IN PARTNERSHIP WITH:
Université Charles de Gaulle (Lille 3)

Université des sciences et technologies de Lille (Lille 1)

Ecole Centrale de Lille

Activity Report 2012

Project-Team SEQUEL

Sequential Learning

IN COLLABORATION WITH: Laboratoire d'informatique fondamentale de Lille (LIFL), Laboratoire d'Automatique, de Génie Informatique et Signal (LAGIS)

RESEARCH CENTER
Lille - Nord Europe

THEME
Optimization, Learning and Statistical Methods

Table of contents

Project-Team SEQUEL

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

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

Creation of the Project-Team: July 01, 2007 .

1. Members

Research Scientists

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

Faculty Members

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

PhD Students

Boris Baldassari [CIFRE with Squoring Technology, since Sep., 2011]
Alexandra Carpentier [ANR-Région Nord-Pas de Calais Grant, until Oct., 2012]
Emmanuel Delande [DGA, until Jan., 2012]
Victor Gabillon [MENESR Grant, since Oct., 2009]
Adrien Hoarau [DGA, since Oct., 2012]
Jean-François Hren [MENESR Grant, until Jun., 2012]
Azadeh Khaleghi [CORDIS grant, since Oct., 2010]
Sami Naamane [CIFRE with France Telecom Grant, since Nov., 2011]
Olivier Nicol [MENESR Grant, since Oct., 2010]
Christophe Salperwyck [CIFRE with France Telecom Grant, until Nov., 2012]
Amir Sani [CORDIS grant, since Oct., 2011]
Marta Soare [Inria-Région Nord pas de Calais grant, since Oct., 2012]

Post-Doctoral Fellows

Hachem Kadri [ANR Lampada, until Aug. 2012]
Nathaniel Korda [ANR Explora, then COMPLacs, since Oct., 2011]
Michal Valko [COMPLacs, until Aug., 2012]
Rapahël Fonteneau [FNRS, since May, 2012]
Prashanth Lakshmanrao Anantha Padmanabha [COMPLacs, since Nov., 2012]

Administrative Assistants

Sandrine Catillon [Secretary (SAR) Inria, shared by 2 projects, until Sep. 2012]
Amélie Supervielle [Secretary (SAR) Inria, shared by 2 projects, since Oct. 2012]

Other

Adrien Hoarau [Master 2 internship, ENS-Cachan, École Polytechnique, Apr. to Aug. 2012]

2. Overall Objectives

2.1. Overall Objectives

SEQUEL means "Sequential Learning". As such, SEQUEL focuses on the task of learning in artificial systems (either hardware, or software) that gather information along time. Such systems are named *(learning) agents* (or learning machines) in the following. These data may be used to estimate some parameters of a model, which in turn, may be used for selecting actions in order to perform some long-term optimization task.

For the purpose of model building, the agent needs to represent information collected so far in some compact form and use it to process newly available data.

The acquired data may result from an observation process of an agent in interaction with its environment (the data thus represent a perception). This is the case when the agent makes decisions (in order to attain a certain objective) that impact the environment, and thus the observation process itself.

Hence, in SEQUEL, the term **sequential** refers to two aspects:

- The **sequential acquisition of data**, from which a model is learned (supervised and non supervised learning),
- the **sequential decision making task**, based on the learned model (reinforcement learning).

Examples of sequential learning problems include:

Supervised learning tasks deal with the prediction of some response given a certain set of observations of input variables and responses. New sample points keep on being observed.

Unsupervised learning tasks deal with clustering objects, these latter making a flow of objects. The (unknown) number of clusters typically evolves during time, as new objects are observed.

Reinforcement learning tasks deal with the control (a policy) of some system which has to be optimized (see [74]). We do not assume the availability of a model of the system to be controlled.

In all these cases, we mostly assume that the process can be considered stationary for at least a certain amount of time, and slowly evolving.

We wish to have any-time algorithms, that is, at any moment, a prediction may be required/an action may be selected making full use, and hopefully, the best use, of the experience already gathered by the learning agent.

The perception of the environment by the learning agent (using its sensors) is generally neither the best one to make a prediction, nor to take a decision (we deal with Partially Observable Markov Decision Problem). So, the perception has to be mapped in some way to a better, and relevant, state (or input) space.

Finally, an important issue of prediction regards its evaluation: how wrong may we be when we perform a prediction? For real systems to be controlled, this issue can not be simply left unanswered.

To sum-up, in SEQUEL, the main issues regard:

- the learning of a model: we focus on models that map some input space $\mathbb{R}^P$ to $\mathbb{R}$,
- the observation to state mapping,
- the choice of the action to perform (in the case of sequential decision problem),
- the performance guarantees,
- the implementation of usable algorithms,

all that being understood in a *sequential* framework.

3. Scientific Foundations

3.1. Introduction

SEQUEL is primarily grounded on two domains:

- the problem of decision under uncertainty,
- statistical analysis and statistical learning, which provide the general concepts and tools to solve this problem.

To help the reader who is unfamiliar with these questions, we briefly present key ideas below.

3.2. Decision-making Under Uncertainty

The phrase "Decision under uncertainty" refers to the problem of taking decisions when we do not have a full knowledge neither of the situation, nor of the consequences of the decisions, as well as when the consequences of decision are non deterministic.

We introduce two specific sub-domains, namely the Markov decision processes which models sequential decision problems, and bandit problems.

3.2.1. Reinforcement Learning

Sequential decision processes occupy the heart of the SEQUEL project; a detailed presentation of this problem may be found in Puterman's book [70].

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 [1] $r_1, r_2, ...$ defined by $r_t = r(x_t, a_t, x_{t+1})$.

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

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

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

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 [66]) 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.

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

In order to maximize a given functional in a sequential framework, one usually applies Dynamic Programming (DP) [64], 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 [65]. 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 ([76]):

- 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. Multi-arm Bandit Theory

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 [71], after which they have found applications in diverse fields, such as control, economics, statistics, or learning theory.

Formally, a K-armed bandit problem ($K \geq 2$) is specified by K real-valued distributions. In each time step a decision maker can select one of the distributions to obtain a sample from it. The samples obtained are considered as rewards. The distributions are initially unknown to the decision maker, whose goal is to maximize the sum of the rewards received, or equivalently, to minimize the regret which is defined as the loss compared to the total payoff that can be achieved given full knowledge of the problem, *i.e.*, when the arm giving the highest expected reward is pulled all the time.

The name "bandit" comes from imagining a gambler playing with K slot machines. The gambler can pull the arm of any of the machines, which produces a random payoff as a result: When arm k is pulled, the random payoff is drawn from the distribution associated to k. Since the payoff distributions are initially unknown, the gambler must use exploratory actions to learn the utility of the individual arms. However, exploration has to be carefully controlled since excessive exploration may lead to unnecessary losses. Hence, to play well, the gambler must carefully balance exploration and exploitation. Auer *et al.* [63] introduced the algorithm UCB (Upper Confidence Bounds) that follows what is now called the "optimism in the face of uncertainty principle". Their algorithm works by computing upper confidence bounds for all the arms and then choosing the arm with the highest such bound. They proved that the expected regret of their algorithm increases at most at a logarithmic rate with the number of trials, and that the algorithm achieves the smallest possible regret up to some sub-logarithmic factor (for the considered family of distributions).

3.3. Statistical analysis of time series

Many of the problems of machine learning can be seen as extensions of classical problems of mathematical statistics to their (extremely) non-parametric and model-free cases. Other machine learning problems are founded on such statistical problems. Statistical problems of sequential learning are mainly those that are concerned with the analysis of time series. These problems are as follows.

3.3.1. Prediction of Sequences of Structured and Unstructured Data

Given a series of observations $x_1, \cdots, x_n$ it is required to give forecasts concerning the distribution of the distribution of the future observations $x_{n+1}, x_{n+2}, \cdots$; in the simplest case, that of the next outcome x_{n+1}. Then x_{n+1} is revealed and the process continues. Different goals can be formulated in this setting. One can either make some assumptions on the probability measure that generates the sequence $x_1, \cdots, x_n, \cdots$, such as that the outcomes are independent and identically distributed (i.i.d.), or that the sequence is a Markov chain, that it is a stationary process, etc. More generally, one can assume that the data is generated by a probability measure that belongs to a certain set $\mathcal{C}$. In these cases the goal is to have the discrepancy between the predicted and the "true" probabilities to go to zero, if possible, with guarantees on the speed of convergence.

Alternatively, rather than making some assumptions on the data, one can change the goal: the predicted probabilities should be asymptotically as good as those given by the best reference predictor from a certain pre-defined set.

Another dimension of complexity in this problem concerns the nature of observations x_i. In the simplest case, they come from a finite space, but already basic applications often require real-valued observations. Moreover, function or even graph-valued observations often arise in practice, in particular in applications concerning Web data. In these settings estimating even simple characteristics of probability distributions of the future outcomes becomes non-trivial, and new learning algorithms for solving these problems are in order.

3.3.2. Hypothesis testing

Given a series of observations of $x_1, \cdots, x_n, \cdots$ generated by some unknown probability measure μ, the problem is to test a certain given hypothesis H_0 about μ, versus a given alternative hypothesis H_1. There are many different examples of this problem. Perhaps the simplest one is testing a simple hypothesis "μ is Bernoulli i.i.d. measure with probability of 0 equals 1/2" versus "μ is Bernoulli i.i.d. with the parameter different from 1/2". More interesting cases include the problems of model verification: for example, testing that μ is a Markov chain, versus that it is a stationary ergodic process but not a Markov chain. In the case when we have not one but several series of observations, we may wish to test the hypothesis that they are independent, or that they are generated by the same distribution. Applications of these problems to a more general class of machine learning tasks include the problem of feature selection, the problem of testing that a certain behaviour (such as pulling a certain arm of a bandit, or using a certain policy) is better (in terms of achieving some goal, or collecting some rewards) than another behaviour, or than a class of other behaviours.

The problem of hypothesis testing can also be studied in its general formulations: given two (abstract) hypothesis H_0 and H_1 about the unknown measure that generates the data, find out whether it is possible to test H_0 against H_1 (with confidence), and if yes then how can one do it.

3.3.3. Change Point Analysis

A stochastic process is generating the data. At some point, the process distribution changes. In the "offline" situation, the statistician observes the resulting sequence of outcomes and has to estimate the point or the points at which the change(s) occurred. In online setting, the goal is to detect the change as quickly as possible.

These are the classical problems in mathematical statistics, and probably among the last remaining statistical problems not adequately addressed by machine learning methods. The reason for the latter is perhaps in that the problem is rather challenging. Thus, most methods available so far are parametric methods concerning piece-wise constant distributions, and the change in distribution is associated with the change in the mean. However, many applications, including DNA analysis, the analysis of (user) behaviour data, etc., fail to comply with this kind of assumptions. Thus, our goal here is to provide completely non-parametric methods allowing for any kind of changes in the time-series distribution.

3.3.4. Clustering Time Series, Online and Offline

The problem of clustering, while being a classical problem of mathematical statistics, belongs to the realm of unsupervised learning. For time series, this problem can be formulated as follows: given several samples $x^1 = (x^1_1, \cdots, x^1_{n_1}), \cdots, x^N = (x^1_N, \cdots, x^N_{n_N})$, we wish to group similar objects together. While this is of course not a precise formulation, it can be made precise if we assume that the samples were generated by k different distributions.

The online version of the problem allows for the number of observed time series to grow with time, in general, in an arbitrary manner.

3.3.5. Online Semi-Supervised Learning

Semi-supervised learning (SSL) is a field of machine learning that studies learning from both labeled and unlabeled examples. This learning paradigm is extremely useful for solving real-world problems, where data is often abundant but the resources to label them are limited.

Furthermore, *online* SSL is suitable for adaptive machine learning systems. In the classification case, learning is viewed as a repeated game against a potentially adversarial nature. At each step t of this game, we observe an example $\mathbf{x_t}$, and then predict its label $\widehat{y}_t$.

The challenge of the game is that we only exceptionally observe the true label y_t. In the extreme case, which we also study, only a handful of labeled examples are provided in advance and set the initial bias of the system while unlabeled examples are gathered online and update the bias continuously. Thus, if we want to adapt to changes in the environment, we have to rely on indirect forms of feedback, such as the structure of data.

3.4. Statistical Learning and Bayesian Analysis

Before detailing some issues in these fields, 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.

Statistical learning is an approach to machine intelligence that 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.

