep computation times of practical interest, a sparse representation is sought.

We currently devote a lot of efforts to LARS-like approximators [76], that we have fitted into the reinforcement learning framework [43].

3.3.2. Monte-Carlo methods

Sequential Monte-Carlo (or particle filtering, see [69]) methods are currently used for various purposes in SEQUEL :

- the estimation of the state of the agent given its current observation as well as its history;
- the estimation of parameters of a model.

3.3.3. 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 let 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 an other approach to this problem. In particular, mixtures of Dirichlet processes [77] 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:

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

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

Keywords: *automatic transcription of speech*, *civil engineering*, *customer affluence modeling*, *environment*, *games*, *multimedia*, *sensor localization*, *transportation systems*.

[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 [84].

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,
- sensor management problem,
- neurosciences.

4.2. Adaptive control

Adaptive control is an important potential 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 [82]. But, reinforcement learning may also handle situations where the dynamics of the system is unknown, situations where the system is partially observable, and non stationary situations. Indeed, in these cases, the behavior is learned by interacting with the environment and thus naturally adapts to the changes of the environment. Furthermore, the adaptive system may also take advantage of expert knowledge when available.

Clearly, the spectrum of potential application 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.

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.

The signal may be mono-channel, audio, or visio, or magnetic, or more generally electro-magnetic (*e.g.*, RFID, or Bluetooth, or wifi, or signals sent by GPS satellites), or else. There might also be several (multi-channel) signals of different nature.

4.4. Functional prediction

One of the current trends in machine learning aim 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. Sensor management problem

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.

4.6. 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

Some software has begun to be developped in SEQUEL . Different threads are followed. For the moment, this software is yet in a rather crude form. We have begun to make it available through our website and via the INRIA forge. It will be developped further in the coming years in its functionalities, as well as in accessibility for general users (including GUIs, documentation, examples, tutorials, ...). This software falls under two varieties: either the implementation of research level algorithms, or the implementation of software tools to make research easier.

5.1.1. 1-class support vector machine

Keywords: *1-class SVM*, *quadratic programming*, *sequential*, *support vector machine*.

Participants: Stéphane Rossignol [correspondant], Michel Moyart.

SMO (Sequential Minimal Optimization) is a numerical optimizer of quadratic programming problems. It does not require the creation of large matrices, allowing thus considering problems with a few millions samples (like, for instance, the automatic transcription of speech problem, see 6.4.4.1). However, it is relatively slow. Stéphane Rossignol worked on an optimized and fast C version of SMO for the 1 class SVM problem. His software will be made available online on the SEQUEL website.

5.1.2. Electronic radar simulation

Keywords: *electronic radar simulation software*.

Participants: Emmanuel Duflos [correspondant], Thomas Huguerre.

The kernel of a simulator of electronic radar has been developped as part of Thomas Huguerre's internship. This kernel has been developped in C++. More work is on-going to make it usable for general users, and make it available through our website.

5.1.3. Crazystone

Keywords: *Go software*.

Participant: Rémi Coulom [correspondant].

Crazystone is an award-winning Go software player, designed and developed by Rémi Coulom.

Being a research tool related to high worldwide competition, it is no longer freely available.

5.1.4. Brennus

Keywords: *Poker software*.

Participants: Raphaël Maîtrepierre [correspondant], Jérémie Mary.

Brennus is a poker bot, that is a program designed to play Poker against other programs, interacting via the Internet (Brennus may play against human players as well). This is the first release of this program. Brennus is related to a new track of research in SEQUEL and the result of Raphaël Maîtrepierre research for his masters thesis, and now his PhD.

6. New Results

6.1. Introduction

New results are organized is the following sections:

1. reinforcement learning,
2. sensor management problem,
3. signal processing.

6.2. Reinforcement learning

Keywords: *Feynmann-Kac flow*, *LARS*, *Lp-norm*, *Monte-Carlo estimation*, *dynamic programming*, *exploration-exploitation trade-off*, *multi-arm bandit*, *non parametric function approximation*, *performance bound*, *policy search*, *value function approximation*, *variance reduction*, *viability problem*.

Participants: Sébastien Bubeck, Amine Chouiha, Pierre-Arnaud Coquelin, Rémi Coulom, Manuel Davy, Sertan Girgin, Jean-François Hren, Robin Jaulmes, Manuel Loth, Raphaël Maîtrepierre, Jérémie Mary, Djalel Mazouni, Rémi Munos, Philippe Preux, Yzao Wang.

6.2.1. Theoretical foundations of reinforcement learning

We have worked on several aspects of reinforcement learning and optimal control, including the use of function approximation to represent the value function or the policy. We have worked in collaboration mainly with Csaba Szepesvári (University of Alberta, Canada), András Antos (Hungarian Academy of Sciences), Jean-Yves Audibert (CERTIS, Ecole des Ponts et Chaussées), Guillaume Deffuant and Sophie Martin (Cemagref, Clermont-Ferrand), Hasnaa Zidani (ENSTA), Olivier Bokanowski (Paris VII), Olivier Teytaud (LRI, Orsay). This work can be summarized as follows:

