INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team SequeL

Sequential Learning

Lille - Nord Europe

THEME COG

Table of contents

SEQUEL *is a joint project with the* LIFL *(UMR 8022 of CNRS, and University of Lille 1, and University of Lille 3) and the* LAGIS *(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]
Manuel Davy [Researcher (CR) CNRS, currently mostly with the start-up Vekia, HdR]
Mohammad Ghavamzadeh [Researcher (CR) INRIA, arrives on Sep 1st, 2008]
Daniil Ryabko [Researcher (CR) INRIA, arrives on Dec 1st, 2007]

Faculty Member

Philippe Preux [Team leader, Professor, Université de Lille, secondment at the INRIA, HdR]
Emmanuel Daucé [Assistant Professor, École Centrale de Marseille, partial secondment in SEQUEL since Sep 1st, 2008]
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

Antoine Labitte [Assistant Engineer, until Sep 30th, 2008]
Tony Ducrocq [Assistant Engineer, since Oct 1st, 2008]

PhD Student

Pierre-Arnaud Coquelin [École Polytechnique, since Oct., 2005, currently mostly with the start-up Vekia]
Robin Jaulmes [DGA Grant, since Oct., 2006]
Manuel Loth [INRIA-Région Nord-pas-de-calais Grant, since Oct., 2006]
Jean-François Hren [MENESR Grant, since Oct., 2007]
Raphaël Maîtrepierre [MENESR Grant, since Oct., 2007]
Sébastien Bubeck [ENS Grant, since Oct., 2007]
Odalric-Ambrym Maillard [ENS Grant, since Oct., 2008]
Nicolas Viandier [INRETS, since Oct., 2007]
Emmanuel Delande [DGA, since Nov., 2008]

Post-Doctoral Fellow

Sertan Girgin [INRIA, left on Jul 31st, 2008]
Alessandro Lazaric [INRIA, begins on Jul. 1st, 2008]
Hachem Kadri [CNRS, begins on Nov. 1st, 2008]
Djalel Mazouni [INRIA until July, ATER since Sep. 2008]

Administrative Assistant

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

Other

Odalric-Ambrym Maillard [Master 2 internship, Apr to Sep 2007]
Aurélie Pruvost [Master 1 internship, May to Aug, 2008]
Philippe Van Eerdenbrugghe [Master 1 internship, Feb to Jun, 2008]
Réginald N'Guyama [Master 1 internship, Feb to Jun, 2008]

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 [72]). 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.

2.2. Highlight of the year

In 2008, we would like to highlight the following three events.

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

This year again, we have had strong results on the game of Go, with Rémi Coulom's Crazy Stone software being the first program in the world to defeat a human expert, with a handicap of only 8 stones, and in December, the same expert with a handicap of only 7 stones. Crazy Stone also won the University of Electro-Communications Cup. (More information in section 6.2.5.2.1.)

We have organized the 8$^{\text{th}}$ European Workshop on Reinforcement Learning. This year issue has witnessed a totally renewed organization, and subsequently, a yet unseen worldwide participation, with major researchers in the field participating at the event. As many attendees have argued, this issue of the workshop has served as a *de facto* first international conference on reinforcement learning. (More information in section 8.3.1.2.)

Finally, Rémi Munos's ANR EXPLORA proposal has been accepted. (More information in section 8.2.2.)

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

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

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 [67].

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

[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.

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

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

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

where $\mathbb{E}$ is the expectation operator and $\gamma \in (0, 1)$ is the discount factor. This value function V^π gives an evaluation of the performance of a given policy π. Other functionals of the sequence of future rewards may be considered, such as the undiscounted reward (see the stochastic shortest path problems [55]) 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) [53], 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 [54]. 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 ([74]):

- 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 [68], 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.* [51] 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

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

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 [61]. 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 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 [60] 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

Keywords: *ambiant intelligence*, *automatic transcription of speech*, *civil engineering*, *customer affluence modeling*, *environment*, *games*, *multimedia*, *optimization in non deterministic environment*, *optimization in time-varying environment*, *optimization in uncertain domain*, *sensor localization*, *transportation systems*.

SEQUEL aims at solving problems of prediction, as well as problems of optimal and adaptive control. As such, the application domains are very numerous.

[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 [71].

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

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

- game softwares: in the 1990's, RL has been the basis of a very successful Backgammon program, TD-Gammon [73] that learned to play at an expert level by basically playing a very marge 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 [66]
- 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 [70], [69] 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.

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 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

Keywords: *Go software*.

Participant: Rémi Coulom [correspondent].

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

Being a research tool related to strong worldwide competition, Crazy Stone is no longer freely available.

6. New Results

6.1. Introduction

New results are organized is the following sections:

1. decision under uncertainty,
2. machine learning,
3. signal processing.

6.2. Decision under uncertainty

Keywords: *LARS*, *Monte-Carlo estimation*, *dynamic programming*, *electronic scanned radar*, *exploration-exploitation trade-off*, *feature discovery*, *l1-regularization*, *learning the representation of data applications*, *multi-arm bandit*, *non parametric Bayesian learning*, *non parametric function approximation*, *performance bound*, *policy search*, *probability of detection*, *radar*, *reinforcement learning*, *scheduling*, *sensor management problem*, *value function approximation*.