Bayesian Analysis applies to data that could be seen as observations in the more general meaning of the term. These data may not only come from classical sensors but also from any *device* recording information. From an operational point of view, like for statistical learning, uncertainty about the data is modeled by a probability measure thus defining the so-called likelihood functions. This last one depend upon parameters defining the state of the world we focus on for decision purposes. Within the Bayesian framework the uncertainty about these parameters is also modeled by probability measures, the priors that are subjective probabilities. Using probability theory and decision theory, one then defines new algorithms to estimate the parameters of interest and/or associated decisions. According to the International Society for Bayesian Analysis (source: http://bayesian.org), and from a more general point of view, this overall process could be summarize as follows: one assesses the current state of knowledge regarding the issue of interest, gather new data to address remaining questions, and then update and refine their understanding to incorporate both new and old data. Bayesian inference provides a logical, quantitative framework for this process based on probability theory.

Kernel method. Generally speaking, a kernel function is a function that maps a couple of points to a real value. Typically, this value is a measure of dissimilarity between the two points. Assuming a few properties on it, the kernel function implicitly defines a dot product in some function space. This very nice formal property as well as a bunch of others have ensured a strong appeal for these methods in the last 10 years in the field of function approximation. Many classical algorithms have been "kernelized", that is, restated in a much more general way than their original formulation. Kernels also implicitly induce the representation of data in a certain "suitable" space where the problem to solve (classification, regression, ...) is expected to be simpler (non-linearity turns to linearity).

The fundamental tools used in SEQUEL come from the field of statistical learning [68]. We briefly present the most important for us to date, namely, kernel-based non parametric function approximation, and non parametric Bayesian models.

3.4.1. Non-parametric methods for Function Approximation

In statistics in general, and applied mathematics, the approximation of a multi-dimensional real function given some samples is a well-known problem (known as either regression, or interpolation, or function approximation, ...). Regressing a function from data is a key ingredient of our research, or to the least, a basic component of most of our algorithms. In the context of sequential learning, we have to regress a function while data samples are being obtained one at a time, while keeping the constraint to be able to predict points at any step along the acquisition process. In sequential decision problems, we typically have to learn a value function, or a policy.

Many methods have been proposed for this purpose. We are looking for suitable ones to cope with the problems we wish to solve. In reinforcement learning, the value function may have areas where the gradient is large; these are areas where the approximation is difficult, while these are also the areas where the accuracy of the approximation should be maximal to obtain a good policy (and where, otherwise, a bad choice of action may imply catastrophic consequences).

We particularly favor non parametric methods since they make quite a few assumptions about the function to learn. In particular, we have strong interests in l_1-regularization, and the (kernelized-)LARS algorithm. l_1-regularization yields sparse solutions, and the LARS approach produces the whole regularization path very efficiently, which helps solving the regularization parameter tuning problem.

3.4.2. Nonparametric Bayesian Estimation

Numerous problems may be solved efficiently by a Bayesian approach. The use of Monte-Carlo methods allows us to handle non–linear, as well as non–Gaussian, problems. In their standard form, they require the formulation of probability densities in a parametric form. For instance, it is a common usage to use Gaussian likelihood, because it is handy. However, in some applications such as Bayesian filtering, or blind deconvolution, the choice of a parametric form of the density of the noise is often arbitrary. If this choice is wrong, it may also have dramatic consequences on the estimation quality. To overcome this shortcoming, one possible approach is to consider that this density must also be estimated from data. A general Bayesian approach then consists in defining a probabilistic space associated with the possible outcomes of the *object* to be estimated. Applied to density estimation, it means that we need to define a probability measure on the probability density of the noise : such a measure is called a *random measure*. The classical Bayesian inference procedures can then been used. This approach being by nature non parametric, the associated frame is called *Non Parametric Bayesian*.

In particular, mixtures of Dirichlet processes [67] provide a very powerful formalism. Dirichlet Processes are a possible random measure and Mixtures of Dirichlet Processes are an extension of well-known finite mixture models. Given a mixture density $f(x|\theta)$, and $G(d\theta) = \sum_{k=1}^{\infty} \omega_k \delta_{U_k}(d\theta)$, a Dirichlet process, we define a mixture of Dirichlet processes as:

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

where $F(x)$ is the density to be estimated. The class of densities that may be written as a mixture of Dirichlet processes is very wide, so that they really fit a very large number of applications.

Given a set of observations, the estimation of the parameters of a mixture of Dirichlet processes is performed by way of a Monte Carlo Markov Chain (MCMC) algorithm. Dirichlet Process Mixture are also widely used in clustering problems. Once the parameters of a mixture are estimated, they can be interpreted as the parameters of a specific cluster defining a class as well. Dirichlet processes are well known within the machine learning community and their potential in statistical signal processing still need to be developed.

3.4.3. Random Finite Sets for multisensor multitarget tracking

In the general multi-sensor multi-target Bayesian framework, an unknown (and possibly varying) number of targets whose states $x_1, ...x_n$ are observed by several sensors which produce a collection of measurements $z_1, ..., z_m$ at every time step k. Well-known models to this problem are track-based models, such as the joint probability data association (JPDA), or joint multi-target probabilities, such as the joint multi-target probability density. Common difficulties in multi-target tracking arise from the fact that the system state and the collection of measures from sensors are unordered and their size evolve randomly through time. Vector-based algorithms must therefore account for state coordinates exchanges and missing data within an unknown time interval. Although this approach is very popular and has resulted in many algorithms in the past, it may not be the optimal way to tackle the problem, since the sate and the data are in fact *sets* and not vectors.

The random finite set theory provides a powerful framework to deal with these issues. Mahler's work on finite sets statistics (FISST) provides a mathematical framework to build multi-object densities and derive the Bayesian rules for state prediction and state estimation. Randomness on object number and their states are encapsulated into random finite sets (RFS), namely multi-target(state) sets $X = \{x_1, ..., x_n\}$ and multi-sensor (measurement) set $Zk = \{z_1, ..., z_m\}$. The objective is then to propagate the multitarget probability density $f_{k|k}(X|Z(k))$ by using the Bayesian set equations at every time step k:

$$\begin{aligned} f_{k+1|k}(X|Z^{(k)}) &= \int f_{k+1|k}(X|W) f_{k|k}(W|Z^{(k)}) \delta W \\ f_{k+1|k+1}(X|Z^{(k+1)}) &= \frac{f_{k+1}(Z_{k+1}|X) f_{k+1|k}(X|Z^{(k)})}{\int f_{k+1}(Z_{k+1}|W) f_{k+1|k}(W|Z^{(k)}) \delta W} \end{aligned} \tag{5}$$

where:

- $X = \{x_1, ..., x_n\}$ is a multi-target state, i.e. a finite set of elements x_i defined on the single-target space $\mathcal{X}$; [2]
- $Z_{k+1} = \{z_1, ..., z_m\}$ is the current multi-sensor observation, i.e. a collection of measures z_i produced at time $k+1$ by all the sensors;
- $Z^{(k)} = \bigcup_{t \leqslant k} Z_t$ is the collection of observations up to time k;
- $f_{k|k}(W|Z^{(k)})$ is the current multi-target posterior density in state W;
- $f_{k+1|k}(X|W)$ is the current multi-target Markov transition density, from state W to state X;
- $f_{k+1}(Z|X)$ is the current multi-sensor/multi-target likelihood function.

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

4. Application Domains

4.1. Introduction

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 processing and functional prediction,
- medical applications,
- web mining,
- computer games.

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

4.2. Adaptive Control

Adaptive control is an important application of the research being done in SEQUEL. Reinforcement learning (RL) precisely aims at controling the behavior of systems and may be used in situations with more or less information available. Of course, the more information, the better, in which case methods of (approximate) dynamic programming may be used [69]. 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 [75] that learned to play at an expert level by basically playing a very large amount of games against itself. Today, various games are studied with RL techniques.
- many optimization problems that are closely related to operation research, but taking into account the uncertainty, and the stochasticity of the environment: see the job-shop scheduling, or the cellular phone frequency allocation problems, resource allocation in general [69]
- we can also foresee that some progress may be made by using RL to design adaptive conversational agents, or system-level as well as application-level operating systems that adapt to their users habits.

  More generally, these ideas fall into what adaptive control may bring to human beings, in making their life simpler, by being embedded in an environment that is made to help them, an idea phrased as "ambient intelligence".
- The sensor management problem consists in determining the best way to task several sensors when each sensor has many modes and search patterns. In the detection/tracking applications, the tasks assigned to a sensor management system are for instance:
  - detect targets,
  - track the targets in the case of a moving target and/or a smart target (a smart target can change its behavior when it detects that it is under analysis),
  - combine all the detections in order to track each moving target,
  - dynamically allocate the sensors in order to achieve the previous three tasks in an optimal way. The allocation of sensors, and their modes, thus defines the action space of the underlying Markov decision problem.

  In the more general situation, some sensors may be localized at the same place while others are dispatched over a given volume. Tasking a sensor may include, at each moment, such choices as where to point and/or what mode to use. Tasking a group of sensors includes the tasking of each individual sensor but also the choice of collaborating sensors subgroups. Of course, the sensor management problem is related to an objective. In general, sensors must balance complex trade-offs between achieving mission goals such as detecting new targets, tracking existing targets, and identifying existing targets. The word "target" is used here in its most general meaning, and the potential applications are not restricted to military applications. Whatever the underlying application, the sensor management problem consists in choosing at each time an action within the set of available actions.
- sequential decision processes are also very well-known in economy. They may be used as a decision aid tool, to help in the design of social helps, or the implementation of plants (see [73], [72] for such applications).

4.3. Signal Processing

Applications of sequential learning in the field of signal processing are also very numerous. A signal is naturally sequential as it flows. It usually comes from the recording of the output of sensors but the recording of any sequence of numbers may be considered as a signal like the stock-exchange rates evolution with respect to time and/or place, the number of consumers at a mall entrance or the number of connections to a web site. Signal processing has several objectives: predict , estimate, remove noise, characterize or classify. The signal is often considered as sequential: we want to predict, estimate or classify a value (or a feature) at time t knowing the past values of the parameter of interest or past values of data related to this parameter. This is typically the case in estimation processes arising in dynamical systems.

Signals may be processed in several ways. One of the best–known way is the time-frequency analysis in which the frequencies of each signal are analyzed with respect to time. This concept has been generalized to the time-scale analysis obtained by a wavelet transform. Both analysis are based on the projection of the original signal onto a well-chosen function basis. Signal processing is also closely related to the probability field as the uncertainty inherent to many signals leads to consider them as stochastic processes: the Bayesian framework is actually one of the main frameworks within which signals are processed for many purposes. It is worth noting that Bayesian analysis can be used jointly with a time-frequency or a wavelet analysis. However, alternatives like belief functions came up these last years. Belief functions were introduced by Demspter few decades ago and have been successfully used in the few past years in fields where probability had, during many years, no alternatives like in classification. Belief functions can be viewed as a generalization of probabilities which can capture both imprecision and uncertainty. Belief functions are also closely related to data fusion.

4.4. Medical Applications

One of the initial motivations of the multi-arm bandit theory stems from clinical trials when one researches the effects of different treatments while maximizing the improvement of the patients' health states.

Medical health-care and in particular patient-management is up today one of the most important applications of the sequential decision making. This is because the treatment of the more complex health problems is typically sequential: A physician repeatedly observes the current state of the patient and makes the decision in order to improve the health condition as measured for example by *qualys* (quality adjusted life years).

Moreover, machine learning methods may be used for at least two means in neuroscience:

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).

4.5. Web Mining

We work on the news/ad recommendation. Theses online learning algorithms reached a critical importance over the last few years due to these major applications. After designing a new algorithm, it is critical to be able to evaluate it without having to plug it into the real application in order to protect user experiences or/and the company's revenue. To do this, people used to build simulators of user behaviors and try to achieve good performances against it. However designing such a simulator is probably much more difficult than designing the algorithm itself! An other common way to evaluate is to not consider the exploration/exploitation dilemma (also known as "Cold Start" for recommender systems). Lately data-driven methods have been developed. We are working on building automatic replay methodology with some theoretical guarantees. This work also exhibits strong link with the choice of the number of contexts to use with recommender systems wrt your audience.

An other point is that web sites must forecast Web page views in order to plan computer resource allocation and estimate upcoming revenue and advertising growth. In this work, we focus on extracting trends and seasonal patterns from page view series. We investigate Holt-Winters/ARIMA like procedures and some regularized models for making short-term prediction (3-6 weeks) wrt to logged data of several big media websites. We work on some news event related webpages and we feel that kind of time series deserves a particular attention. Self-similarity is found to exist at multiple time scales of network traffic, and can be exploited for prediction. In particular, it is found that Web page views exhibit strong impulsive changes occasionally. The impulses cause large prediction errors long after their occurrences and can sometime be predicted (e.g., elections, sport events, editorial changes,holidays) in order to improve accuracies. It also seems that some promising model could arise from using global trends shift in the population.