- **Establishing links between statistical learning and reinforcement learning**. Performance bounds on the policies deduced by approximate dynamic programming methods (such as approximate value iteration, approximate policy iteration) when using sampling devices are established in terms of the capacity (using VC dimension, covering numbers) of the function space considered in the approximations. See[26], [25], [10]
- **Analysis of dynamic programming using L_p-norms**. This work extends usual analysis in L_∞-norm to L_p-norms, which opens promizing new directions towards an analysis of dynamic programming and reinforcement learning combined with function approximation. See [21][20].
- **Policy gradient estimation in continuous time**. This method allows to search directly for a locally optimal controller in a class of parameterized policies, in the case of continuous-time state-dynamics [7]. An application of this method to a control problem in finance has been worked on [71].
- **Analysis of the exploration-exploitation tradeoff using variance estimate**. We investigate the multi-armed bandit framework using new deviation inequalities that takes into account the variance estimate. This results in a great sharpening of the regret bounds. See [27], [11], [51].
- **Use of bandit algorithms for performing tree search.** We investigate the recursive use of bandit algorithms for designing efficient tree exploration policies. The resulting methods explore the tree in an asymmetric way expanding and exploring first the most promising branches. We analyzed the UCT algorithm [80] (UCB algorithm [70] applied to trees) and use it for tree-search in the game of go (see sec. 6.2.5.1 below). With Pierre-Arnaud Coquelin, we further investigated improved bandit algorithms for tree search, providing performance guarantees [32]. Several master students have done an internship on related topics (Jean-Francois Hren [61] and Amine Chouia [60]) and are currently beginning their PhD on several extensions of this domain. Yizao Wang has done his second year master internship at SequeL in collaboration with Jean-Yves Audibert, from the CERTIS (ENPC), on bandit algorithms applied to tree search when using variance estimates [68]. This domain stands as one of our priviledged research directions.

6.2.2. Numerical approximation of viability problems

We use several ideas from the ultra-bee schemes used for transport equation with discountinuous solutions and the dynamic programming approach combined with function approximation to approximate the viability kernel of a viability problem [75][31].

6.2.3. Estimation of the gradient of a Feymann-Kac flow

This is a joint work between Pierre-Arnaud Coquelin and Romain Deguest at Centre de Mathématiques Appliquées de l'École Polytechnique (CMAP).

We have proposed a numerical method to estimate the gradient of a Feynman-Kac flow. The idea is to achieve a sensitivity analysis along a Markov chain canonically associated to the Feynman-Kac model. One can use classical methods, such as likelihood ratio method or infinitesimal perturbation analysis, to estimate the gradient. We have proposed more efficient algorithms, i.e that have a lower variance, to estimate this gradient.

The range of applications is quite broad: maximum likelihood parameter estimation in Hidden Markov Model, Direct Policy search in Partially Observable Markov Decision Process, Policy optimization in risk sensitive cost Markov decision process problem and sensitivity analysis of rare event with respect to some parameter of the model [54].

6.2.4. Non parametric function approximation for RL: the Equi-Gradient TD algorithm

After 2006 work on the adaptation of the kernelized-LARS algorithm to fit the reinforcement learning problem [43], which we named the "equi-gradient descent" algorithm (EGD). A major advance in 2007 has been to extend the traditionnal LARS approach to an infinite number of features. This opens-up many applications of EGD to model fitting, such as radial basis funciton networks, neural networks, wavelets, ... beyond the reinforcement learning scope. Current work goes on on this point.

We have also proposed a unified view of many algorithms (TD, residual TD, iLSTD, LSTD, LSPE and our equi-gradient TD algorithm), showing that they all fit into a single, and simple, formalism [44].

6.2.5. Exploration vs. exploitation

6.2.5.1. The game of Go

After the 2006 major breakthrough in go realized by Rémi Coulom's Crazystone program, the latter has evolved further. He won a bronze medal in the 9×9 game and a silver medal in the 19×19 game at the 12^{th} Computer Olympiads in Amsterdam.

The main research advance in 2007 has consisted in incorporating domain knowledge into his Go-playing program by supervised learning from expert games. This is a follow-up on his previous research on Monte-Carlo tree search. This new technique lead to huge strength improvements. Even on the large 19x19 board, his program is now stronger than the strongest classical non-Monte-Carlo programs [33], [34].

In parallel, an other track was followed in the team by Rémi Munos who collaborated with Yizao Wang, first year master student at CMAP, École Polytechnique, Sylvain Gelly, PhD student, and Olivier Teytaud at INRIA TAO. The resulting program MoGo is currently the world best computer-go program [16], [38] [78].

6.2.5.2. The game of Poker

We began a work on games with incomplete information. The keypoint is to adapt some techniques used in to balance exploration and exploitation, and take advantage of some theoretical bounds on bandit problems to create new solutions to adapt the computer strategie to its opponent, during the game. We chose poker to apply these ideas. Poker blends interesting topics of research (partially observable problem, time-varying problem) to some high economics interests.

Raphaël Maîtrepierre, Jérémie Mary, and Rémi Munos created Brennus, a computer Texas Hold'em poker player[64], [65]. This new bots, based on really new techniques in the field of poker games, perfomed quite well: it defeated all opponents of the 2006 challenge and ranked 8th/17 at the AAAI poker challenge, held in august 2007, in Vancouver[4]. The very interesting point is that only the newest pseudo-equilibrium bots defeated Brennus. This means that Brennus is able to find a strategy not too far from optimal play while it keeps the ability to adapt. Moreover, new trends in computer game is to face multiplayer games. In this area the traditionnal approach on pseudo-equilibrium does not scale-up.

6.2.6. Algorithmic issues in reinforcement learning

An activity is also going on regarding the practical application of reinforcement learning, based on recently published ideas, and the combination of various approaches. In particular, we have worked on the following lines:

- investigate natural gradient to obtain better approximation in value-based approach, as well as in a policy search approach.
- investigate the transfer of knowledge, learning a task in a simple setting and using what has been learnt in a larger problem setting.
- investigate automatic feature selection to find better representations of the problem that make it easier to learn the value function or the policy.

[4] see AAAI 2007 Poker challenge website

This work requires some efforts to obtain experimental results and to be able to draw experimental guidelines for the practitionner. This work is on-going and preliminary results are available in [55]. We use the GRID 5000 to speed-up the experimental work.