Participants: Sébastien Bubeck, Pierre-Arnaud Coquelin, Rémi Coulom, Emmanuel Duflos, Mohammad Ghavamzadeh, Sertan Girgin, Jean-François Hren, Manuel Loth, Raphaël Maîtrepierre, Jérémie Mary, Djalel Mazouni, Rémi Munos, Philippe Preux, Daniil Ryabko, Philippe Vanheeghe.

6.2.1. Reinforcement Learning

6.2.1.1. Learnability in Reinforcement Learning in non-Markovian environments

We have addressed the problem of reinforcement learning in arbitrary environments, not restricted to (PO)MDPs. The general problem is as follows: an agent is interacting with an unknown environment, and is occasionally rewarded for its behaviour. It seeks to maximize its cumulative rewards. The first problem that arises in such a general setting is that the environment may not forgive first wrong (or exploratory) actions of the agent. In other words, the agent may fall into a pit from which it will never be able to get out, to explore other parts of the environment, and to learn how to get the rewards. We formalize the problem and find a characterization [23] of environments that "forgive" initial wrong actions and allow the agent to learn sufficiently fast to be able to find the way to maximize its rewards.

6.2.1.2. Function approximation

6.2.1.2.1. Regularization in Reinforcement Learning

In [30], we studied how to add L_2-regularization to value function approximation in RL. The problem setting is to find a good policy in a batch or active learning scenario for infinite-horizon expected total discounted reward Markovian decision problems with continuous state and finite action spaces. We developed two novel policy evaluation algorithms by adding L_2-regularization to two widely-used policy evaluation methods in RL: Bellman residual minimization (BRM) [75], [52] and least-squares temporal difference learning (LSTD) [56]. We showed how our algorithms can be implemented efficiently when the value-function approximator belongs to a reproducing kernel Hilbert space. We also proved finite-sample performance bounds for our algorithms. In particular, we showed that they are able to achieve a rate that is as good as the corresponding regression rate when the value functions belong to a known smoothness class. We further showed that this rate of convergence carries through to the performance of a policy found by running policy iteration with our regularized policy evaluation methods. The results indicate that from the point of view of convergence rates RL is not harder than regression estimation.

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

We have worked further the Equi-Gradient Temporal Difference algorithm, designed originally in 2006 [64]. This has led to a new kernelized LARS-like algorithm, based on an l_1 regularization, which builds the regularization path. The striking new feature of this algorithm is that it automatically optimizes the hyper-parameters of the kernels, thus being able to deal with an infinite number of features. The algorithm is named the "Equi-correlated network" [65]. It can also be seen as a one hidden layer neural network, in which the hidden layer is growing and shrinking, according to the flow of data. This algorithm has been tested on regression tasks, as well as in the approximate dynamic programming setting.

6.2.1.2.3. Function approximation and representation learning

We have devoted a large amount of work to the automatic feature discovery problem, in reinforcement learning. We argue for a very strong link between this issue and non parametric function approximation. We have investigated various approaches for that, using genetic programming [31], cascade-correlation network [33], [18], [32], and the Equi-Correlation Network. We have also published the first study in which feature discovery is embedded in a direct policy search approach of reinforcement learning [18].

6.2.2. Policy gradient estimation in POMDPs

With Pierre-Arnaud Coquelin (Vekia) and Romain Deguest (Colombia University), we considered a Partially Observable Markov Decision Problem where decisions are based on a Particle Filter for estimating the belief state given past observations. We developped a policy gradient approach for parameterized policy optimization based on a sensitivity analysis of the performance measure with respect to the parameters of the policy (see [28]).

6.2.3. Exploration vs. exploitation

The exploration/exploitation balance problem is a long-standing issue in artificial intelligence. This problem has the beginning of a very strong activity in SEQUEL, in relation with sequential decision problems, as well as in the bandit framework.

6.2.3.1. Bandits

- **Many-armed bandits** In collaboration with Yizao Wang (University of Michigan), and Jean-Yves Audibert (Ecole des Ponts), R. Munos considered the so-called *many-armed bandit problem* which is a multi-armed bandit problem where the number of arms is larger than the possible number of experiments. We made a stochastic assumption on the mean-reward of a new selected arm which characterizes its probability of being a near-optimal arm. We derived algorithms based on upper-confidence-bounds applied to a restricted set of randomly selected arms and provide upper-bounds on the resulting expected regret. We also derive a lower-bound which matches (up to a logarithmic factor) the upper-bound in some cases (see [43]).
- **Hierarchical Optimistic Optimization** With Sébastien Bubeck, Gilles Stoltz, and Csaba Szepesvári, R. Munos analyzed a global optimization algorithm, based on bandit algorithms on measurable spaces, whose rate of convergence (in terms of regret per round) may be as small as $O(1/sqrt(n))$ (when the smoothness of the function around its maxima is known) where n is the number of evaluation of the function, independently of the dimension of the space (see [27]).