4.6. Games

The problem of artificial intelligence in games consists in choosing actions of players in order to produce artificial opponents. Most games can be formalized as Markov decision problems, so they can be approached with reinforcement learning.

In particular, SEQUEL was a pioneer of Monte Carlo Tree Search, a technique that obtained spectacular successes in the game of Go. Other application domains include the game of poker and the Japanese card game of hanafuda.

5. Software

5.1. Introduction

In 2012, SEQUEL continued the development of software for computer games (notably Go) and also developed two novel libraries for functional regression and data mining.

5.2. Computer Games

Participant: Rémi Coulom.

We continued the development of three main softwares for computer games:

- ***Crazy Stone*** is a top-level Go-playing program that has been developed by Rémi Coulom since 2005. Crazy Stone won several major international Go tournaments in the past. In 2012, a new version was released in Japan. This new version won a game with a 4-stone handicap against a professional player during the European Go Congress in Bonn, Germany. It is distributed as a commercial product by *Unbalance Corporation* (Japan). 6-month work in 2012. URL: http://remi.coulom.free.fr/CrazyStone/
- ***Crazy Hanafuda*** is a program to play the Japanese card game of Hanafuda. One month of work in 2012. A licence agreement was signed with Unbalance Corporation in January. The Windows 8 version of the program was released commercially in November.
- ***CLOP*** [30] is a tool for automatic parameter optimization of game-playing programs. Distributed as freeware (GPL). One month of work in 2012. Available at: http://remi.coulom.free.fr/CLOP/

5.3. Vowpal Wabbit

Participants: Jérémie Mary, Romaric Gaudel, Thomas Chabin.

Vowpal Wabbit is a GPL project leaded by John Langford at Yahoo! Research and now at Microsoft. The goal is to build a very fast, distributed and large scale machine learning software. https://github.com/JohnLangford/vowpal_wabbit/wiki. We worked on the optimization of the parser and on the memory structures of the i/o. The modifications have been accepted for commit in the main branch and allow an average division by two of all execution times.

6. New Results

6.1. Decision-making Under Uncertainty

6.1.1. Reinforcement Learning

Transfer in Reinforcement Learning: a Framework and a Survey [56]

Transfer in reinforcement learning is a novel research area that focuses on the development of methods to transfer knowledge from a set of source tasks to a target task. Whenever the tasks are *similar*, the transferred knowledge can be used by a learning algorithm to solve the target task and significantly improve its performance (e.g., by reducing the number of samples needed to achieve a nearly optimal performance). In this chapter we provide a formalization of the general transfer problem, we identify the main settings which have been investigated so far, and we review the most important approaches to transfer in reinforcement learning.

Online Regret Bounds for Undiscounted Continuous Reinforcement Learning [44]

We derive sublinear regret bounds for undiscounted reinforcement learning in continuous state space. The proposed algorithm combines state aggregation with the use of upper confidence bounds for implementing optimism in the face of uncertainty. Beside the existence of an optimal policy which satisfies the Poisson equation, the only assumptions made are Holder continuity of rewards and transition probabilities.

Semi-Supervised Apprenticeship Learning [23]

In apprenticeship learning we aim to learn a good policy by observing the behavior of an expert or a set of experts. In particular, we consider the case where the expert acts so as to maximize an unknown reward function defined as a linear combination of a set of state features. In this paper, we consider the setting where we observe many sample trajectories (i.e., sequences of states) but only one or a few of them are labeled as experts' trajectories. We investigate the conditions under which the remaining unlabeled trajectories can help in learning a policy with a good performance. In particular, we define an extension to the max-margin inverse reinforcement learning proposed by Abbeel and Ng (2004) where, at each iteration, the max-margin optimization step is replaced by a semi-supervised optimization problem which favors classifiers separating clusters of trajectories. Finally, we report empirical results on two grid-world domains showing that the semi-supervised algorithm is able to output a better policy in fewer iterations than the related algorithm that does not take the unlabeled trajectories into account.

Fast Reinforcement Learning with Large Action Sets Using Error-Correcting Output Codes for MDP Factorization [31] [48]

The use of Reinforcement Learning in real-world scenarios is strongly limited by issues of scale. Most RL learning algorithms are unable to deal with problems composed of hundreds or sometimes even dozens of possible actions, and therefore cannot be applied to many real-world problems. We consider the RL problem in the supervised classification framework where the optimal policy is obtained through a multiclass classifier, the set of classes being the set of actions of the problem. We introduce error-correcting output codes (ECOCs) in this setting and propose two new methods for reducing complexity when using rollouts-based approaches. The first method consists in using an ECOC-based classifier as the multiclass classifier, reducing the learning complexity from O(A2) to O(Alog(A)) . We then propose a novel method that profits from the ECOC's coding dictionary to split the initial MDP into O(log(A)) separate two-action MDPs. This second method reduces learning complexity even further, from O(A2) to O(log(A)) , thus rendering problems with large action sets tractable. We finish by experimentally demonstrating the advantages of our approach on a set of benchmark problems, both in speed and performance.

Analysis of Classification-based Policy Iteration Algorithms [13]

We introduce a variant of the classification-based approach to policy iteration which uses a cost-sensitive loss function weighting each classification mistake by its actual regret, i.e., the difference between the action-value of the greedy action and of the action chosen by the classifier. For this algorithm, we provide a full finite-sample analysis. Our results state a performance bound in terms of the number of policy improvement steps, the number of rollouts used in each iteration, the capacity of the considered policy space (classifier), and a capacity measure which indicates how well the policy space can approximate policies that are greedy w.r.t. any of its members. The analysis reveals a tradeoff between the estimation and approximation errors in this classification-based policy iteration setting. Furthermore it confirms the intuition that classification-based policy iteration algorithms could be favorably compared to value-based approaches when the policies can be approximated more easily than their corresponding value functions. We also study the consistency of the algorithm when there exists a sequence of policy spaces with increasing capacity.

Minimax PAC-Bounds on the Sample Complexity of Reinforcement Learning with a Generative Model [5] [24]

We consider the problem of learning the optimal action-value function in discounted-reward Markov decision processes (MDPs). We prove new PAC bounds on the sample-complexity of two well-known model-based reinforcement learning (RL) algorithms in the presence of a generative model of the MDP: value iteration and policy iteration. The first result indicates that for an MDP with N state-action pairs and the discount factor $\gamma \in [0, 1)$ only $O(N \log(N/\delta)/[(1-\gamma)^3\epsilon^2])$ state-transition samples are required to find an ϵ-optimal estimation of the action-value function with the probability (w.p.) $1-\delta$. Further, we prove that, for small values of ϵ, an order of $O(N \log(N/\delta)/[(1-\gamma)^3\epsilon^2])$ samples is required to find an ϵ-optimal policy w.p. $1-\delta$. We also prove a matching lower bound of $\Omega(N \log(N/\delta)/[(1-\gamma)^3\epsilon^2])$ on the sample complexity of estimating the optimal action-value function. To the best of our knowledge, this is the first minimax result on the sample complexity of RL: The upper bound matches the lower bound interms of N , ϵ, δ and $1/(1-\gamma)$ up to a constant factor. Also, both our lower bound and upper bound improve on the state-of-the-art in terms of their dependence on $1/(1-\gamma)$.

Optimistic planning in Markov decision processes [25]

The reinforcement learning community has recently intensified its interest in online planning methods, due to their relative independence on the state space size. However, tight near-optimality guarantees are not yet available for the general case of stochastic Markov decision processes and closed-loop, state-dependent planning policies. We therefore consider an algorithm related to $AO*$ that optimistically explores a tree representation of the space of closed-loop policies, and we analyze the near-optimality of the action it returns after n tree node expansions. While this optimistic planning requires a finite number of actions and possible next states for each transition, its asymptotic performance does not depend directly on these numbers, but only on the subset of nodes that significantly impact near-optimal policies. We characterize this set by introducing a novel measure of problem complexity, called the near-optimality exponent. Specializing the exponent and performance bound for some interesting classes of MDPs illustrates the algorithm works better when there are fewer near-optimal policies and less uniform transition probabilities.

Risk Bounds in Cost-sensitive Multiclass Classification: an Application to Reinforcement Learning [61]

We propose a computationally efficient classification-based policy iteration (CBPI) algorithm. The key idea of CBPI is to view the problem of computing the next policy in policy iteration as a classification problem. We propose a new cost-sensitive surrogate loss for each iteration of CBPI. This allows us to replace the non-convex optimization problem that needs to be solved at each iteration of the existing CBPI algorithms with a convex one. We show that the new loss is classification calibrated, and thus is a sound surrogate loss, and find a calibration function (i.e., a function that represents the convergence rate of the true loss in terms of the convergence rate of the surrogate-loss) for this loss. To the best of our knowledge, this is the first calibration result (with convergence rate) in the context of multi-class classification. As a result, we are able to extend the theoretical guarantees of the existing CBPI algorithms that deal with a non-convex optimization at each iteration to our convex and efficient algorithm, and thereby, obtain the first computationally efficient and theoretically sound CBPI algorithm.

Least-Squares Methods for Policy Iteration [55]

Approximate reinforcement learning deals with the essential problem of applying reinforcement learning in large and continuous state-action spaces, by us- ing function approximators to represent the solution. This chapter reviews least-squares methods for policy iteration, an important class of algorithms for approxi- mate reinforcement learning. We discuss three techniques for solving the core, pol- icy evaluation component of policy iteration, called: least-squares temporal difference, least-squares policy evaluation, and Bellman residual minimization. We introduce these techniques starting from their general mathematical principles and detailing them down to fully specified algorithms. We pay attention to online variants of policy iteration, and provide a numerical example highlighting the behavior of representative offline and online methods. For the policy evaluation component as well as for the overall resulting approximate policy iteration, we provide guarantees on the performance obtained asymptotically, as the number of processed samples and executed iterations grows to infinity. We also provide finite-sample results, which apply when a finite number of samples and iterations is considered. Finally, we outline several extensions and improvements to the techniques and methods reviewed

On Classification-based Approximate Policy Iteration [53]

Efficient methods for tackling large reinforcement learning problems usually exploit special structure, or regularities, of the problem at hand. For example, classification-based approximate policy iteration explicitly controls the complexity of the policy space, which leads to considerable improvement in convergence speed whenever the optimal policy is easy to represent. Conventional classification-based methods, however, do not benefit from regularities of the value function, because they typically use rollout-based estimates of the action-value function. This Monte Carlo-style approach for value estimation is data-inefficient and does not generalize the estimated value function over states. We introduce a general framework for classification-based approximate policy iteration (CAPI) which exploits regularities of both the policy and the value function. Our theoretical analysis extends existing work by allowing the policy evaluation step to be performed by any reinforcement learning algorithm (including temporal-difference style methods), by handling nonparametric representations of policies, and by providing tighter convergence bounds on the estimation error of policy learning. In our experiments, instantiations of CAPI outperformed powerful purely value-based approaches.

Conservative and Greedy Approaches to Classification-based Policy Iteration [37]

The existing classification-based policy iteration (CBPI) algorithms can be divided into two categories: *direct policy iteration* (DPI) methods that directly assign the output of the classifier (the approximate greedy policy w.r.t. the current policy) to the next policy, and *conservative policy iteration* (CPI) methods in which the new policy is a mixture distribution of the current policy and the output of the classifier. The conservative policy update gives CPI a desirable feature, namely the guarantee that the policies generated by this algorithm improve at each iteration. We provide a detailed algorithmic and theoretical comparison of these two classes of CBPI algorithms. Our results reveal that in order to achieve the same level of accuracy, CPI requires more iterations, and thus, more samples than the DPI algorithm. Furthermore, CPI may converge to suboptimal policies whose performance is not better than DPI's.

A Dantzig Selector Approach to Temporal Difference Learning [36]

LSTD is a popular algorithm for value function approximation. Whenever the number of features is larger than the number of samples, it must be paired with some form of regularization. In particular, l1-regularization methods tend to perform feature selection by promoting sparsity, and thus, are well- suited for high-dimensional problems. However, since LSTD is not a simple regression algorithm, but it solves a fixed-point problem, its integration with l1-regularization is not straightforward and might come with some drawbacks (e.g., the P-matrix assumption for LASSO-TD). In this paper, we introduce a novel algorithm obtained by integrating LSTD with the Dantzig Selector. We investigate the performance of the proposed algorithm and its relationship with the existing regularized approaches, and show how it addresses some of their drawbacks.