6.2.7. Neurosciences

This is a joint work between Pierre-Arnaud Coquelin, Andrea Brovelli and Driss Boussaoud of the "Institut de Neurosciences Cognitives de la Méditerranée" (INCM). The goal was to find the neural learning model used by a monkey during a sequential associative learning task. We developed an approach based on maximum likelihood estimation in Hidden Markov Model. The results were very interesting and the approach is quite novel in this field [28], [12].

6.3. Sensor management problem: the management of electronically scanned array radars

Keywords: *electronic scanned radar*, *non parametric bayesian learning*, *probability of detection*, *radar*, *radar scheduling*, *reinforcement learning*, *reinforcement learning*, *sensor management problem*.

Participants: Emmanuel Duflos, Thomas Bréhard, Thomas Huguerre, Philippe Vanheeghe, Pierre-Arnaud Coquelin.

This is a special case of a sensor management problem. here, we suppose that all the sensors have the same type: Electronically Scanned Array (ESA) radars. These radars are in a context of target detection, and target tracking. This is a common application in sensor management and typically a military

6.3.1. Radar scheduling

First results have been developed in the framework of the scheduling of radars in a multitarget environment. The scheduling is based on the modeling of the probability of detection of a target. The detection process has been improved in order to maximize this probability. This optimization leads to a specific form for the probability of detection which allows analytical derivation of scheduling strategies for one radar in a multitarget environment: if the radar has to spend a given time T to detect N (N is suppposed to be known) targets, how much time must it allocate to each target if the criterion to optimize is the sum of the probabilities of detection? Some results have been extended to the multisensor/multitarget environment ([37]).

6.3.2. Myopic Approach

A method based on the modeling of the probability of detection of the radar and on the posterior probability of presence of a target in a radar resolution cell (knowing the past actions and the past measures) P_p has been developped (the context is supposed to be multitarget, and monosensor). After each analysis of a direction, the posterior probability of presence of a target in a cell is updated. The presence of a target in a cell being modeled by a Bernouilli random variable of probability P_p, the choice of the next action is based on the minimization of the variance of this random variable. A first algorithm, very simple, although not very realistic has first been proposed: the targets are supposed not to move during the observation. This algorithm was then modified to take into account the movement of the target which is supposed to be markovian with respect to the cells [62], [39].

6.3.3. The sensor management problem as a reinforcement learning problem

The sensor management problem (SMP) is a practical application to which we have decided to pay a large effort in SEQUEL . We want to consider the SMP as a particular reinforcement learning problem. Hence, the first step in this effort is to formulate the SMP as some Markov decision problem.

6.3.3.1. Q-learning for radar Management

The use of classical methods of Q-learning for ESA radars has also been evaluated. This method was recently described in [81]. This method works quite well when the cells number is not to large which is not the case with an ESA radar [62], [39].

6.3.3.2. Policy Search

A new and original approach consisting in deriving the optimal parameterized policy based on stochastic gradient estimation has also be developed [52]. Two different technics, namely the Infinitesimal Approximation (IPA) and the Likelihood Ratio (LR), have been used to address the problem. This work is based on the PhD results of Pierre-Arnaud Coquelin.

6.4. Signal analysis and processing

Keywords: *Dirichlet mixture process*, *GPS*, *RFID tags*, *anomaly detection*, *clustering*, *land vehicle localization*, *locutor recognition*, *material localization*, *model selection*, *rupture detection*, *speech processing*, *speech transcription*, *state estimation*, *supervised learning*, *unsupervised learning*.

Participants: Manuel Davy, Emmanuel Duflos, Philippe Vanheeghe, Stéphane Rossignol.

6.4.1. Bayesian unsupervised learning

6.4.1.1. Time varying clustering

Time varying clustering with first order stationary Pitman-Yor processes [30]. There is a need to develop models to cluster evolving data, where the number and the composition of the clusters may evolve and adapt sequentially. We have developped a new class of Pitman-Yor processes which have a given fixed marginal distribution at each time, and a given cluster dynamic model.

6.4.1.2. Latent variable estimation

Maximum likelihood in latent variable models. For this kind of problems, the EM algorithm is the well known and efficient approach used by many researchers. However, the EM algorithm is gradient-based, meaning that it converges to local optima. We have developped [18] a Monte Carlo algorithm to be used whenever the EM approach fails. This applies a Sequential Monte Carlo strategy which can be, to some extent, related to genetic algorithms with the major difference that convergence results hold in our approach.

6.4.1.3. Bayesian functional clustering

With E. Jackson (PhD student) and A. Doucet (Professor at the U. of British Columbia), we have investigated Bayesian functional clustering. Given observations obtained by sampling different functions at random locations (and different from one function to an other), we have developed an algorithm that clusters these functions into coherent groups. For instance, each observation may be a signal (in one or more dimensions), such that the sampling instants are different from one signal to an other. We have then modeled the underlying signals by using Gaussian processes and the clustering itself is performed by using a Dirichlet mixture process [40]. The target applications concern the clustering of expression data of messenger-RNA, as well as sampling data originating from geostatistics.

6.4.2. Joint segmentation of piecewise constant autoregressive processes

Certain electrical devices include a set of electric cables which, under certain circumstances, may give rise to electrical arcs (a typical example is the command circuitry of an airplane). This is basically a problem of detecting ruptures in a multichannel environment.

We have designed a Bayesian algorithm to detect these ruptures which models the signals on each cable individually, by an autoregressive model and that assumes a correlation between the instants of rupture in the different cables [15].