6.2.4. Optimistic Planning

With Jean-Francois Hren, R. Munos considered the question: given finite computational resources (*e.g.*, CPU time), which may not be known ahead of time, what is the best way to explore the set of all possible sequences of decisions, such that once all resources have been used, the algorithm would be able to propose an action (or a sequence of actions) whose performance is as close as possible to optimality? We proposed an analyzed an algorithm, *optimistic planning*, which explores first the most promising sequences (see [34]).

6.2.5. Applications

6.2.5.1. The sensor management problem

In the sensor management problem, we continue the line of research along radar management and parameterized policy search.

We deepened the approach consisting in deriving optimal parametrized policies based on a stochastic gradient estimation. We assumed in this work that it is possible to learn the optimal policy off-line (in simulation) using models of the environement and of the sensor(s). The learned policy can then be used to manage the sensor(s). In order to approximate the gradient in a stochastic context, we introduce a new method to approximate the gradient, based on Infinitesimal Perturbation Approximation (IPA). The effectiveness of this general framework has been illustrated by the management of an Electronically Scanned Array Radar (see [26]).

6.2.5.2. Applications to games

6.2.5.2.1. The game of Go

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

Rémi Coulom continued developing the top-level go-playing program Crazy Stone. Crazy Stone has dominated international computer-go tournaments for one year, after winning the first UEC Cup in Tokyo (December, 2007), it won the tournament of the European Go Congress (July, 2008), and won a 8-stone handicap game against professional player Kaori Aoba in Japan, during the FIT'2008 conference (September, 2008). Cray Stone then won the University of Electro-Communications Cup, Japan (December 2008), and, a second time, defeated Kaori Aoba with a handicap of only 7 stones this time. These victories against a human expert, with a moderate ap for the human, constitute a major milestone, since no go-playing program had ever won a game against a professional player with such a handicaps before.

The new automated planning techniques pioneered in Crazy Stone are based on Monte-Carlo tree search. They are now studied by several major research groups all around the world. They have been applied successfully to many domains, so this breakthrough reaches far beyong the game of Go.

6.2.5.2.2. The game of Poker

In 2008, we further worked on our Artificial Poker Player, mainly on the opponent modelisation in limit game using bandits techniques. We published some results at ECAI'08 [35].

Here, the main idea is to use UCB like algorithms to construct a meta strategy from several basis stategies. Each of the basis strategies is not a good poker player, but the combinaison of them (uniformly) is close to a Nash Equilibrium, so the resulting meta strategy is not so bad. After this initial stage, UCB starts to identify some strategies which should be played more than uniformly, so the meta strategy starts to adapt to its opponent. One of the advantages of this technique is the fast adaptation to the identified weaknesses and the ability to identify changes in the opponent play (using changepoint detection).

This idea has been implemented in Brennus (in C++) by Raphaël Maîtrepierre. J. Mary also worked with a Master 1 student on an Ajax web interface to play poker online againts Brennus.

From a more theoretical point of view, we also started to study the no-limit case. After an invitation of Martin Zinkevitch (chair of the poker bot competetion at AAAI07), we decided to work on the computation of Nash Equilibrium. The reason is that in two players competitions, such bots are more efficient than opponent modelling ones, even if human players find them boring, and they don't achieve maximisation of their gains.

We focus on the fast computation of Nash Equilibrium by regret minimisation in the case of a continous control. The idea is to use a parametrised distribution over the action states. At this time we begin to have results on a simplified game of poker (pure poker, with only one round of betting)

6.3. Machine Learning

Keywords: *Foundations of machine learning*, *prediction*, *statistical inference*.

Participant: Daniil Ryabko.

6.3.1. Sequence prediction in the most general form.

We have proposed and started to address [24], [39] the question of sequence prediction in what is perhaps its most general form: under what conditions on the class of environments there exists a predictor that predicts every environment in the class. The first results obtained in this direction already generalize such diverse approaches to sequence prediction as predicting finite-memory processes, predicting computable processes, and "merging of opinions". In particular, we have shown that if a set of environments is separable with respect to the expected average Kullback-Leibler divergence, then there exists a predictor that predicts (asymptotically) every environment in this set. Finite-step performance guarantees have also been obtained. These results outline a very promising direction for future research, since obtaining general conditions that guarantee the existence of a predictor is a first step to constructing predictors for mew application domains for which currently efficient predictors are not known.

6.3.2. Statistical inference.

Sequential inference is not restricted to prediction and planning. Classical problems of sequential inference include hypothesis testing, testing for homogeneity or component independence of time series, and the change point problem. We have addressed these problems in the setting when the stochastic processes are stationary ergodic: an assumption that is much more general than the assumptions considered in the literature on these problems before (which are: i.i.d. processes, finite-memory processes, or those that satisfy certain mixing conditions). We have constructed [41], [40], [25] change point estimates, identity tests and process classifiers, that are asymptotically consistent for arbitrary stationary ergodic processes.