Finite-Sample Analysis of Least-Squares Policy Iteration [14]

In this paper, we report a performance bound for the widely used least-squares policy iteration (LSPI) algorithm. We first consider the problem of policy evaluation in reinforcement learning, that is, learning the value function of a fixed policy, using the least-squares temporal-difference (LSTD) learning method, and report finite-sample analysis for this algorithm. To do so, we first derive a bound on the performance of the LSTD solution evaluated at the states generated by the Markov chain and used by the algorithm to learn an estimate of the value function. This result is general in the sense that no assumption is made on the existence of a stationary distribution for the Markov chain. We then derive generalization bounds in the case when the Markov chain possesses a stationary distribution and is β-mixing. Finally, we analyze how the error at each policy evaluation step is propagated through the iterations of a policy iteration method, and derive a performance bound for the LSPI algorithm.

Approximate Modified Policy Iteration [47]

Modified policy iteration (MPI) is a dynamic programming (DP) algorithm that contains the two celebrated policy and value iteration methods. Despite its generality, MPI has not been thoroughly studied, especially its approximation form which is used when the state and/or action spaces are large or infinite. In this paper, we propose three implementations of approximate MPI (AMPI) that are extensions of well-known approximate DP algorithms: fitted-value iteration, fitted-Q iteration, and classification-based policy iteration. We provide error propagation analyses that unify those for approximate policy and value iteration. On the last classification-based implementation, we develop a finite-sample analysis that shows that MPI's main parameter allows to control the balance between the estimation error of the classifier and the overall value function approximation.

Bayesian Reinforcement Learning [57]

This chapter surveys recent lines of work that use Bayesian techniques for reinforcement learning. In Bayesian learning, uncertainty is expressed by a prior distribution over unknown parameters and learning is achieved by computing a posterior distribution based on the data observed. Hence, Bayesian reinforcement learning distinguishes itself from other forms of reinforcement learning by explicitly maintaining a distribution over various quantities such as the parameters of the model, the value function, the policy or its gradient. This yields several benefits: a) domain knowledge can be naturally encoded in the prior distribution to speed up learning; b) the exploration/exploitation tradeoff can be naturally optimized; and c) notions of risk can be naturally taken into account to obtain robust policies.

6.1.2. Multi-arm Bandit Theory

Learning with stochastic inputs and adversarial outputs [15]

Most of the research in online learning is focused either on the problem of adversarial classification (i.e., both inputs and labels are arbitrarily chosen by an adversary) or on the traditional supervised learning problem in which samples are independent and identically distributed according to a stationary probability distribution. Nonetheless, in a number of domains the relationship between inputs and outputs may be adversarial, whereas input instances are i.i.d. from a stationary distribution (e.g., user preferences). This scenario can be formalized as a learning problem with stochastic inputs and adversarial outputs. In this paper, we introduce this novel stochastic-adversarial learning setting and we analyze its learnability. In particular, we show that in a binary classification problem over an horizon of n rounds, given a hypothesis space H with finite VC-dimension, it is possible to design an algorithm that incrementally builds a suitable finite set of hypotheses from H used as input for an exponentially weighted forecaster and achieves a cumulative regret of order $O(\sqrt{nVC(H)\log n})$ with overwhelming probability. This result shows that whenever inputs are i.i.d. it is possible to solve any binary classification problem using a finite VC-dimension hypothesis space with a sub-linear regret independently from the way labels are generated (either stochastic or adversarial). We also discuss extensions to multi-class classification, regression, learning from experts and bandit settings with stochastic side information, and application to games.

A Truthful Learning Mechanism for Multi-Slot Sponsored Search Auctions with Externalities [35]

Sponsored search auctions constitute one of the most successful applications of *microeconomic mechanisms*. In mechanism design, auctions are usually designed to incentivize advertisers to bid their truthful valuations and, at the same time, to assure both the advertisers and the auctioneer a non–negative utility. Nonetheless, in sponsored search auctions, the click–through–rates (CTRs) of the advertisers are often unknown to the auctioneer and thus standard incentive compatible mechanisms cannot be directly applied and must be paired with an effective learning algorithm for the estimation of the CTRs. This introduces the critical problem of designing a learning mechanism able to estimate the CTRs as the same time as implementing a truthful mechanism with a revenue loss as small as possible compared to an optimal mechanism designed with the true CTRs. Previous works showed that in single–slot auctions the problem can be solved using a suitable exploration–exploitation mechanism able to achieve a per–step regret of order $O(T^{-1/3})$ (where T is the number of times the auction is repeated). In this paper we extend these results to the general case of contextual multi–slot auctions with position– and ad–dependent externalities. In particular, we prove novel upper–bounds on the revenue loss w.r.t. to a VCG auction and we report numerical simulations investigating their accuracy in predicting the dependency of the regret on the number of rounds T, the number of slots K, and the number of advertisements n.

Regret Bounds for Restless Markov Bandits [43]

We consider the restless Markov bandit problem, in which the state of each arm evolves according to a Markov process independently of the learner's actions. We suggest an algorithm that after T steps achieves $\widetilde{O}(\sqrt{T})$ regret with respect to the best policy that knows the distributions of all arms. No assumptions on the Markov chains are made except that they are irreducible. In addition, we show that index-based policies are necessarily suboptimal for the considered problem.

Online allocation and homogeneous partitioning for piecewise constant mean approximation [42]

In the setting of active learning for the multi-armed bandit, where the goal of a learner is to estimate with equal precision the mean of a finite number of arms, recent results show that it is possible to derive strategies based on finite-time confidence bounds that are competitive with the best possible strategy. We here consider an extension of this problem to the case when the arms are the cells of a finite partition P of a continuous sampling space X in Rd. Our goal is now to build a piecewise constant approximation of a noisy function (where each piece is one region of P and P is fixed beforehand) in order to maintain the local quadratic error of approximation on each cell equally low. Although this extension is not trivial, we show that a simple algorithm based on upper confidence bounds can be proved to be adaptive to the function itself in a near-optimal way, when |P| is chosen to be of minimax-optimal order on the class of alpha-Holder functions.

The Optimistic Principle applied to Games, Optimization and Planning: Towards Foundations of Monte-Carlo Tree Search [17]

This work covers several aspects of the optimism in the face of uncertainty principle applied to large scale optimization problems under finite numerical budget. The initial motivation for the research reported here originated from the empirical success of the so-called Monte-Carlo Tree Search method popularized in computer-go and further extended to many other games as well as optimization and planning problems. Our objective is to contribute to the development of theoretical foundations of the field by characterizing the complexity of the underlying optimization problems and designing efficient algorithms with performance guarantees. The main idea presented here is that it is possible to decompose a complex decision making problem (such as an optimization problem in a large search space) into a sequence of elementary decisions, where each decision of the sequence is solved using a (stochastic) multi-armed bandit (simple mathematical model for decision making in stochastic environments). This so-called hierarchical bandit approach (where the reward observed by a bandit in the hierarchy is itself the return of another bandit at a deeper level) possesses the nice feature of starting the exploration by a quasi-uniform sampling of the space and then focusing progressively on the most promising area, at different scales, according to the evaluations observed so far, and eventually performing a local search around the global optima of the function. The performance of the method is assessed in terms of the optimality of the returned solution as a function of the number of function evaluations. Our main contribution to the field of function optimization is a class of hierarchical optimistic algorithms designed for general search spaces (such as metric spaces, trees, graphs, Euclidean spaces, ...) with

different algorithmic instantiations depending on whether the evaluations are noisy or noiseless and whether some measure of the "smoothness" of the function is known or unknown. The performance of the algorithms depend on the local behavior of the function around its global optima expressed in terms of the quantity of near-optimal states measured with some metric. If this local smoothness of the function is known then one can design very efficient optimization algorithms (with convergence rate independent of the space dimension), and when it is not known, we can build adaptive techniques that can, in some cases, perform almost as well as when it is known.

Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation [6]

We consider optimal sequential allocation in the context of the so-called stochastic multi-armed bandit model. We describe a generic index policy, in the sense of Gittins (1979), based on upper confidence bounds of the arm payoffs computed using the Kullback-Leibler divergence. We consider two classes of distributions for which instances of this general idea are analyzed: The kl-UCB algorithm is designed for one-parameter exponential families and the empirical KL-UCB algorithm for bounded and finitely supported distributions. Our main contribution is a unified finite-time analysis of the regret of these algorithms that asymptotically matches the lower bounds of Lai and Robbins (1985) and Burnetas and Katehakis (1996), respectively. We also investigate the behavior of these algorithms when used with general bounded rewards, showing in particular that they provide significant improvements over the state-of-the-art.

Minimax strategy for Stratified Sampling for Monte Carlo [8]

We consider the problem of stratified sampling for Monte-Carlo integration. We model this problem in a multi-armed bandit setting, where the arms represent the strata, and the goal is to estimate a weighted average of the mean values of the arms. We propose a strategy that samples the arms according to an upper bound on their standard deviations and compare its estimation quality to an ideal allocation that would know the standard deviations of the strata. We provide two pseudo-regret analyses: a distribution-dependent bound of order $O(n^{-3/2})$ that depends on a measure of the disparity of the strata, and a distribution-free bound $O(n^{-4/3})$ that does not. We also provide the first problem independent (minimax) lower bound for this problem and demonstrate that MC-UCB matches this lower bound both in terms of number of samples n and in terms of number of strata K. Finally, we link the pseudo-regret with the difference between the mean squared error on the estimated weighted average of the mean values of the arms, and the optimal oracle strategy: this provides us also with a problem dependent and a problem independent rate for this measure of performance and, as a corollary, asymptotic optimality.

Upper-Confidence-Bound Algorithms for Active Learning in Multi-Armed Bandits [7]

In this paper, we study the problem of estimating uniformly well the mean values of several distributions given a finite budget of samples. If the variance of the distributions were known, one could design an optimal sampling strategy by collecting a number of independent samples per distribution that is proportional to their variance. However, in the more realistic case where the distributions are not known in advance, one needs to design adaptive sampling strategies in order to select which distribution to sample from according to the previously observed samples. We describe two strategies based on pulling the distributions a number of times that is proportional to a high-probability upper-confidence-bound on their variance (built from previous observed samples) and report a finite-sample performance analysis on the excess estimation error compared to the optimal allocation. We show that the performance of these allocation strategies depends not only on the variances but also on the full shape of the distributions.

Bandit Algorithms boost motor-task selection for Brain Computer Interfaces [32] [10]

Brain-computer interfaces (BCI) allow users to "communicate" with a computer without using their muscles. BCI based on sensori-motor rhythms use imaginary motor tasks, such as moving the right or left hand, to send control signals. The performances of a BCI can vary greatly across users but also depend on the tasks used, making the problem of appropriate task selection an important issue. This study presents a new procedure to automatically select as fast as possible a discriminant motor task for a brain-controlled button. We develop for this purpose an adaptive algorithm, *UCB-classif*, based on the stochastic bandit theory. This shortens the training stage, thereby allowing the exploration of a greater variety of tasks. By not wasting time on

inefficient tasks, and focusing on the most promising ones, this algorithm results in a faster task selection and a more efficient use of the BCI training session. Comparing the proposed method to the standard practice in task selection, for a fixed time budget, *UCB-classif* leads to an improved classification rate, and for a fixed classification rate, to a reduction of the time spent in training by 50%.

Adaptive Stratified Sampling for Monte-Carlo integration of Differentiable functions [26]

We consider the problem of adaptive stratified sampling for Monte Carlo integration of a differentiable function given a finite number of evaluations to the function. We construct a sampling scheme that samples more often in regions where the function oscillates more, while allocating the samples such that they are well spread on the domain (this notion shares similitude with low discrepancy). We prove that the estimate returned by the algorithm is almost similarly accurate as the estimate that an optimal oracle strategy (that would know the variations of the function *everywhere*) would return, and provide a finite-sample analysis.

Risk-Aversion in Multi-Armed Bandits [46]

In stochastic multi-armed bandits the objective is to solve the exploration-exploitation dilemma and ultimately maximize the expected reward. Nonetheless, in many practical problems, maximizing the expected reward is not the most desirable objective. In this paper, we introduce a novel setting based on the principle of risk-aversion where the objective is to compete against the arm with the best risk-return trade-off. This setting proves to be intrinsically more difficult than the standard multi-arm bandit setting due in part to an exploration risk which introduces a regret associated to the variability of an algorithm. Using variance as a measure of risk, we introduce two new algorithms, we investigate their theoretical guarantees, and we report preliminary empirical results.