6.4.3. Localization

6.4.3.1. Sequential learning of sensors localization

This work is done in collaboration with Prof Carl Haas of the University of Waterloo (Canada). This collaboration is related to a problem appearing in civil engineering: how can we automatically localize the building materials on a construction site? This is a real problem because a lot of time (hence of money) is lost to look for these materials that have often been moved away. The proposed solution is to equipped 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 send 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. When the nodes are moving the problem is still more complicated. Our work shows how the Transferable Belief Model can be used to learn the position of the communication nodes and to detect potential movements. This study also shows how to deal with the computation.

This work has also been applied for land vehicle localization. The vehicle is equipped with three sensors, including a GPS sensor.

6.4.3.2. Accurate Localization using Satellites in Urban Canyons

This work is done in collaboration with Juliette Marais, junior researcher at INRETS and Fleury Donnay NAHIMANA, a PhD Student supervised by Emmanuel Duflos and Juliette Marais.

Lots of Global Navigation Satellite System (GNSS) applications deal today with transportation. However, main transport applications, either by rail or road, are used in dense urban areas or, to the least, in suburban areas. In either one, the conditions of reception of every available satellite signals are not ideal. The consequences of environmental obstructions are unavailability of the service and multipath reception that degrades, in particular, the accuracy of the positioning. In order to enhance GNSS performances, several research axes can be found in the literature that can deal with multi-sensors uses, electronic enhancement or receiver processing. We focus here on the multisensor approach where each satellite is considered as a sensor.

Today most of the GNNS receivers, like the well-known GPS, consider that the received noise is gaussian and use a Kalman filter. This assumption is false in urban canyon and we must find new models for the noise and derive new methods to estimate the position in an accurate way from the signals send by the satellite and froml all other information sent by each satellite. Such a problem is all the more a typical one since the future Gallileo constellation will provide the receivers with information as the integrity of the signals, leading to new services for industry.

This problem can be modelled in the framework of the sensor management problem each satellite being considered as a sensor with several *modes*. Moreover the receiver being generally in movement, it is necessary to estime with respect to time the non stationnary noise probability density function in the same time as we estimate the position.

We have shown that in narrow urban canyons the noise resulting from multipath is multimodal and can be modelled in a first approximation by a gaussian Mixture Model leading to a non linear and non gaussian estimation process [45]. When urban canyons are large, reception conditions are near those that can be found outside the town which means a gaussian noise. When moving the overall localization process must therefore be modelled by a Jump Markov System. An estimation process based on a particle filter has been implemented. Results of simulation show an improvement of the performances with respect to the classical receiver.

6.4.4. Supervised learning

6.4.4.1. Speech processing

Stéphane Rossignol (post-doc fellowship) and Manuel Davy, in collaboration with France Télécom, continued to work on the automatic transcription of speech using kernel methods (mostly, the 1 class SVM) [57], [58]. Their works focused on the classification problem. This is a particularly complicated problem, due to the

facts that it comprises a few thousands classes and that these classes are very unbalanced (some classes are represented by only a few data; others by millions). Stéphane Rossignol has demonstrated that it is possible to train a 1-class SVM on the hugest classes in a relatively short amount of time. He has shown as well that the dissimilarity measures used in order to classify a sample in one of the trained 1-class SVM are effective even for this extremely unbalanced problem. On a small test dataset, he has shown that the obtained performance are at least as good as the performance obtained using the existing France Télécom system. In the first half of 2007, Stéphane Rossignol and Manuel Davy began to evaluate and improve the classification models they built in 2006. From June 2007 to September 2007, Michel Moyart has been a trainee of the ISEN engineer school on this topic [56]. He improved the code, allowing the evaluations to be performed on a huge set of test data and the improvement procedure to be fast enough. These evaluations and improvements are on their way.

6.4.4.2. Anomaly detection

Stéphane Rossignol and Manuel Davy, within the framework of the contract "d'aide au transfert technologique" number 510416 between the CNRS and a company manufacturing test instruments for the professional audio industry, worked on the detection and characterization of loudspeakers flaws in the production line by using kernel methods [59]. They worked as well on the test signal and on the TFR to use in order to effectively underline the characteristics of each flaw. And they worked on the overall methodology to follow: some flaws are hardly separable, requesting thus an additional feature extraction step; furthermore, the flaws are slightly evolving in time; and so forth. These various problems request the system to be as flexible as possible. Stéphane Rossignol and Manuel Davy have demonstrated that it is possible to use the 1-class SVM technique in order to discriminate between the flaws. A completely functional system is on the way.

6.4.4.3. Characterization of environmental sounds

Stéphane Rossignol worked on the automatic characterization of environmental sounds using kernel methods. First, concerning the segmentation and indexing of musical audio signals, he improved the performance of the techniques he developed during his Ph-D thesis and afterwards. Using the KCD (Kernel Change Detection) technique in order to underline the transitions between notes allows to reduce the number of false alarms by 80 %. Second, he worked with Asma Rabaoui on the application of the One-class SVM technique to Audio Surveillance Systems. Nine classes of environmental sounds (gunshots, cries, barkings, etc.) have been effectively discriminated using the One-class SVM technique. More than 90 % of good classification rate has been obtained [48], [46], [67].

6.4.4.4. Supervised learning for altimetry radar data classification

Bayesian supervised classification with generative models. In classification problems, it sometimes happens that a model that describes the data generation process is available (this is called a generative model). Learning a bayesian classifier comes down to learning the posterior distribution of the model parameters, which requires to define prior distributions over these paramters. However, simple choices like the Gaussian are often inaccurate. Therefore, we propose to use a Dirichlet Process mixture prior so as to gain more learning flexibility while not ending up with an overly complex model. This has been applied to radar altimetry data classification. To be more specific, satellites like Topex-Poseidon carry radars to measure very precisely their height. However, the response of the earth surface to the electromagnetic pulse carries more information than just the altitude: its shape depends on the type of surface it hits (ocean, forest, etc.) and it is interesting to obtain this extra information [35].