6.4. Signal analysis and processing

Keywords: *geo-localization*, *global navigation satellite system*, *sensor localization*, *urban canyons*.

Participants: Emmanuel Duflos, Philippe Vanheeghe, Nicolas Viandier.

6.4.1. Localization

6.4.1.1. 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. In 2008 we have obtained a PICS (International Project for Scientific Cooperation) from the CNRS 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.

6.4.1.2. Accurate Localization using Satellites in Urban Canyons

This work is done in collaboration with Juliette Marais, junior researcher at INRETS, Fleury Donnay Nahimana and Nicolas Viandier. Nicolas Viandier and Fleury Donnay Nahimana are both PhD students 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 axis 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.

The real originality of this work is to search for solutions to increase the localization precision without adding new sensors. We focus on signal processing enhancement to estimate sequentially the noise on pseudo-distances. We thus propose new noise models allowing to take into account the non gaussian characteristics. The simplest model is a gaussian mixture and the most complex one, still under analysis, is an infinite Dirichlet Process Mixture. We also take into account into the position estimation algorithm that the reception state of each satellite varies with respect to time. This state of reception is also estimated using a Dirichlet distribution. Particle Filtering is used to implement the estimator. The method a validated using real data ([36]). This work is based on the theoric approach developped by Caron in [57] and [15].

Donnay Fleury Nahimana has participated to the Young European Arena of Research competition (YEAR 2008), Ljubjana, Slovenie, Avril 2008. He was awarded the gold medal for his works in the Transport Category[4].

In order to make simulations in a realistic way, we have built and implemented a realistic 3D model of a part of a district of Lille. This model has been developped by Reginald N'Guyama, a surveyor student, during a 6 month internship.

7. Contracts and Grants with Industry

7.1. Contracts and Grants with Industry

7.1.1. Contract with Vekia Innovation

Vekia Innovation is the new name of the startup we created in 2007, Predict & Control.

Jérémie Mary started a 20 K-euros contract with Vekia Innovation startup about prediction of phone calls for phone centers. This is planned to be a 6 months collaboration. It consists in working on several aspects with the necessity to work on three points :

- Long term prediction: how many operators should be to hired;
- Construction of operator planning, depending on the abilities of the operators, and on their special needs (type of contract, disponibilities).
- On-line load repartition depending of the current load, and of the competences of the operators.

[4] See http://ec.europa.eu/research/transport/news/article_6928_en.html

From a scientific point of view, there are several aspects:

- statistical modeling of the work of the operators. This modeling should take into account calendar events, the type of work being asked (phone, mails, emails), special events (like discounts, and ad-campaigns). The construction and the validation of this modeling will use huge amounts of collected data,
- Use of this model with efficient algorithms (computation must be fast) in order to construct plannings. We also want to work on the replanning within a day, according to the current charge.

This work will lead to a toolbox written in Python to handle this kind of problems. A SequeL young engineer (Tony Ducrocq) is affected for six monthes to produce the code on this study.

8. Other Grants and Activities

8.1. Regional activities

8.1.1. Ambiant intelligence campus (Campus Inteligence Ambiante)

Participants: Emmanuel Duflos, Philippe Vanheeghe, Emmanuel Delande, Jérémie Mary, Rémi Munos, Philippe Preux.

SequeL is also taking part in the "Contrat de Plan État-Région" project "Campus Intelligence Ambiante" (CIA). Sequel participation deals with the study of adaptive system in the domain of ambiant intelligence.

8.1.2. Pôle de Compétitivité PICOM

Participants: Jérémie Mary, Philippe Preux.

SequeL is taking part in a project named "Ubiquitous Virtual Seller" of the Pôle de Compétitivité "Industrie du Commerce".

This 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 sell strategy, ... The proposal is currently under expertize prior funding.

8.2. National activities

8.2.1. DGA / Thalès

Participants: Emmanuel Duflos, Philippe Vanheeghe, Emmanuel Delande.

A DGA PhD grant has been accepted for a new PhD student (E. Delande) who joins on Nov. 1st, 2008. He will work on the optimal sensor management problem; he will also work in collaboration with M. Prenat of Thalès with whom we have a long standing relationship dealing with this problem.

8.2.2. ANR EXPLORA

Participants: Sébastien Bubeck, Mohammad Ghavamzadeh, Manuel Loth, Jérémie Mary, Rémi Munos, Philippe Preux, Daniil Ryabko.

Keywords: resource allocation, numerical simulation, exploration / exploitation dilemma, regret minimization, bandit algorithms, population of bandits, learning from experts, tree and graph search, sequential decision making under uncertainty, optimization, game theory, reinforcement learning, multi-agent learning.

In 2008, Rémi Munos has managed the proposal for a new ANR project, named EXPLORA. EXPLORA means **EXPLO**ration – **EXPLO**itation for efficient **R**esource **A**llocation. Applications to optimization, control, learning, and games.

The participants are affiliated to: INRIA-Lille Nord Europe (SEQUEL), INRIA-Saclay (TAO), HEC Paris (GREGHEC), Les Ponts (CERTIS), Paris 5 (CRIP5), Paris 8 (LAMSADE).