Bandit Theory meets Compressed Sesing for high dimensional Stochastic Linear Bandit [27]

We consider a linear stochastic bandit problem where the dimension K of the unknown parameter θ is larger than the sampling budget n. In such cases, it is in general impossible to derive sub-linear regret bounds since usual linear bandit algorithms have a regret in $O(K\sqrt{n})$. In this paper we assume that θ is S-sparse, i.e. has at most S non-zero components, and that the space of arms is the unit ball for the L_2 norm. We combine ideas from Compressed Sensing and Bandit Theory and derive an algorithm with a regret bound in $O(S\sqrt{n})$. We detail an application to the problem of optimizing a function that depends on many variables but among which only a small number of them (initially unknown) are relevant.

Thompson Sampling: an Asymptotically Optimal Finite Time Analysis [38]

The question of the optimality of Thompson Sampling for solving the stochastic multi-armed bandit problem had been open since 1933. In this paper we answer it positively for the case of Bernoulli rewards by providing the first finite-time analysis that matches the asymptotic rate given in the Lai and Robbins lower bound for the cumulative regret. The proof is accompanied by a numerical comparison with other optimal policies, experiments that have been lacking in the literature until now for the Bernoulli case.

Regret bounds for Restless Markov Bandits [43]

We consider the restless Markov bandit problem, in which the state of each arm evolves according to a Markov process independently of the learner's actions. We suggest an algorithm that after T steps achieves $O(\sqrt{T})$ regret with respect to the best policy that knows the distributions of all arms. No assumptions on the Markov chains are made except that they are irreducible. In addition, we show that index-based policies are necessarily suboptimal for the considered problem.

Minimax number of strata for online Stratified Sampling given Noisy Samples [28]

We consider the problem of online stratified sampling for Monte Carlo integration of a function given a finite budget of n noisy evaluations to the function. More precisely we focus on the problem of choosing the number of strata K as a function of the budget n. We provide asymptotic and finite-time results on how an oracle that has access to the function would choose the number of strata optimally. In addition we prove a lower bound on the learning rate for the problem of stratified Monte-Carlo. As a result, we are able to state, by improving the bound on its performance, that algorithm MC-UCB, is minimax optimal both in terms of the number of samples n and the number of strata K, up to a $\log(nK)$ factor. This enables to deduce a minimax optimal bound on the difference between the performance of the estimate output by MC-UCB, and the performance of the estimate output by the best oracle static strategy, on the class of Holder continuous functions, and up to a factor $\log(n)$.

Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence [33]

We study the problem of identifying the best arm(s) in the stochastic multi-armed bandit setting. This problem has been studied in the literature from two different perspectives: fixed budget and fixed confidence. We propose a unifying approach that leads to a meta-algorithm called unified gap-based exploration (UGapE), with a common structure and similar theoretical analysis for these two settings. We prove a performance bound for the two versions of the algorithm showing that the two problems are characterized by the same notion of complexity. We also show how the UGapE algorithm as well as its theoretical analysis can be extended to take into account the variance of the arms and to multiple bandits. Finally, we evaluate the performance of UGapE and compare it with a number of existing fixed budget and fixed confidence algorithms.

6.2. Statistical Analysis of Time Series

6.2.1. Prediction of Sequences of Structured and Unstructured Data

Reducing statistical time-series problems to binary classification [45]

We show how binary classification methods developed to work on i.i.d. data can be used for solving statistical problems that are seemingly unrelated to classification and concern highly-dependent time series. Specifically, the problems of time-series clustering, homogeneity testing and the three-sample problem are addressed. The algorithms that we construct for solving these problems are based on a new metric between time-series distributions, which can be evaluated using binary classification methods. Universal consistency of the proposed algorithms is proven under most general assumptions. The theoretical results are illustrated with experiments on synthetic and real-world data.

6.2.2. Hypothesis Testing

Testing composite hypotheses about discrete ergodic processes [21]

Given a discrete-valued sample $X_1, \cdots, X_n$ we wish to decide whether it was generated by a distribution belonging to a family H_0, or it was generated by a distribution belonging to a family H_1. In this work we assume that all distributions are stationary ergodic, and do not make any further assumptions (in particular, no independence or mixing rate assumptions). We find some necessary and some sufficient conditions, formulated in terms of the topological properties of H_0 and H_1, for the existence of a consistent test. For the case when H_1 is the complement of H_0 (to the set of all stationary ergodic processes) these necessary and sufficient conditions coincide, thereby providing a complete characterization of families of processes membership to which can be consistently tested, against their complement, based on sampling. This criterion includes as special cases several known and some new results on testing for membership to various parametric families, as well as testing identity, independence, and other hypotheses.

Uniform hypothesis testing for finite-valued stationary processes [22]

Given a discrete-valued sample $X_1, \cdots, X_n$ we wish to decide whether it was generated by a distribution belonging to a family H_0, or it was generated by a distribution belonging to a family H_1. In this work we assume that all distributions are stationary ergodic, and do not make any further assumptions (e.g. no independence or mixing rate assumptions). We would like to have a test whose probability of error (both Type I and Type II) is uniformly bounded. More precisely, we require that for each ϵ there exist a sample size n such that probability of error is upper-bounded by ϵ for samples longer than n. We find some necessary and some sufficient conditions on H_0 and H_1 under which a consistent test (with this notion of consistency) exists. These conditions are topological, with respect to the topology of distributional distance.

6.2.3. Change Point Analysis

Locating Changes in Highly Dependent Data with Unknown Number of Change Points [39]

The problem of multiple change point estimation is considered for sequences with unknown number of change points. A consistency framework is suggested that is suitable for highly dependent time-series, and an asymptotically consistent algorithm is proposed. In order for the consistency to be established the only assumption required is that the data is generated by stationary ergodic time-series distributions. No modeling, independence or parametric assumptions are made; the data are allowed to be dependent and the dependence can be of arbitrary form. The theoretical results are complemented with experimental evaluations.

6.2.4. Clustering Time Series, Online and Offline

Online Clustering of Processes [40]

The problem of online clustering is considered in the case where each data point is a sequence generated by a stationary ergodic process. Data arrive in an online fashion so that the sample received at every time-step is either a continuation of some previously received sequence or a new sequence. The dependence between the sequences can be arbitrary. No parametric or independence assumptions are made; the only assumption is that the marginal distribution of each sequence is stationary and ergodic. A novel, computationally efficient algorithm is proposed and is shown to be asymptotically consistent (under a natural notion of consistency). The performance of the proposed algorithm is evaluated on simulated data, as well as on real datasets (motion classification).

Incremental Spectral Clustering with the Normalised Laplacian [52]

Partitioning a graph into groups of vertices such that those within each group are more densely connected than vertices assigned to different groups, known as graph clustering, is often used to gain insight into the organization of large scale networks and for visualization purposes. Whereas a large number of dedicated techniques have been recently proposed for static graphs, the design of on-line graph clustering methods tailored for evolving networks is a challenging problem, and much less documented in the literature. Motivated by the broad variety of applications concerned, ranging from the study of biological networks to graphs of scientific references through to the exploration of communications networks such as the World Wide Web, it is the main purpose of this paper to introduce a novel, computationally efficient, approach to graph clustering in the evolutionary context. Namely, the method promoted in this article is an incremental eigenvalue solution for the spectral clustering method described by Ng. et al. (2001). Beyond a precise description of its practical implementation and an evaluation of its complexity, its performance is illustrated through numerical experiments, based on datasets modelling the evolution of a HIV epidemic and the purchase history graph of an e-commerce website.

6.2.5. Online Semi-Supervised Learning

Learning from a Single Labeled Face and a Stream of Unlabeled Data [41]

Face recognition from a single image per person is a challenging problem because the training sample is extremely small. We consider a variation of this problem. In our problem, we recognize only one person, and there are no labeled data for any other person. This setting naturally arises in authentication on personal computers and mobile devices, and poses additional challenges because it lacks negative examples. We formalize our problem as one-class classification, and propose and analyze an algorithm that learns a non-parametric model of the face from a single labeled image and a stream of unlabeled data. In many domains, for instance when a person interacts with a computer with a camera, unlabeled data are abundant and easy to utilize. This is the first paper that investigates how these data can help in learning better models in the single-image-per-person setting. Our method is evaluated on a dataset of 43 people and we show that these people can be recognized 90% of time at nearly zero false positives. This recall is 25+% higher than the recall of our best performing baseline. Finally, we conduct a comprehensive sensitivity analysis of our algorithm and provide a guideline for setting its parameters in practice.

6.3. Statistical Learning and Bayesian Analysis

6.3.1. Non-parametric Methods for Function Approximation

Linear Regression with Random Projections [16]

We investigate a method for regression that makes use of a randomly generated subspace G_P (of finite dimension P) of a given large (possibly infinite) dimensional function space F, for example, $L_2([0,1]^d)$. G_P is defined as the span of P random features that are linear combinations of a basis functions of F weighted by random Gaussian i.i.d. coefficients. We show practical motivation for the use of this approach, detail the link that this random projections method share with RKHS and Gaussian objects theory and prove, both in deterministic and random design, approximation error bounds when searching for the best regression function in G_P rather than in F, and derive excess risk bounds for a specific regression algorithm (least squares regression in G_P). This paper stresses the motivation to study such methods, thus the analysis developed is kept simple for explanations purpose and leaves room for future developments.

6.3.2. Nonparametric Bayesian Estimation

DPM pour l'inférence dans les modèles dynamiques non linéaires avec des bruits de mesure alpha-stable [50]

Stable random variables are often use to model impulsive noise; Recently it has be shown that communication at very high frequency suffer from such a noise. Stable noise cannot however be considered as usual noise in estimation processes because the variance does not usually exists nor an analytic expression for the probability density function. In this work we show how to manage such a problem using a bayesian nonparametric approach. We develop a Sequential Monte Carlo based algorithm to realize the estimation in a non linear dynamical system. The measurement noise is a non-stationnary stable process and it is modeled using a Dirichlet Process Mixture.

6.3.3. Random Finite Sets for Multisensor Multitarget Tracking

Multi-sensor PHD filtering with application to sensor management [2]

The aim of multi-object filtering is to address the multiple target detection and/or tracking problem. This thesis focuses on the Probability Hypothesis Density (PHD) filter, a well-known tractable approximation of the Random Finite Set (RFS) filter when the observation process is realized by a single sensor. The first part proposes the rigorous construction of the exact multi-sensor PHD filter and its simplified expression, without approximation, through a joint partitioning of the target state space and the sensors. With this new method, the exact multi-sensor PHD can be propagated in simple surveillance scenarii. The second part deals with the sensor management problem in the PHD framework. At each iteration, the Balanced Explorer and Tracker (BET) builds a prediction of the posterior multi-sensor PHD thanks to the Predicted Ideal Measurement Set (PIMS) and produces a multi-sensor control according to a few simple operational principles adapted to surveillance activities

6.4. Applications

6.4.1. Signal Processing

Dirichlet Process Mixtures for Density Estimation in Dynamic Nonlinear Modeling: Application to GPS Positioning in Urban Canyons [19]

In global positioning systems (GPS), classical localization algorithms assume, when the signal is received from the satellite in line-of-sight (LOS) environment, that the pseudorange error distribution is Gaussian. Such assumption is in some way very restrictive since a random error in the pseudorange measure with an unknown distribution form is always induced in constrained environments especially in urban canyons due to multipath/masking effects. In order to ensure high accuracy positioning, a good estimation of the observation error in these cases is required. To address this, an attractive flexible Bayesian nonparametric noise model based on Dirichlet process mixtures (DPM) is introduced. Since the considered positioning problem involves elements of non-Gaussianity and nonlinearity and besides, it should be processed on-line, the suitability of the proposed modeling scheme in a joint state/parameter estimation problem is handled by an efficient Rao-Blackwellized particle filter (RBPF). Our approach is illustrated on a data analysis task dealing with joint estimation of vehicles positions and pseudorange errors in a global navigation satellite system (GNSS)-based localization context where the GPS information may be inaccurate because of hard reception conditions.

Dislocation detection in field environments: A belief functions contribution [20]

Dislocation is defined as the change between discrete sequential locations of critical items in field environments such as large construction projects. Dislocations on large sites of materials and critical items for which discrete time position estimates are available represent critical state changes. The ability to detect dislocations automatically for tens of thousands of items can ultimately improve project performance significantly. Detecting these dislocations in a noisy information environment where low cost radio frequency identification tags are attached to each piece of material, and the material is moved sometimes only a few meters, is the main focus of this study. We propose in this paper a method developed in the frame of belief functions to detect dislocations. The belief function framework is well-suited for such a problem where both uncertainty and imprecision are inherent to the problem. We also show how to deal with the calculations. This method has been implemented in a controlled experimental setting. The results of these experiments show the ability of the proposed method to detect materials dislocation over the site reliably. Broader application of this approach to both animate and inanimate objects is possible.