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

As an INRIA team, SEQUEL has not signed any contract by way of the INRIA. However, various works in 2007 have been done under contract, such as France Telecom, Auchan, and others (confidentiality required).

7.1.1. Speech analysis with France Telecom

This work was described in sec. 6.4.4.1.

7.1.2. Affluence prediction in Auchan supermarkets

In 2007, we have dealt with a large-scale application of functional prediction, involving state-of-the-art supervised and non supervised learning methods. More precisely, we have worked with the group Auchan which is a major international group which operates more that 150 Hypermarkets worldwide. Among other important issues, the prediction of the number of customers reaching the cashiers at a given time of the day has been worked on in 2007. For each day, this is a functional prediction problem (number of open cashiers against time slot of each particular day), which may be seen as non-stationary, because the customer habits evolve.

8. Other Grants and Activities

8.1. Regional activities

Emmanuel Duflos takes part in the PEPSAT regional project, as the coordinator for the LAGIS. PEPSAT deals with global navigation satellite systems and is supported by the Région Nord-Pas de Calais.

8.2. National activities

8.2.1. ANR Kernsig

Participants: Manuel Davy, Thomas Bréhard.

This project is headed by Prof. S. Canu with the INSA-Rouen. It deals with the study of kernel methods for signal processing.

In 2007, Manuel Davy and Alain Rakotomamojy (LITIS, Rouen) have worked on the regularization path of the 1-class SVM [47].

8.2.2. ARC CODA

Participants: Rémi Munos, Pierre-Arnaud Coquelin, Djalel Mazouni.

A two years ARC project named CODA (for "Optimal control of an anaerobic digestor") in collaboration with the INRA laboratory in Narbonne, the INRIA project-team COMORE in Sophia-Antipolis, and the spin-off Naskeo Environment, started in 2007. A post-doc fellow (Djalel Mazouni) has been hired for one year.

Several approaches for solving this partially observable Markov decision problem have been developed by two PhD students, Pierre-Arnaud Coquelin [31], [54] using a sensitivity analysis combined with particle filtering approach, and Robin Jaulmes [17], [41] using a Bayesian setting.

We refer the interested reader to the website `http://sequel.futurs.inria.fr/munos/arc-coda` for more information, and up-to-date information.

8.3. International activities

Philippe Vanheeghe has visited the Centre for Pavement and Transportation Technology (CPATT), headed by prof. Carl Haas at the University of Waterloo, Canada, from Feb. 23rd to Mar. 11th, 2007. This deals with sensor management in order to locate building materials in building areas using RFID tags.

8.4. Visits and invitations

- Martin Zinkevich spent a week in Villeneuve d'Ascq, from Oct 22th to Oct 27th to work on Poker software, with Jérémie Mary and Raphaël Maitrepierre. M. Zinkevich has chaired the 2007 AAAI Poker challenge. He used to be a member of the University of Alberta at Edmonton. He is now with Yahoo Research in the Silicon Valley.
- Rémi Coulom has been invited by Takeshi Ito, University of Electro-Communications, Tokyo, from 6 to 9 Nov, 2007, with regards to Crazystone, his Go playing software. He gives an invited talk at the 12^{th} Game Programming Workshop.
- Stepan Albrecht, University of Bohemia, is visiting Manuel Davy during 3 months, with regards to music analysis.

9. Dissemination

9.1. Scientific Communauty animation

- Manuel Davy is associate editor for the IEEE Transactions on Signal Processing review. He has also reviewed papers for IEEE Trans. on Signal Processing, IEEE Signal Processing Letter, Speech communications, Signal Processing, IEEE Trans. on Circuits and Systems I.
- Emmanuel Duflos and Philippe Vanheeghe have organized two special sessions on the sensor management problem in the Fusion 2007 onference, held in July in Quebec City, Canada.
- Emmanuel Duflos is working towards bringing the FUSION 2010 conference in Lille. Accordingly, he sent a formal proposition to the International Society Of Information Fusion to organize the FUSION Conference in Lille in 2010. The organization is supported at the moment by the Ecole Centrale de Lille and INRIA-Futurs. The proposition must be improved in 2008. The final decision will be taken in July 2008.
- Emmanuel Duflos is also involved in the organization of the 5^{th} Computational Engineering in Systems Applications conference which will be held in 2009, in South Korea.
- Rémi Munos is member of the scientific board of the Journal of Machine Learning Research, Machine Learning Journal, Artificial Intelligence Journal, Revue d'Intelligence Artificielle. He has been a member of the PC of the 2006 conferences Neural Information Processing Systems, International Conference on Machine Learning, Conférence Francophone sur l'Apprentissage Automatique.
- Rémi Munos was an invited speaker at the Workshop on Reinforcement Learning, held in Tübingen, in July 2007. He was also invited at the "Diffusion des savoirs" seminar (Ecole Normale Supérieure), January 2007
- Rémi Munos is Associate Researcher with CREA (Centre de Recherche en Epistémologie Appliquée), École Polytechnique, since September 2007
- Rémi Munos is Co-chair of the IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning, 2007.
- Philippe Preux is a member of the program committee of the IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning, the 2007 European Conference on Machine Learning, Artificial Evolution, Reconnaissance des Formes et Intelligence Artificielle 2008, Extraction et Gestion des Connaissances 2008.
- Philippe Preux, Rémi Coulom, together with Samuel Delepoulle, have edited a special issue of the "Revue d'Intelligence Artificielle" on Markov Decision Processes [66]

9.2. Teaching

We list the courses that are related to the research activities in SEQUEL that happened in 2007.