The proposal 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.

Our contributions will be theoretical (convergence issues, regret bounds), algorithmic (the design of algorithms adapting automatically to the unknown underlying structure of the problem), and numerical (we aim at solving real world large scale problems).

This is a fundamental research project. It brings together academic partners covering a broad spectrum of expertise, from the most theoretical aspects (statistics, optimal control, game theory, statistical learning, decision theory, stochastic processes) to more applied and experimental skills (game programming, parallel computing).

8.2.3. ANR Kernsig

Participants: Manuel Davy, Emmanuel Duflos, Hachem Kadri.

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

A Post-doc has been recruited on Nov. 1st, 2008 by E. Duflos to work on incremental functional regression.

8.2.4. ARC CODA

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

This is a two years ARC project (2007 - 2008) named CODA (for "Optimal control of an anaerobic digestor") done in collaboration with the INRA Laboratory LBE in Narbonne, the INRIA project-team COMORE in Sophia-Antipolis, and the spin-off Naskeo Environment.

A post-doc fellow (Djalel Mazouni) has been hired for one year in 2007/2008.

Anaerobic digestion is a biological process in which microorganisms break down biodegradable material in the absence of oxygen. It is widely used to treat wastewater sludges and organic wastes because it provides volume and mass reduction of the input material, as well as production of biogas (such as methane), a renewable energy source.

The complex digestion process makes the dynamics unstable which motivates the need to develop efficient methods for stabilizing the dynamical reactions while trying to optimize some performance measure (such as the biogas production).

The goal of this project consists in designing adaptive control methods for an anaerobic digestor from the approximate knowledge of the state dynamics, and the partial information of observed data coming from a real-world reactor.

Several approaches for solving this partially observable Markov decision problem have been developed by two PhD students, Pierre-Arnaud Coquelin [59], [58] using a sensitivity analysis combined with particle filtering approach, and Robin Jaulmes [63], [62] 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

8.3.1. Scientific event organizations

SEQUEL has organized two important scientific events in 2008, namely the 8th European Workshop on Reinforcement Learning, and the 10th Machine Learning Summer School.

8.3.1.1. Machine Learning Summer School

The SequeL team organised the 10^{th} Machine Learning Summer School 2008 (MLSS'08) from 1^{st} to 15^{th} September at the "Ile de Ré". Supervision of the organisation was done by Manuel Davy and Jérémie Mary.

MLSS is a major event of the Machine Learning community. It combines theory from areas as diverse as Statistics, Mathematics, Engineering, and Information Technology with many practical and relevant real life applications. The aim of the summer school is to cover the entire spectrum from theory to practice. It is mainly targeted at research students, academics, and IT professionals from all over the world.

There is a selection process both for students and speakers. This year, the number of applications was around 200 (7% Undergraduates, 13% Masters, 67% PhDs, 3% Post Doc, 7% Academics, 3% Industrials) from 38 countries.

The total number of accepted participants was 102 from 34 countries (25 from Germany, 21 from France, 15 from Italia, 10 from Netherlands, 16 from UK, 11 from the US, 10 from Switzerland, 7 from India, 7 from Canada, and many from others countries but with less than 5 participants). Note that we report the countries of studies, not the countries of citizenship (China would be more representated if we do so).

They received lectures for two weeks (8 hours a day including 2h of practical sessions) from 11 speakers from University of Harvard, Alberta, Bristish Columbia, Waterloo, Berkeley, Bordeaux, New York, Lille, INSA Rouen and ENS Ulm.

8.3.1.2. 8th European Workshop on Reinforcement Learning

As the eighth in the series, the event is a distant follower of the first workshop that was held in Brussels, Belgium, in 1994. Since then, with an average bi-annual frequency, EWRL has gathered mostly European researchers, aiming at being a very open forum dealing with the current researches in reinforcement learning. While keeping this openness in the organization, we have thought time has come to make EWRL something bigger; we have tried to gather the world-wide community, to have a great scientific event entirely dedicated to the research in reinforcement learning. Still, we have wished to keep it wide open to students, PhD students, but also, future PhD students, and let them hear and meet some of the top researchers in the field.

The workshop itself has gathered 105 participants during 4 full days. Among attendees, 44 PhD students, 10 post-doctoral fellows, 39 academics, 8 undergarduate students, and 4 researchers working in private companies. While 47 % of the participants were French, 16 % were Belgian, 9 % were Canadians, 7 % were German, 7 % were Netherlanders, and other participants came from the USA, China, South-Africa, Israel, and other European countries; overall 13 different countries were represented.

The program of the workshop has also included three invited speakers, namely Richard S. Sutton, from the University of Alberta in Edmonton, Canada, Dimitri Bertsekas, from the Massachussetts Institute of Technology, USA, and Jan Peters, from the Max-Planck Institute, Tübingen, Germany.

A restricted number of travel grants was available for students. They have been granted after a selection process based on the submission of a resume. 6 students have benefited from these grants.