Towards dictionary learning from images with non Gaussian noise [29]

We address the problem of image dictionary learning from noisy images with non Gaussian noise. This problem is difficult. As a first step, we consider the extreme sparse code given by vector quantization, i.e. each pixel is finally associated to 1 single atom. For Gaussian noise, the natural solution is K-means clustering using the sum of the squares of differences between gray levels as the dissimilarity measure between patches. For non Gaussian noises (Poisson, Gamma,...), a new measure of dissimilarity between noisy patches is necessary. We study the use of the generalized likelihood ratios (GLR) recently introduced by Deledalle et al. 2012 to compare non Gaussian noisy patches. We propose a K-medoids algorithm generalizing the usual Linde-Buzo-Gray K-means using the GLR based dissimilarity measure. We obtain a vector quantization which provides a dictionary that can be very large and redundant. We illustrate our approach by dictionaries learnt from images featuring non Gaussian noise, and present preliminary denoising results.

6.4.2. Medical Applications

Outlier detection for patient monitoring and alerting. [12]

We develop and evaluate a data-driven approach for detecting unusual (anomalous) patient-management decisions using past patient cases stored in electronic health records (EHRs). Our hypothesis is that a patient-management decision that is unusual with respect to past patient care may be due to an error and that it is worthwhile to generate an alert if such a decision is encountered. We evaluate this hypothesis using data obtained from EHRs of 4486 post-cardiac surgical patients and a subset of 222 alerts generated from the data. We base the evaluation on the opinions of a panel of experts. The results of the study support our hypothesis that the outlier-based alerting can lead to promising true alert rates. We observed true alert rates that ranged from 25% to 66% for a variety of patient-management actions, with 66% corresponding to the strongest outliers.

6.4.3. Web Mining

Managing advertising campaigns – an approximate planning approach [11]

We consider the problem of displaying commercial advertisements on web pages, in the "cost per click" model. The advertisement server has to learn the appeal of each type of visitor for the different advertisements in order to maximize the profit. Advertisements have constraints such as a certain number of clicks to draw, as well as a lifetime. This problem is thus inherently dynamic, and intimately combines combinatorial and statistical issues. To set the stage, it is also noteworthy that we deal with very rare events of interest, since the base probability of one click is in the order of 10^4. Different approaches may be thought of, ranging from computationally demanding ones (use of Markov decision processes, or stochastic programming) to very fast ones.We introduce NOSEED, an adaptive policy learning algorithm based on a combination of linear programming and multi-arm bandits. We also propose a way to evaluate the extent to which we have to handle the constraints (which is directly related to the computation cost). We investigate the performance of our system through simulations on a realistic model designed with an important commercial web actor.

ICML Exploration & Exploitation challenge: Keep it simple! [18]

Recommendation has become a key feature in the economy of a lot of companies (online shopping, search engines...). There is a lot of work going on regarding recommender systems and there is still a lot to do to improve them. Indeed nowadays in many companies most of the job is done by hand. Moreover even when a supposedly smart recommender system is designed, it is hard to evaluate it without using real audience which obviously involves economic issues. The ICML Exploration & Exploitation challenge is an attempt to make people propose efficient recommendation techniques and particularly focuses on limited computational resources. The challenge also proposes a framework to address the problem of evaluating a recommendation algorithm with real data. We took part in this challenge and achieved the best performances; this paper aims at reporting on this achievement; we also discuss the evaluation process and propose a better one for future challenges of the same kind.

6.4.4. Games

CLOP: Confident Local Optimization for Noisy Black-Box Parameter Tuning [30]

Artificial intelligence in games often leads to the problem of parameter tuning. Some heuristics may have coefficients, and they should be tuned to maximize the win rate of the program. A possible approach is to build local quadratic models of the win rate as a function of program parameters. Many local regression algorithms have already been proposed for this task, but they are usually not robust enough to deal automatically and efficiently with very noisy outputs and non-negative Hessians. The CLOP principle, which stands for Confident Local OPtimization, is a new approach to local regression that overcomes all these problems in a simple and efficient way. CLOP discards samples whose estimated value is confidently inferior to the mean of all samples. Experiments demonstrate that, when the function to be optimized is smooth, this method outperforms all other tested algorithms.

6.5. Other Results

Sequential approaches for learning datum-wise sparse representations [9]

In supervised classification, data representation is usually considered at the dataset level: one looks for the "best" representation of data assuming it to be the same for all the data in the data space. We propose a different approach where the representations used for classification are tailored to each datum in the data space. One immediate goal is to obtain sparse datum-wise representations: our approach learns to build a representation specific to each datum that contains only a small subset of the features, thus allowing classification to be fast and efficient. This representation is obtained by way of a sequential decision process that sequentially chooses which features to acquire before classifying a particular point; this process is learned through algorithms based on Reinforcement Learning. The proposed method performs well on an ensemble of medium-sized sparse classification problems. It offers an alternative to global sparsity approaches, and is a natural framework for sequential classification problems. The method extends easily to a whole family of sparsity-related problems which would otherwise require developing specific solutions. This is the case in particular for cost-sensitive and limited-budget classification, where feature acquisition is costly and is often performed sequentially. Finally, our approach can handle non-differentiable loss functions or combinatorial optimization encountered in more complex feature selection problems.

Multiple Operator-valued Kernel Learning [60]

Positive definite operator-valued kernels generalize the well-known notion of reproducing kernels, and are naturally adapted to multi-output learning situations. This paper addresses the problem of learning a finite linear combination of infinite-dimensional operator-valued kernels which are suitable for extending functional data analysis methods to nonlinear contexts. We study this problem in the case of kernel ridge regression for functional responses with an lr-norm constraint on the combination coefficients. The resulting optimization problem is more involved than those of multiple scalar-valued kernel learning since operator-valued kernels pose more technical and theoretical issues. We propose a multiple operator-valued kernel learning algorithm based on solving a system of linear operator equations by using a block coordinatedescent procedure. We experimentally validate our approach on a functional regression task in the context of finger movement prediction in brain-computer interfaces.

7. Bilateral Contracts and Grants with Industry

7.1. Orange Labs

Participant: Jérémie Mary.

There has been various activities between SEQUEL and Orange Labs.

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

7.2. Effigenie

Participant: Jérémie Mary.

We are currenly working on better prediction of news websites audiences in order to plan some better strategies for marketing services. A prediction module should be produced in 2013.

7.3. Squoring Technology

Participants: Boris Baldassari, Philippe Preux.

Boris Baldassari has been hired by Squoring Technology (Toulouse) as a PhD student in May 2011. He works on the use of machine learning to improve the quality of the software development process. During his first year as a PhD student, Boris investigated the existing norms and measures of quality of software development process. He also dedicated some times to gather some relevant datasets, which are made of either the sequence of source code releases over a multi-years period, or all the versions stored on an svn repository (svn or alike). Information from mailing-lists (bugs, support, ...) may also be part of these datasets. Tools in machine learning capable of dealing with this sort of data have also been investigated. Goals that may be reached in this endeavor have also been precised.

7.4. TBS

Participants: Jérémie Mary, Philippe Preux.

A new project has started on September 2012 in collaboration with the TBS company. The goal is to understand and predict the audiences of some news related websites. These websites tend to present an ergodic frequentation with respect to a context. The main goal is to separate the effect of the context (big events, election, ...) and the impact of the policies of the news websites. This research is done using data from major french media websites and also involves research of tendances on the web (like Google Trends/ Google Flu). Used algorithms mix methods from time series prediction (ARIMA and MARSS models) and some machine learning methods (L1 penalization, SVM).

7.5. Unbalance Corporation

Participant: Rémi Coulom.

Unbalance Corporation (http://www.unbalance.co.jp/) is a Japanese publisher of game software. We have two license agreements with this company, for the games of Go and Hanafuda.

8. Partnerships and Cooperations

8.1. Regional Initiatives

8.1.1. Connectome, and large graph mining

Participant: Philippe Preux.

- *Title*: Connectome and epilepsy
- *Type*: No funding yet (self-funded project)
- *Coordinator*: Louise Tyvaert, Department of clinical neurophysiology, CHRU Lille, Université de Lille 2, France
- *Others partners*: Mostrare, Inria Lille
- *Duration*: Began in spring 2012
- *Abstract*: The long term goal of this collaboration is to investigate the use of machine learning tools to analyse connectomes, and possibly related EEG signals, to determine, for a given patient, the region of the brain from which originate epilepsy strokes. As a first step, we concentrate on connectome, that is a graph representation of the connectivity in the brain. We study the properties of these graphs from a formal point of view, and try to match these properties with brain activity, and brain disorders.
- *Activity Report*: being a multi-disciplinary project, the first thing was to understanding each others. Connectomes having been acquired at the hospital via MRI and image processing, the resulting graphs have been processed using a spatially regularized spectral clustering approach; we were able to recover well-known brain areas automatically. Indeed, one of the first issues to clarify is the relevance of the graph representation of these MRI data (connectomes), an issue unclear in the medicine community. These first results have been submitted for publication at the IEEE 2013 symposium on Bio-Imaging (ISBI'2013).

8.2. National Initiatives

8.2.1. DGA/Thales

Participants: Emmanuel Duflos, Philippe Vanheeghe, Emmanuel Delande.

- *Title*: Multi-sensor PHD filtering with application to sensor management (http://www.theses.fr/2012ECLI0001)
- *Type*: PhD grant
- *Coordinator*: LAGIS - Inria Lille - Nord Europe (SequeL)
- *Others partners*: DGA and Thales Communications
- *Web site*: http://www.theses.fr/2012ECLI0001
- *Duration*: *EDIT THIS: 3 years*
- *Abstract*: The defense of this PhD thesis was held in January 2012.
- *Activity Report*:

8.2.2. ANR-Lampada

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

- *Title*: Learning Algorithms, Models an sPArse representations for structured DAta
- *Type*: National Research Agency (ANR-09-EMER-007)
- *Coordinator*: Inria Lille - Nord Europe (Mostrare)
- *Others partners*: Laboratoire d'Informatique Fondamentale de Marseille, Laboratoire Hubert Curien ; Saint Etienne, Laboratoire d'Informatique de Paris 6.
- *Web site*: http://lampada.gforge.inria.fr/
- *Duration*: ends mid-2014
- *Abstract*: Lampada is a fundamental research project on machine learning and structured data. It focuses on scaling learning algorithms to handle large sets of complex data. The main challenges are 1) high dimension learning problems, 2) large sets of data and 3) dynamics of data. Complex data we consider are evolving and composed of parts in some relations. Representations of these data embed both structure and content information and are typically large sequences, trees and graphs. The main application domains are web2, social networks and biological data.

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

  The representation research topic includes condensed data representation, sampling, prototype selection and representation of streams of data. Machine learning methods include edit distance learning, reinforcement learning and incremental methods, density estimation of structured data and learning on streams.
- *Activity Report*: Philippe Preux has collaborated with Ludovic Denoyer and Gabriel Dulac-Arnold from LIP'6 to investigate further the idea of datum-wise representation, introduced in 2011, and originally published at ECML/PKDD'2011. This eventually led to a deeped presentation in the *Machine Learning Journal*.

  They also studied the reinforcement learning problem in the case of a large but not infinite number of actions (hundreds, or thousands discrete actions). They introduced the use of Error-correcting output codes to deal with this setting, proposed, and studied two RL algorithms that take advantage of an ECOC-based representation of actions. The idea was published at ECML/PKDD'2012 and other conferences (EWRL workshop held as part of the ICML conference, and French ones).

Hachem Kadri and Philippe Preux have continued their work on machine learning for functional data. They introduced an algorithm for multiple operators learning. Along with Mohammad Ghavamzadeh, they only introduced a operator-based aprroach for structured output.

Danil Ryabko and colleagues have obtained new results on nonparametric clustering of time-series data. In particular, a fully online clustering algorithm has been developed; we have also shown how to use binary classification methods for clustering time series.

8.2.3. ANR EXPLO-RA

Participants: Alexandra Carpentier, Mohammad Ghavamzadeh, Jean-François Hren, Alessandro Lazaric, Rémi Munos, Daniil Ryabko.