- Rémi Munos teaches a class in reinforcement learning in the M2 “Mathematics-Vision-Learning” (MVA) at the ENS-Cachan; he also teaches a cognitive science class in M1 at the EHESS (Paris).
- Philippe Preux teaches in the M2 of computer science at the University of Lille a class on reinforcement learning.
- Stéphane Rossignol gives a class (16 hours) in Data Mining to the students of the Research Master “Génie Industriel” of the Ecole Centrale de Lille.

Otherwise, each of the 5 professors and assistant professors of the SEQUEL team teaches 192 hours per year, mostly at master level. Taught classes include machine learning, data mining, and signal processing classes.

10. Bibliography

Major publications by the team in recent years

[1] A. KLAPURI, M. DAVY (editors). *Signal Processing Methods for Music Transcription*, Springer, New York, 2006.

[2] 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.

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

[4] C. DUBOIS, M. DAVY. *Joint Detection and Tracking of Time-Varying Harmonic Components: a Flexible Bayesian Approach*, in "IEEE transactions on Speech, Audio and Language Processing", vol. 15, n^{o} 4, May 2006, p. 1283–1295.

[5] E. GOBET, R. MUNOS. *Sensitivity analysis using Itô Malliavin calculus and martingales. Application to stochastic optimal control*, in "SIAM journal on Control and Optimization", vol. 43(5), 2005, p. 1676-1713.

[6] R. MUNOS. *Geometric variance reduction in Markov chains. Application to value function and gradient estimation*, in "Journal of Machine Learning Research", vol. 7, 2006, p. 413–427.

[7] R. MUNOS. *Policy gradient in continuous time*, in "Journal of Machine Learning Research", vol. 7, 2006, p. 771–791.

[8] R. MUNOS. *Performance Bounds in Lp norm for Approximate Value Iteration*, in "SIAM J. Control and Optimization", 2007.

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

Year Publications

Articles in refereed journals and book chapters

[10] A. ANTOS, C. SZEPESVÁRI, R. MUNOS. *Learning near-optimal policies with Bellman-residual minimization based fitted policy iteration and a single sample path*, in "Machine Learning Journal", to appear, 2007.

[11] J.-Y. AUDIBERT, R. MUNOS, C. SZEPESVÁRI. *Tuning Bandit Algorithms in Stochastic Environments*, in "Theoretical Computer Science", to appear, 2007.

[12] A. BROVELLI, P.-A. COQUELIN, D. BOUSSAOUD. *Estimating the hidden learning representations*, in "Journal of physiology Paris", to appear, 2007.

[13] F. CARON, M. DAVY, A. DOUCET, E. DUFLOS, P. VANHEEGHE. *Bayesian Inference for Linear Dynamic Models with Dirichlet Process Mixtures*, in "IEEE trans. on Signal Processing", to appear, 2007.

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

[15] N. DOBIGEON, J.-Y. TOURNERET, M. DAVY. *Joint segmentation of piecewise constant autoregressive processes by using a hierarchical model and a Bayesian sampling approach*, in "IEEE transactions on Signal Processing", vol. 55, n^{o} 4, April 2007, p. 1251–1263.

[16] S. GELLY, R. MUNOS. *L'ordinateur, champion de go ?*, in "Pour la Science", vol. 354, 2007, p. 28-35.

[17] R. JAULMES, J. PINEAU, D. PRECUP. *Apprentissage actif dans les processus décisionnels de Markov partiellement observables*, in "Revue d'Intelligence Artificielle", vol. 21, n^{o} 1, 2007, p. 9–34.

[18] A. JOHANSEN, A. DOUCET, M. DAVY. *Particle Methods for Maximum Likelihood Estimation in Latent Variable Models*, in "Statistics and Computing", to appear, 2007.

[19] 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", 2007.

[20] R. MUNOS. *Analyse en norme L_p de l'algorithme d'itérations sur les valeurs avec approximations*, in "Revue d'Intelligence Artificielle", vol. 21, n^{o} 1, 2007, p. 55–76.

[21] R. MUNOS. *Performance Bounds in Lp norm for Approximate Value Iteration*, in "SIAM J. Control and Optimization", 2007.

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

[23] K. OTA, E. DUFLOS, P. VANHEEGHE. *Bayesian Inference for Speech Density Estimation by the Dirichlet Process Mixture*, in "Studies in Informatics and Control", vol. 16, n^{o} 2, September 2007, p. 227–244.

[24] E. SAHIN, S. GIRGIN, L. BAYINDIR, A. E. TURGUT. *Swarm Robotics*, in "Swarm Intelligence. Introduction and Applications", C. BLUM, D. MERKLE (editors), Natural Computing Series, To appear, Springer Verlag, Berlin, Germany, 2008.

Publications in Conferences and Workshops

[25] A. ANTOS, C. SZEPESVÁRI, R. MUNOS. *Fitted Q-iteration in Continuous Action-space MDPs*, in "Proc. Neural Information Processing Systems (NIPS)", to appear, December 2007.

[26] A. ANTOS, C. SZEPESVÁRI, R. MUNOS. *Value-Iteration based Fitted Policy Iteration: learning with a single trajectory*, in "IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning", 2007, p. 330–337.

[27] J.-Y. AUDIBERT, R. MUNOS, C. SZEPESVÁRI. *Tuning Bandit Algorithms in Stochastic Environments*, in "Proc. 18th International Conference on Algorithmic Learning Theory (ALT)", M. HUTTER, R. SERVEDIO, E. TAKIMOTO (editors), Lecture Notes in Artificial Intelligence, vol. 4754, Springer, October 2007, p. 150–165.