Wishing to be open to any researcher, registration to EWRL was free. The organization was funded by SEQUEL, the "Collège Doctoral Européen Lille Nord-Pas de Calais" who has funded the travel grants, the INRIA Research Center "Lille-Nord Europe", and the computer science laboratory (LIFL) of the University of Lille.

After the workshop, a post-selection of 21 papers are published by Springer, in a Lecture Notes in Artificial Intelligence volume [47].

8.3.1.3. NIPS workshop

Along with Yaakov Engel, Shie Mannor (Assistant Professor of Electrical and Computer Engineering at McGill University, and Canada Research Chair in Machine Learning), and Pascal Poupart (Assistant Professor at the School of Computer Science at the University of Waterloo), M. Ghavamzadeh is organizing a one-day workshop at NIPS 2008 on "Model Uncertainty and Risk in Reinforcement Learning", Whistler, British Columbia, Canada, December 13, 2008.

8.3.1.4. Special session at fusion

E. Duflos and Ph. Vanheeghe have organized a special session dedicated to sensor management at the Fusion conference, held in July, in Köln, Germany.

8.3.2. Programme Interdisciplinaire de Coopération Scientifique

A "Programme Interdisciplinaire de Coopération Scientifique" (PICS) has been accepted for 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.

8.3.3. Associate team

On Jan 1st, 2008 was created an "Équipe Associée" with University of Alberta at Edmonton (Canada), with Richard Sutton's group.

Several visits have been made under this funding, by Rémi Munos, Sébastien Bubeck, Mohammad Ghavamzadeh, Csaba Szepesvari, and Richard Sutton. Richard Sutton was invited to give an invited talk at EWRL'2008 (see sec. 8.3.1.2). This has resulted in work, which has begun to be published [27].

8.4. Visits and invitations

- Rémi Munos has visited the University of Alberta, at Edmonton, to work further with C. Szepesvári,
- Sébastien Bubeck has visited the University of Alberta, at Edmonton, to work further with C. Szepesvári.
- Philippe Preux was invited by B. Chaib-Draa, at the University of Laval, in Québec, Canada.

9. Dissemination

9.1. Scientific community animation

- Emmanuel Duflos is involved in the organization of the 5th Computational Engineering in Systems Applications conference which will be held in 2009, in South Korea.
- Emmanuel Duflos is reviewing submissions to the journals IEEE Signal Processing, IEEE Special Topics in Signal Processing, IEEE Transaction on Geosciences and Remote Sensing, Information Fusion, IEEE Transactions on Control Systems Technology, IEEE Transaction on Automatic Control and Journal of Applied Geophysics.
- Emmanuel Duflos is a member of the Fusion'2008 International Program Committee.
- Rémi Munos is reviewing papers for the following journals: Maths of Operations Research, Revue d'Intelligence Artificielle, IEEE Systems, Man and Cybernetics, Journal of Machine Learning Research

  He also reviewed papers for the following conferences: Neural Information Processing Systems 2008, International Conference on Machine Learning 2008, Conférence Francophone sur l'Apprentissage Automatique 2008, Conference on Learning Theory 2008, Uncertainty in Artificial Intelligence 2008.

- Rémi Munos was an invited speaker at:
  - Cours à l'Ecole de Printemps en Informatique Théorique sur l'apprentissage automatique (Porquerolle),
  - Workshop on Fast Reinforcement Learning (Barbados),
  - Séminaire du Xerox Research Centre Europe,
  - Journées Modélisation Aléatoire et Stochastique (Rennes),
  - Séminaire Apprentissage (Ecole Normale Supérieure),
  - Séminaire Apprentissage et Optimisation (LRI, Orsay),
  - Séminaire proba/stat à Bordeaux.
- Rémi Munos is a member of the PC Co-chair of the IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning, 2009.
- Rémi Munos has been elected as a member of the INRIA "Commission d'évaluation"; he has also reviewed a project for the ANR[5].
- Philippe Preux has reviewed papers for the journal "Autonomous Robotics"; he is a member of the program committee of, and has reviewed submissions for, IEEE Approximate Dynamic Programming and Reinforcement Learning 2009, "Extraction et Gestion des Connaissances" 2009.
- Rémi Munos and Philippe Preux are experts with the AERES[6].
- Philippe Preux also serves as a member of the "Jury Gilles Kahn 2008" which aims at awarding the "best" computer science PhD dissertation of the year.
- S. Girgin, M. Loth, R. Munos, Ph. Preux and D. Ryabko have organized all aspects of the 8th European Workshop on reinforcement Learning, including the edition of a volume in the Lecture Notes in Artificial Intelligence, series of Springer.
- Rémi Coulom was elected on the board of the International Computer Games Association for the position of "programmers representative", in 2008.
- Daniil Ryabko is a reviewer for the journals IEEE Trans. on Information Theory, Theoretical Computer Science, and the conferences COLT, ALT, ECML, UAI, and ICALP.

9.2. Teaching

We list the courses 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 of computer science at the University of Lille a class on reinforcement learning.
- Jérémie Mary and Rémi Coulom are teaching data mining in master at the University of Lille.