- Title: EXPLOration - EXPLOitation for efficient Resource Allocation with Applications to optimization, control, learning, and games
- Type: National Research Agency
- Coordinator: Inria Lille - Nord Europe (SequeL, Rémi Munos)
- Others partners: Inria Saclay - Ile de France (TAO), HEC Paris (GREGHEC), Ecole Nationale des Ponts et Chaussées (CERTIS), Université Paris 5 (CRIP5), Université Paris Dauphine (LAMSADE).
- Duration: 2008-2012.
- See also: https://sites.google.com/site/anrexplora/
- Activity Report: We developed bandit algorithm for planning in Markov Decision Processes based on the optimism in the face of uncertainty principle.

8.2.4. ANR CO-ADAPT

Participants: Alexandra Carpentier, Rémi Munos.

- *Title*: Brain computer co-adaptation for better interfaces
- *Type*: National Research Agency
- *Coordinator*: Maureen Clerc
- *Other Partners*: Inria Odyssee project (Maureen Clerc), the INSERM U821 team (Olivier Bertrand), the Laboratory of Neurobiology of Cognition (CNRS) (Boris Burle) and the laboratory of Analysis, topology and probabilities (CNRS and University of Provence) (Bruno Torresani).
- *Web site*: https://twiki-sop.inria.fr/twiki/bin/view/Projets/Athena/CoAdapt/WebHome
- *Duration*: 2009-2013
- *Abstract*: The aim of CoAdapt is to propose new directions for BCI design, by modeling explicitly the co-adaptation taking place between the user and the system. The goal of CoAdapt is to study the co-adaptation between a user and a BCI system in the course of training and operation. The quality of the interface will be judged according to several criteria (reliability, learning curve, error correction, bit rate). BCI will be considered under a joint perspective: the user's and the system's. From the user's brain activity, features must be extracted, and translated into commands to drive the BCI system. From the point of view of the system, it is important to devise adaptive learning strategies, because the brain activity is not stable in time. How to adapt the features in the course of BCI operation is a difficult and important topic of research. We will investigate Reinforcement Learning (RL) techniques to address the above questions.
- *Activity Report*: See https://twiki-sop.inria.fr/twiki/bin/view/Projets/Athena/CoAdapt/WebHome

8.2.5. ANR AMATIS

Participant: Pierre Chainais.

- *Title*: Multifractal Analysis and Applications to Signal and Image Processing
- *Type*: National Research Agency
- *Coordinator*: Univ. Paris-Est-Créteil (S. Jaffard)
- *Duration*: 2011-2015
- *Other Partners*: Univ. Paris-Est Créteil, Univ. Sciences et Technologies de Lille and Inria (Lille, ENST (Telechom ParisTech), Univ. Blaise Pascal (Clermont-Ferrand), and Univ. Bretagne Sud (Vannes), Statistical Signal Processing group at the Physics Department at the Ecole Normale Supérieure de Lyon, one researcher from the Math. Department of Institut National des Sciences Appliquees de Lyon and two researchers from the Laboratoire d'Analyse, Topologie et Probabilités (LAPT) of Aix-Marseille University.
- *Abstract*: Multifractal analysis refers to two concepts of different natures : On the theoretical side, it corresponds to pointwise singularity characterization and fractional dimension determination ; on the applied side, it is associated with scale invariance characterization, involving a family of parameters, the scaling function, used in classification or model selection. Following the seminal ideas of Parisi and Frisch in the mid-80s, these two components are usually related by a Legendre transform, stemming from a heuristic argument relying on large deviation and statistical thermodynamics principles : The multifractal formalism. This led to new theoretical approaches for the study of singularities of functions and measures, as well as efficient tools for classification and models selection, that allowed to settle longstanding issues (e.g., concerning the modeling of fully developed turbulence). Though this formalism had been shown to hold for large classes of functions of widely different origins, the generality of its level of validity remains an open issue. Despite its popularity in applications, the interactions between theoretical developments and applications are unsatisfactory. Its use in image processing for instance is still in its infancy. This is partly due to discrepancy between the theoretical contributions mostly grounded in functional analysis and geometric measure theory, and applications naturally implying a stochastic or statistical framework. The AMATIS project aims at addressing these issues, by proposing a consistent and documented framework combining different theoretical approaches and bridging the gap towards applications. To that end, it will both address a number of challenging theoretical issues and devote significant efforts to elaborating a WEB platform with softwares and documentation. It will combine the efforts of mathematicians with those of physicists and experts in signal and image processing. Dissemination among and interactions between scientific fields are also intended via the organization of summer schools and workshop.
- *Activity Report*: a collaboration with P. Bas (CR CNRS, LAGIS) has started on the steganalysis of textured images. While steganography aims at hiding a message within some support, e.g. a numerical image, steganalysis aims at detecting the presence or not of any hidden message in the support. Steganalysis involves two main tasks: first identify relevant features which may be sensitive to the presence of a hidden message, then use supervised classification to build a detector. While the steganalysis of usual images has been well studied, the case of textured images, for which multifractal models may be relevant, is much more difficult. Indeed, textured images have a rich and disordered content which favors hiding information in an unperceptible manner. A student internship of 6 months at Master level has finished in November. The purpose was to explore the potential of new multiscale wavelet based discriminant features for steganalysis.

8.2.6. National Partners

- Inria Nancy - Grand Est, Team MAIA, France.
  - Bruno Scherrer *Collaborator*
    We have had collaboration on the topics of *approximate dynamic programming and statistical learning* and *high-dimensional reinforcement learning* this year. On the first topic, we have published a conference paper [47] and a technical report [62], and on the second one we have published a conference paper [36] together.

- Supélec, IMS Research Group, Metz, France.
  - Matthieu Geist *Collaborator*
    We have had collaboration on the topics of *approximate dynamic programming and statistical learning* and *high-dimensional reinforcement learning* this year. On the first topic, we have published a conference paper [47] and a technical report [62], and on the second one we have published a conference paper [36] together.
- LIP'6, UPMC, Paris, France.
  - Ludovic Denoyer *Collaborator*
    We have a collaboration on the topic of *reinforcement learning, sparse representation*. We have worked on the datum-wise representation of data, as well as the handling of large but non infinite sets of actions. See section 8.2.2 for further details.

8.3. European Initiatives

8.3.1. FP7 Projects

PASCAL-2

Participants: the whole SEQUEL team is involved.

- *Title*: Pattern Analysis, Statistical Modeling, and Computational Learning
- *Type*: Cooperation (ICT), Network of Excellence (NoE)
- *Coordinator*: Univ. Southampton
- *Others partners*: Many european organizations, universities, and research centers.
- *Web site*: http://www.pascal-network.org/
- *Duration*: March 2008 - February 2013

PASCAL-2 Pump Priming Programme

Participants: Mohammad Ghavamzadeh, Rémi Munos.

- *Title*: Sparse Reinforcement Learning in High Dimensions
- *Type*: PASCAL-2 Pump Priming Programme
- *Partners*: Inria Lille - Nord Europe, Shie Mannor (Technion, Israel)
- *Web site*: http://sites.google.com/site/sparserl/home
- *Duration*: November 2009 - September 2012
- *Abstract*: With the explosive growth and ever increasing complexity of data, developing theory and algorithms for learning with high-dimensional data has become an important challenge in statistical machine learning. Although significant advances have been made in recent years, most of the research efforts have been focused on supervised learning problems. We propose to design, analyze, and implement reinforcement learning algorithms for high-dimensional domains. We will investigate the possibility of using the recent results in l1-regularization and compressive sensing in reinforcement learning.
- *Activity report*: The project ended early this year. The list of publications obtained within the project is listed at https://sites.google.com/site/sparserl/publications.

CompLACS

Participants: Mohammad Ghavamzadeh, Nathan Korda, Prashanth Lakshmanrao Anantha Padmanabha, Alessandro Lazaric, Rémi Munos, Philippe Preux, Daniil Ryabko, Michal Valko.

- *Title*: Composing Learning for Artificial Cognitive Systems
- *Type*: Cooperation (ICT), Specific Targeted Research Project (STREP)
- *Coordinator*: University College of London
- *Other partners*: University College London, United Kingdom (John Shawe-Taylor, Stephen Hailes, David Silver, Yee Whye Teh), University of Bristol, United Kingdom (Nello Cristianini), Royal Holloway, United Kingdom (Chris Watkins), Radboud Universiteit Nijmegen, The Netherlands (Bert Kappen), Technische Universitat Berlin, Germany (Manfred Opper), Montanuniversitat Leoben, Austria (Peter Auer), Max-Planck Institute of Biological Cybernetics, Germany (Jan Peters).
- *Web site*: http://www.complacs.org/
- *Duration*: March 2011 - February 2015
- *Abstract*: One of the aspirations of machine learning is to develop intelligent systems that can address a wide variety of control problems of many different types. However, although the community has developed successful technologies for many individual problems, these technologies have not previously been integrated into a unified framework. As a result, the technology used to specify, solve and analyse one control problem typically cannot be reused on a different problem. The community has fragmented into a diverse set of specialists with particular solutions to particular problems. The purpose of this project is to develop a unified toolkit for intelligent control in many different problem areas. This toolkit will incorporate many of the most successful approaches to a variety of important control problems within a single framework, including bandit problems, Markov Decision Processes (MDPs), Partially Observable MDPs (POMDPs), continuous stochastic control, and multi-agent systems. In addition, the toolkit will provide methods for the automatic construction of representations and capabilities, which can then be applied to any of these problem types. Finally, the toolkit will provide a generic interface to specifying problems and analysing performance, by mapping intuitive, human-understandable goals into machine-understandable objectives, and by mapping algorithm performance and regret back into human-understandable terms.
- *Activity report*: We worked on WorkPackage 2 (multi-armed bandits and extensions) and we designed hierarchical bandit-based planning algorithms for MDPs and POMDPs.

8.4. International Initiatives

8.4.1. Inria Associate Teams

SEQRL

- *Title*: Decision-making under Uncertainty with Applications to Reinforcement Learning, Control, and Games
- *Inria principal investigator*: Rémi Munos
- *International Partner*:
  - *Institution*: University of Alberta (Canada)
  - *Laboratory*: Department of Computer Science
  - *Principal investigator*: Csaba Szepesvári
- *Duration*: January 2010 - January 2013
- *Website*: http://sites.google.com/site/associateteamualberta/home

- *Abstract*: This associate team aims at bridging researchers from the SequeL team-project at Inria Lille with the Department of Computing Science of the University of Alberta in Canada. Our common interest lies in machine learning, especially reinforcement learning, bandit algorithms and statistical learning with applications to control and computer games. The department of Computing Science at the University of Alberta is internationally renown as a leading research institute on these topics. The research work spans from theory to applications. Grounded on an already existing scientific collaboration, this associate team will make it easier to collaborate further between the two institutes, and thus strengthen this relationship. We foresee that the associate team will boost our collaboration, create new opportunities for financial support, and open-up a long-term fruitful collaboration between the two institutes. The collaboration will be through organizing workshops and exchanging researchers, postdoctoral fellows, and Ph.D. students between the two institutes.
- *Activity report*: This year we had two Ph.D. students from the university of Alberta, Yasin Abbasi and Bernardo Avila Pires, who visited SequeL for six and four weeks, respectively. We send our Ph.D. student Amir Sani to a workshop organized by the university of Alberta and McGill university in Barbados in April. Mohammad Ghavamzadeh had a one week visit to the university of Alberta to work with Csaba Szepesvári and Bernardo Avila Pires.
- *Joint Publications*: We have one conference paper submitted [53] and one in preparation [61] this year.