[28] A. BROVELLI, P.-A. COQUELIN, D. BOUSSAOUD. *Estimating the hidden learning representations*, in "Proc. of Neurocomp, Paris", 2007.

[29] T. BRÉHARD, M. DAVY. *Etude de l'influence de la modelisation sur les performances du filtre particulaire Rao-Blackwellise*, in "Proc. colloque du Groupe de recherche et d'études du traitement du signal et des images (GRETSI), Troyes, France", September 2007, -.

[30] F. CARON, M. DAVY, A. DOUCET. *Generalized Polya Urn for Time-varying Dirichlet Process Mixtures*, in "Proc. 23rd Conference on Uncertainty in Artificial Intelligence (UAI), Vancouver, Canada", July 2007, p. 33–40.

[31] P.-A. COQUELIN, S. MARTIN, R. MUNOS. *A dynamic programming approach to viability problems*, in "IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning", April 2007, p. 178–184.

[32] P.-A. COQUELIN, R. MUNOS. *Bandit Algorithms for Tree Search*, in "Proc. 23rd Conference on Uncertainty in Artificial Intelligence (UAI), Vancouver, Canada", July 2007, p. 67–74.

[33] R. COULOM. *Computing Elo Ratings of Move Patterns in the Game of Go*, in "Proc. of the Computer Games Workshop, Amsterdam, The Netherlands", J. H. VAN DEN HERIK, M. WINANDS, J. UITERWIJK, M. SCHADD (editors), June 2007.

[34] R. COULOM. *Monte-Carlo Tree Search in Crazy Stone*, in "Proc. of the 12th Game Programming Workshop, Hakone, Japan", T. ITO, A. KISHIMOTO (editors), invited conference, Information Processing Society of Japan, November 2007.

[35] M. DAVY, J.-Y. TOURNERET. *Classification Bayesienne Supervisée par Processus de Dirichlet*, in "Proc. colloque du Groupe de recherche et d'études du traitement du signal et des images (GRETSI), Troyes, France", September 2007.

[36] F. DESOBRY, M. DAVY, W. FIZTGERALD. *Density Kernels on Unordered Sets for Kernel-Based Signal Processing*, in "IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, USA", vol. 2, April 2007, p. II-417–II-420.

[37] E. DUFLOS, M. DE VILMORIN, P. VANHEEGHE. *Time Allocation of a Set of Radars in a Multitarget Environment*, in "Proc. of the FUSION Conference, Quebec (Canada)", INTERNATIONAL SOCIETY ON INFORMATION FUSION (editor), July 2007.

[38] S. GELLY, Y. WANG, R. MUNOS, O. TEYTAUD. *Modification of UCT with Patterns in Monte-Carlo Go*, in "IEEE International Symposium on Computational Intelligence and Game", April 2007, p. 175–182.

[39] T. HUGUERRE, E. DUFLOS, T. BRÉHARD, P. VANHEEGHE. *An Optimal Detection Strategy for ESA Radars*, in "Proc. of the COGnitive systems with Interactive Sensors Conference", Société de l'Electricité, de l'Electronique et des Technologies de l'Information et de la Communication, November 2007.

[40] E. JACKSON, M. DAVY, A. DOUCET, W. FITZGERALD. *Bayesian Unsupervised Signal Classification By Dirichlet Process Mixtures of Gaussian Processes*, in "IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP), Honolulu, USA", vol. 3, April 2007, p. III-1077–III-1080.

[41] R. JAULMES, J. PINEAU, D. PRECUP. *A formal framework for robot learning and control under model uncertainty*, in "Proc. International Conference on Robotics and Automation", 2007.

[42] M. LAMBERT, R. JAULMES, A. GODIN, E. MOLINÉ, D. DUFOURD. *A methodology for assessing robot autonomous functionalities*, in "Proc. Intelligent Autonomous Vehicle conference", 2007.

[43] M. LOTH, M. DAVY, P. PREUX. *Sparse temporal difference learning using LASSO*, in "IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning", April 2007, p. 352–359.

[44] M. LOTH, P. PREUX, M. DAVY. *A Unified View of TD Algorithms; Introducing Full-Gradient TD and Equi-Gradient Descent TD*, in "Proc. 15th European Symposium on Artificial Neural Networks (ESANN)", April 2007, p. 289–294.

[45] F. NAHIMANA, E. DUFLOS, J. MARAIS. *A Jump Markov System for modelling a realistic error model depending on satellite reception state in urban environment*, in "Proc. of the Institute of Navigation - Global Navigation Satellite System (ION GNSS)", September 2007.

[46] A. RABAOUI, M. DAVY, S. ROSSIGNOL, Z. LACHIRI, N. ELLOUZE. *Sélection des paramètres acoustiques pour la classification des sons environnementaux avec des SVMs mono-classe*, in "Proc. colloque du Groupe de recherche et d'études du traitement du signal et des images (GRETSI)", September 2007, p. 173-176.

[47] A. RAKOTOMAMOJY, M. DAVY. *One-class SVM regularization path and comparison with alpha seeding*, in "Proc. 15th European Symposium on Artificial Neural Networks (ESANN), Brugge, Belgium", April 2007, p. 271–276.

[48] S. ROSSIGNOL, M. DAVY. *Détection de ruptures à l'aide des « SVM 1 classe » pour la segmentation des signaux sonores musicaux*, in "Proc. colloque du Groupe de recherche et d'études du traitement du signal et des images (GRETSI)", September 2007, p. 165-168.

[49] O. TEYTAUD, S. GELLY, J. MARY. *Active learning in regression, with application to stochastic dynamic programming.*, in "Proc. of the Conference d'Apprentissage Automatique (CAP)", 2007.