Otherwise, each of the 4 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.

[5] ANR stands for "National Research Agency" which funds national research projects.
[6] French national agency that evaluates research laboratories and university diplomas.

[2] 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", vol. 71, 2008, p. 89–129.

[3] J.-Y. AUDIBERT, R. MUNOS, C. SZEPESVÁRI. *Tuning Bandit Algorithms in Stochastic Environments*, in "Theoretical Computer Science", To appear, 2008.

[4] 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.

[5] 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.

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

[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", vol. 46, n^{o} 2, 2008, p. 541–561.

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

[10] 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.

[11] 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

[12] K. OTA. *Studies in Signal Processing for Robust Speech Recognition in Noisy and Reverberant Environment*, Ph. D. Thesis, 2008.

Articles in International Peer-Reviewed Journal

[13] 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", vol. 71, 2008, p. 89–129.

[14] J.-Y. AUDIBERT, R. MUNOS, C. SZEPESVÁRI. *Tuning Bandit Algorithms in Stochastic Environments*, in "Theoretical Computer Science", To appear, 2008.

[15] 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.

[16] F. CARON, B. RISTIC, E. DUFLOS, P. VANHEEGHE. *Least Committed Basic Belief Density Induced by a Multivariate Gaussian: Formulation with Applications*, in "International Journal of Approximate Reasoning", vol. 48, n^{o} 2, 2008, p. 419–436.

[17] B. FERGANI, M. DAVY, A. HOUACINE. *Speaker diarization using one-class support vector machines*, in "Speech Communication", vol. 50, n^{o} 5, May 2008, p. 355-365.

[18] S. GIRGIN, P. PREUX. *Basis Expansion in Natural Actor Critic Methods*, in "Lecture Notes in Artificial Intelligence (LNAI)", vol. 5323, June 2008, p. 111–124.

[19] A. JOHANSEN, A. DOUCET, M. DAVY. *Particle Methods for Maximum Likelihood Estimation in Latent Variable Models*, in "Statistics and Computing", vol. 18, n^{o} 1, March 2008, p. 47–57.

[20] 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", 2008.

[21] 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.

[22] R. MUNOS, C. SZEPESVÁRI. *Finite time bounds for sampling based fitted value iteration*, in "Journal of Machine Learning Research", vol. 9, 2008, p. 815–857.

[23] 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.

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

[25] D. RYABKO, J. SCHMIDHUBER. *Using Data Compressors to Construct Order Tests for Homogeneity and Component Independence*, in "Applied Mathematics Letters", (to appear), 2008.

International Peer-Reviewed Conference/Proceedings

[26] T. BREHARD, P.-A. COQUELIN, E. DUFLOS, P. VANHEEGHE. *Optimal policies search for sensor management : Application to the ESA radar*, in "Proceedings of the 11th International Conference on Information Fusion", 2008, p. 1–8.

[27] 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.

[28] P.-A. COQUELIN, R. DEGUEST, R. MUNOS. *Particle Filter-based Policy Gradient for POMDPs*, in "Proceedings of Advances in Neural Information Processing Systems", vol. 22, MIT Press, 2008.

[29] R. COULOM. *Whole-History Rating: A Bayesian Rating System for Players of Time-Varying Strength*, in "Proceedings of the 6th International Conference on Computer and Games, Beijing, China", H. J. VAN DEN HERIK, X. XU, Z. MA (editors), Lecture Notes in Computer Science, Springer, October 2008.

[30] A. M. FARAHMAND, M. GHAVAMZADEH, CS. SZEPESVÁRI, S. MANNOR. *Regularized Policy Iteration*, in "Proceedings of Advances in Neural Information Processing Systems", vol. 22, MIT Press, 2008.

[31] S. GIRGIN, P. PREUX. *Feature discovery in reinforcement learning using Genetic Programming*, in "Proc. 11th Euro-GP", M. O'NEILL, L. VANNESCHI, S. GUSTAFSON, A. E. ALCÁZAR, I. DE FALCO, A. D. CIOPPA, E. TARANTINO (editors), LNCS, this paper was nominated to receive the best paper award, vol. 4971, Springer, March 2008, p. 218–229.

[32] S. GIRGIN, P. PREUX. *Incremental Basis Function Expansion in Reinforcement Learning using Cascade-Correlation Networks*, in "Proc. International Conference on Machine Learning and Applications (ICML-A)", IEEE Press, December 2008, p. 75–82.

[33] S. GIRGIN, P. PREUX. *Incremental basis function expansion in reinforcement learning using cascade-correlation networks*, in "Proc. ECAI workshop ERLARS", N. SIEBEL, J. PAULI (editors), 2008.

[34] J.-F. HREN, R. MUNOS. *Optimistic Planning of Deterministic Systems*, in "Recent Advances in Reinforcement Learning", Lecture Notes in Artificial Intelligence, vol. 5323, Springer, 2008, p. 151–164.