[50] U. VON LUXBURG, S. BUBECK, S. JEGELKA, M. KAUFMANN. *Consistent Minimization of Clustering Objective Functions*, in "Proc. Neural Information Processing Systems", to appear, December 2007.

Internal Reports

[51] J.-Y. AUDIBERT, R. MUNOS, CS. SZEPESVÁRI. *Variance estimates and exploration function in multi-armed bandit*, Research Report 07-31, Certis - Ecole des Ponts, 2007.

[52] T. BRÉHARD, P.-A. COQUELIN, E. DUFLOS. *Optimal Policies Search for Sensor Management : Application to the AESA Radar*, Research Report, n^{o} 6361, INRIA, 11 2007, https://hal.inria.fr/inria-00188292.

[53] S. BUBECK, U. VON LUXBURG. *Overfitting of Clustering and how to avoid it*, Technical report, November 2007.

[54] P.-A. COQUELIN, R. DEGUEST, R. MUNOS. *Numerical methods for sensitivity analysis of Feynman-Kac models*, Technical report, INRIA, 2007, http://hal.inria.fr/inria-00125427.

[55] S. GIRGIN, P. PREUX. *Feature Discovery in Reinforcement Learning using Genetic Programming*, Research Report, n^{o} 6358, INRIA, 11 2007, https://hal.inria.fr/inria-00187997.

[56] M. MOYART. *Reconnaissance vocale – Rapport de Stage*, Technical report, n^{o} ISEN N4 P49 2006/2007, INRIA Futurs, September 2007.

[57] S. ROSSIGNOL, M. DAVY. *Méthodes à noyaux pour la reconnaissance de parole – Second rapport d'avancement*, Deliverable, Technical report, January 2007.

[58] S. ROSSIGNOL, M. DAVY. *Rapport de recherche final et manuel d'utilisation des programmes*, Deliverable, Technical report, January 2007.

[59] S. ROSSIGNOL. *Detection and Characterization of loudspeakers flaws in the production line*, Technical report, INRIA Futurs, 2007.

Miscellaneous

[60] A. CHOUIA. *Bandit Algorithms for Tree Search with Uncertainty*, Technical report, Université Paris Dauphine, 2007.

[61] J.-F. HREN. *Algorithmes d'exploration pour l'optimisation de fonctions bruitées*, Technical report, Université Lille 1, 2007.

[62] T. HUGUERRE. *Apprentissage par Renforcement pour la gestion de Systèmes Multicapteurs*, Technical report, Université d'Artois, 2007.

[63] R. JAULMES, E. MOLINÉ, B. GRAVIER. *HNG : Une architecture de contrôle hybride, générique et distribuée, pour la robotique à autonomie ajustable*, (submitted), 2007.

[64] R. MAITREPIERRE. *Apprentissage par renforcement d'une stratégie de jeu pour le poker*, Technical report, Université Lille 1, 2007.

[65] R. MAÎTREPIERRE, J. MARY, R. MUNOS. *Reinforcement Learning in Poker*, poster at the AAAI, 2007.

[66] P. PREUX, S. DELEPOULLE, R. COULOM. *Prise de décision séquentielle*, Revue d'Intelligence Artificielle, vol. 1, n^{o} 1, January 2007.

[67] A. RABAOUI, M. DAVY, S. ROSSIGNOL, Z. LACHIRI, N. ELLOUZE. *Using One-Class SVMs and Wavelets for Audio Surveillance Systems*, January 2007.

[68] Y. WANG. *Bandit Algorithms for Tree Search with Uncertainty*, Technical report, ENS Cachan, 2007.

References in notes

[69] A. DOUCET, N. DE FREITAS, N. GORDON (editors). *Sequential Monte Carlo Methods in Practice*, Springer, 2001.

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

[71] C. BARRERA-ESTEVE, F. BERGERET, E. GOBET, A. MEZIOU, R. MUNOS, D. REBOUL-SALZE. *Numerical methods for the pricing of Swing options: a stochastic control approach*, in "Methodology and Computing in Applied Probability", vol. 8, 2006, p. 517-540.

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

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

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

[75] O. BOKANOWSKI, S. MARTIN, R. MUNOS, H. ZIDANI. *An anti-diffusive scheme for viability problems*, in "Applied Numerical Mathematics, special issue on Numerical methods for viscosity solutions and applications", vol. 45-9, 2006, p. 1147-1162.

[76] B. EFRON, T. HASTIE, I. JOHNSTONE, R. TIBSHIRANI. *Least Angle Regression*, in "Annals of Statistics", vol. 32, n^{o} 2, 2004, p. 407–499.

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

[78] S. GELLY, Y. WANG, R. MUNOS, O. TEYTAUD. *Modification of UCT with Patterns in Monte-Carlo Go*, Technical report, n^{o} RR-6062, 2006, http://hal.inria.fr/inria-00117266.

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

[80] L. KOCSIS, CS. SZEPESVÁRI. *Bandit based Monte-Carlo Planning*, in "Proc. European Conference on Machine Learning (ECML)", Lecture Notes in Computer Science, vol. 4212, Springer-Verlag, 2006, p. 282–293.

[81] C. KREUCHER, A. HERO.. *Non-myopic Approaches to Scheduling Agile Sensors for Multitarget Detection, Tracking, and Identification*, in "The Proceedings of the 2005 IEEE Conference on Acoustics, Speech, and Signal Processing (ICASSP) Special Section on Advances in Waveform Agile Sensor Processing, volume V, March 18 - 23", 2005, p. 885–888, http://www.eecs.umich.edu/~hero/Preprints/2005ICASSP_a.pdf.

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

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

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

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

[86] 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.