[35] R. MAÎTREPIERRE, J. MARY, R. MUNOS. *Adaptative Play in Texas Hold'em poker.*, in "Proc. European Conference on Artificial Intelligence (ECAI)", 2008, p. 333–337.

[36] D. NAHIMANA, E. DUFLOS, J. MARAIS. *Reception state estimation of GNSS satellites in urban environment using particle filtering*, in "Proceedings of the 11th International Conference on Information Fusion", 2008, p. 1–5.

[37] K. OTA, E. DUFLOS, P. VANHEEGHE, M. YANAGIDA. *Reception state estimation of GNSS satellites in urban environment using particle filtering*, in "Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. ICASSP 2008", 2008, p. 1553–1556.

[38] K. OTA, E. DUFLOS, P. VANHEEGHE, M. YANAGIDA. *Speech recognition with speech density estimation by the Dirichlet Process Mixture*, in "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Las Vegas, Nevada, (USA)", 2008, p. 1553 - 1556.

[39] D. RYABKO. *Some Sufficient Conditions on an Arbitrary Class of Stochastic Processes for the Existence of a Predictor*, in "Proc. 19th International Conf. on Algorithmic Learning Theory (ALT'08), Budapest, Hungary", LNAI, Springer, Berlin, 2008.

[40] D. RYABKO. *Testing Statistical Hypotheses About Ergodic Processes*, in "Proc. 2008 IEEE Region 8 International Conference on Computational Technologies in Electrical and Electronics Engineering (SIBIRCON 2008), Novosibirsk, Russia", IEEE, 2008, p. 257–260.

[41] D. RYABKO, B. RYABKO. *On Hypotheses Testing for Ergodic Processes*, in "Proc. 2008 IEEE Information Theory Workshop, Porto, Portugal", IEEE, 2008, p. 281–283.

[42] N. VIANDIER, D. NAHIMANA, J. MARAIS, E. DUFLOS. *GNSS performance enhancement in urban environment based on pseudo-ranie error mode*, in "Proceedings of the IEEE/ION PLAN", May 2008, p. 377–382.

[43] Y. WANG, J.-Y. AUDIBERT, R. MUNOS. *Algorithms for Infinitely Many-Armed Bandits*, in "Proceedings of Advances in Neural Information Processing Systems", vol. 22, MIT Press, 2008.

National Peer-Reviewed Conference/Proceedings

[44] R. MAÎTREPIERRE, J. MARY, R. MUNOS. *Adaptative Play in Texas Hold'em poker.*, in "10e Conférence d'Apprentissage - CAp 2008", 2008, p. 137–149.

Scientific Books (or Scientific Book chapters)

[45] R. MUNOS. *Programmation dynamique avec approximation de la fonction valeur*, O. SIGAUD, O. BUFFET (editors), Hermes, 2008.

[46] 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, Springer Verlag, Berlin, Germany, 2008, p. 113–124.

Books or Proceedings Editing

[47] S. GIRGIN, M. LOTH, R. MUNOS, P. PREUX, D. RYABKO (editors). *Recent Advances in reinforcement Learning*, Lecture Notes in Artificial Intelligence, vol. 5323, Springer, 2008.

Research Reports

[48] M. LOTH, P. PREUX. *Equi-correlation networks: nonlinear regression by following the L_1 regularization path*, under submission, Technical report, n^{o} RR-, INRIA, March 2008.

Other Publications

[49] M. LOTH, P. PREUX. *Reinforcement learning by direct optimal value estimation and regret minimization*, in "8th European Workshop on Reinforcement Learning", June 2008.

[50] A. RABAOUI, M. DAVY, S. ROSSIGNOL, Z. LACHIRI, N. ELLOUZE. *Using One-Class SVMs and Wavelets for Audio Surveillance Systems*, (submitted), 2008.

References in notes

[51] 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.

[52] L. BAIRD. *Residual Algorithms: Reinforcement learning with function approximation*, in "Proceedings of the Twelfth International Conference on Machine Learning", 1995, p. 30–37.

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

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

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

[56] S. BRADTKE, A. BARTO. *Linear least-squares algorithms for temporal difference learning*, in "Machine Learning", vol. 22, 1996, p. 33–57.

[57] F. CARON, A. DOUCET, 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 2007, p. 2703–2719.

[58] 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.

[59] 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.

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

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

[62] 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.

[63] 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.

[64] 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.

[65] M. LOTH, P. PREUX. *The Equi-Correlation Network: A new kernelized-LARS with automatic kernel parameters tuning*, (submitted), 2008.

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

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

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

[69] J. RUST. *How Social Security and Medicare Affect Retirement Behavior in a World of Incomplete Market*, in "Econometrica", available via urlhttp://gemini.econ.umd.edu/jrust/papers.html, vol. 65, n^{o} 4, July 1997, p. 781–831.

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

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

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

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

[74] 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.

[75] R. J. WILLIAMS, L. BAIRD. *Tight Performance Bounds on Greedy Policies Based on Imperfect Value Functions*, in "Proceedings of the Tenth Yale Workshop on Adaptive and Learning Systems", 1994.