8.4.2. Inria International Partners

- University of Alberta, Edmonton, Alberta, Canada.
  - Prof. Csaba Szepesvári *Collaborator*
  - Bernardo Avila Pires *Collaborator*
    With Csaba Szepesvári we managed the associate team with the university of Alberta. We have had several visits to SequeL and UAlberta this year. We also have a conference paper [61] on *risk bounds in cost-sensitive multiclass classification* in preparation with Csaba Szepesvári and Bernardo Avila Pires.
- McGill University, Montreal, Quebec, Canada.
  - Prof. Joelle Pineau *Collaborator*
  - Prof. Doina Precup *Collaborator*
  - Amir massoud Farahmand *Collaborator*
    Mohammad Ghavamzadeh and Rémi Munos wrote a proposal with Joelle Pineau, Doina Precup, and Amir Farahmand to start an associate team with the McGill university. Mohammad Ghavamzadeh also have a conference paper submitted [53] on *classification-based approximate policy iteration* with Amir Farahmand and Doina Precup.
- Technion - Israel Institute of Technology, Haifa, Israel.
  - Prof. Shie Mannor *Collaborator*
    Mohammad Ghavamzadeh continued his collaboration with Shie Mannor. This year, we co-authored a book chapter on *Bayesian reinforcement learning* [57].
- University of Waterloo, Waterloo, Ontario, Canada.
  - Prof. Pascal Poupart *Collaborator*
    Mohammad Ghavamzadeh continued his collaboration with Pascal Poupart. This year, we co-authored a book chapter on *Bayesian reinforcement learning* [57].
- University of Waterloo, Waterloo, Ontario, Canada.
  - Prof. Carl Haas *Collaborator*
- University of Waterloo, Waterloo, Ontario, Canada.
  - Prof. Giovani Cascante *Collaborator*

- Politecnico di Milano, Italy.
  - Prof. Marcello Restelli *Collaborator*
  - Prof. Nicola Gatti *Collaborator*
    We continued our collaboration on transfer in reinforcement learning and we developed a novel collaboration focused on the interplay between bandit theory and mechanism design, notably in the sponsored search auction application domain [35].
- Technicolor Research, Palo Alto.
  - Branislav Kveton *Collaborator*
    We have an ongoing collaboration related to the sequential graph-based learning. This involves both theory and the application to industry, such as sequential face recognition. Currently we investigate the problem of face detection from a single labeled face and the streams of unlabeled data.

8.5. International Research Visitors

- Ronald Ortner, from University of Leoben, Austria.
  Period: spent his sabbatical Jan-Oct 2012 with us. Some papers as a result of this collaboration are [43], [44]; some more are under submission.
- Gusztav Morvai, senior research at Budapest University of Technology and Economics.
  Period: Oct 18-24, 2012
- Tor Lattimore, Ph.D. student at Australian National University.
  Period: Nov. 2-9, 2012
- Bernardo Avila Pires
  Period: May 2012 (one month)
  He worked with Mohammad Ghavamzadeh on *risk bounds in cost-sensitive multiclass classification*. The outcome of this collaboration has been a conference paper in preparation [61] so far.
- Joelle Pineau
  Period: September 2012 (one week)
  Prof. Pineau visited SequeL for one week as a part of her sabbatical. During her stay, in addition to have discussions with SequeL team members and giving two talks on her research, she wrote a proposal with Mohammad Ghavamzadeh and Rémi Munos to start an associate team between SequeL and McGill university.
- Pr. Giovanni Cascante, University of Waterloo, Waterloo, Ontario, Canada.
  Period: June 2012
  He worked with Philippe Vanheeghe and Emmanuel Duflos on parameters estimation in acoustic probing in civil engineering. The outcome of this collaboration has been a project master (from November 2012) and a proposition of research project under evaluation the University of Waterloo so far.

8.5.1. Internships

- Louis Dacquet, student at Ecole Centrale Lille.
  Period: April-June 2012.
  He worked with Pierre Chainais on *blind image deconvolution*.
- Alexandre Kazmierowski, student at Ecole Telecom ParisTech.
  Period: June-July 2012.
  He worked with Pierre Chainais and Antoine Gloria (SIMPAF project) on textured models for heterogeneous media and homogeneization theory in PDEs.

- Phuong Nguyen, Ph.D. student at Australian National University.
  Period: 15 February - 30 April 2012
  He worked with Daniil Ryabko on state representation for reinforcement learning. As a result, one paper is submitted and one is being prepared.
- Florian Gas, Student at the Ecole Centrale de Lille, France.
  Period: May 2012 - July 2012.
  He worked with Emmanuel Duflos on foundations of Sequential Monte Carlo Methods in high dimension

9. Dissemination

9.1. Scientific Animation

9.1.1. Awards

Shih-Chieh Huang, supervised by Rémi Coulom received the Taiwan Computer Game Association PhD Thesis Award during the 2012 Taiwan Computer Game Workshop on June 30, 2012.

9.1.2. Tutorials

- *A. Lazaric* and *M. Ghavamzadeh* co-chaired a tutorial on *Statistical Learning Theory in Reinforcement Learning and Approximate Dynamic Programming* at the Twenty-Ninth International Conference on Machine Learning (ICML), 2012, which was held in Edinburgh, Scotland in June. Here is the webpage of the tutorial
  http://chercheurs.lille.inria.fr/~ghavamza/ICML2012-Tutorial.html

9.1.3. Workshops and Schools

- *J. Mary* co-organized the "New Challenges for Exploration and Exploitation" workshop and competition together with A. Garivier, L. Li, R. Munos, O. Nicol, R. Ortner, and Ph. Preux.
- *H. Kadri* was the main organizer of the ICML workshop on "Object, functional and structured data: towards next generation kernel-based methods" along with Fl. d'Alché-Buc, M. Pontil, and A. Rakotomamonjy.
- *E. Duflos* co-organized the *one-day workshop on Non Parametric Bayesian for Signal and Image Processing* () in Paris (in the frame of the GDR ISIS), with François Caron. The guest speaker was Mickael Jordan from the University of Berkeley.

9.1.4. Invited Talks

- *P. Chainais*, Journées Bordelaises d'Analyse Mathématique des Images, Bordeaux, Host: Prof. J.F. Aujol & C. Dossal (November 2012).
- *P. Chainais*, Nat'Images, Nice, Host: G. Peyré (July 2012).
- *M. Ghavamzadeh*, University of Waterloo, Canada - AI Seminar, Host: Prof. Pascal Poupart (2012).
- *M. Ghavamzadeh*, McGill University, Canada - School of Computer Science, Host: Prof. Joelle Pineau (2012).
- *M. Ghavamzadeh*, University of Alberta, Canada - AI Seminar, Host: Prof. Csaba Szepesvári (2012).
- *M. Ghavamzadeh*, Workshop on "Large-Scale Online Learning and Decision-Making", London (2012).
- *D. Ryabko*, The Fifth Workshop on Information Theoretic Methods in Science and Engineering (WITMSE 2012), Amsterdam, The Netherlands, Aug. 2012s.
- *Ph. Preux*, Université de Clermont-Ferrand, June 2012.
- *M. Valko*, University of Oxford, UK, Host: David Silver (April 2012).

- *M. Valko*, Large-scale Online Learning and Decision Making, UK, Host: Prof. Marc Tommasi (April 2012).
- *M. Valko*, LAMPADA workshop, France, Host: Jakub Zavodny (July 2012).
- *A. Lazaric*, Politecnico di Milano, Italy - AI Seminar, Host: Prof. Nicola Gatti (April 2012).

9.1.5. Review Activities

- ***Participation to the program committees of international conferences***
  - International Conference on Patter Recognition Applications and Methods (ICPRAM 2012)
  - Algorithmic Learning Theory (ALT 2012)
  - AAAI Conference on Artificial Intelligence (AAAI 2012)
  - European Workshop on Reinforcement Learning (EWRL 2012)
  - Annual Conference on Neural Information Processing Systems (NIPS 2012)
  - International Conference on Artificial Intelligence and Statistics (AISTATS 2012)
  - European Conference on Machine Learning (ECML 2012)
  - International Conference on Machine Learning (ICML 2012 and 2013)
  - International Conference on Uncertainty in Artificial Intelligence (UAI 2012)
  - French Conference on Planning, Decision-making, and Learning in Control Systems (JFPDA 2012)
  - FUSION 2012
- ***International journal and conference reviewing activities*** (in addition to the conferences in which we belong to the PC)
  - IEEE Transactions on Image Processing
  - Journal of Statistical Physics
  - Digital Signal Processing
  - IEEE Statistical Signal Processing SSP'2012
  - European Signal Processing Conference EUSIPCO 2012
  - IEEE Transactions on Information Theory
  - Annual Conference on Neural Information Processing Systems (NIPS 2012)
  - International Conference on Machine Learning (ICML 2012)
  - European Conference on Machine Learning (ECML 2012)
  - Uncertainty in Artificial Intelligence (UAI 2012)
  - Machine Learning Journal (MLJ)
  - Journal of Machine Learning Research (JMLR)
  - Journal of Artificial Intelligence Research (JAIR)
  - IEEE Transactions on Automatic Control (TAC)
  - IEEE Transactions of Signal Processing
  - Journal of Autonomous Agents and Multi-Agent Systems (JAAMAS)

9.1.6. Evaluation activities, expertise

- *P. Chainais* is a grant proposal reviewer for the ANR SIMI2.
- *Ph. Preux* is expert for the AERES, ANR, ANRT, and CNRS.
- *M. Ghavamzadeh* is in the Editorial Board Member of Machine Learning Journal (MLJ, 2011-present).

- *M. Ghavamzadeh* is in the Steering Committee Member of the European Workshop on Reinforcement Learning (EWRL, 2011-present).
- *P. Preux*, *R. Gaudel* and *J. Mary* are experts for *Crédit Impôt Recherche* (CIR).
- *E. Duflos* is a project proposal reviewer for ANR.

9.1.7. Other Scientific Activities

- *R. Munos* is Vice Président du Comité des Projets at Inria Lille-Nord Europe since September 2011.
- *D. Ryabko* is a member of COST-GTRI committee at Inria.
- *D. Ryabko* is a general advisor at Inria Lille.
- *R. Gaudel* manages the proml diffusion list.
- *E. Duflos* is Director of Research of Ecole Centrale de Lille since September 2011.
- *E. Duflos* is the Head of the Signal and Image Team of LAGIS (UMR CNRS 8219) since January 2012.
- *R. Gaudel* is board member of LIFL.

9.2. Teaching

- *A. Lazaric*, PhD, “Advanced topics in Machine Learning”, 24 hours, Department of Electronics and Informatics, Politecnico di Milano (Italy).
- *P. Chainais*, Ecole Centrale de Lille, “Machine Learning”, 36 hours, 3rd year.
- *P. Chainais*, Ecole Centrale de Lille, “Wavelets and Applications", 24 hours, 2nd year.
- *P. Chainais*, Ecole Centrale de Lille, “Introduction to Matlab", 16 hours, 3rd year.
- *P. Chainais*, Ecole Centrale de Lille, “Signal processing", 22 hours, 1st year.
- *P. Chainais*, Ecole Centrale de Lille, “Data Compression", 16 hours, 2nd year.
- *P. Chainais* is Responsible for a new 3rd year program called Decision making & Data analysis.
- *Ph. Preux*, “Decision under uncertainty”, 46 hours, M2, Master in Computer Science, Université de Lille 1.
- *R. Munos*, Master: “Introduction to Reinforcement Learning”, 30 hours, M2, Master “Mathematiques, Vision, Apprentissage”, ENS Cachan.
- *R. Gaudel*, Master: “Data Mining”, 24 hours, M2, Master “Mathematiques et Informatique Appliqués aux Sciences Humaines et Sociales”, Université Lille 3.
- *R. Gaudel*, Master: “Web Mining”, 24 hours, M2, Master “Mathematiques et Informatique Appliqués aux Sciences Humaines et Sociales”, Université Lille 3.
- *R. Gaudel*, Licence: “Programmation”, 2×16 hours, L1, Licence “Mathematiques et Informatique Appliqués aux Sciences Humaines et Sociales”, Université Lille 3.
- *R. Gaudel*, Licence: “Information and Communication Technologies”, 2×12 hours, L1, Licence “Sociologie, Histoire, Développement Social”, Université Lille 3.
- *R. Gaudel*, Licence: “Artificial Intelligence”, 27 hours, L2, Licence “Mathematiques et Informatique Appliqués aux Sciences Humaines et Sociales”, Université Lille 3.
- *R. Gaudel*, Licence: “C2i”, 25 hours, L1-3, any Licence, Université Lille 3.
- *J. Mary*, Master : “Programmation et analyse de donnée en R”, 48h eq TD, M1, Université de Lille 3, France.
- *J. Mary*, Master : “Graphes et Réseaux”, 32h eq TD,L1, Université de Lille 3, France.
- *J. Mary*, Master : “Système”, 12h eq TD,L1, Université de Lille 3, France.
- *E. Duflos*, Master (3rd year of Engineer School): “Advanced Estimation" , 20 hours, M2, Option "Data Analysis and Decision", Ecole Centrale de Lille.

- *E. Duflos*, Master (3rd year of Engineer School): “Multi-Objects Filreting" , 16 hours, M2, Option "Data Analysis and Decision", Ecole Centrale de Lille.

9.3. Supervision

- PhD : *Jean Francois Hren*, Planification optimiste pour systèmes dèterministes, Université de Lille 1, June 2012.
- PhD : *Alexandra Carpentier*, Toward optimal sampling in low and high dimension, Université de Lille 1, Octobre 2012.
- PhD: *Christophe Salperwyck*, *Apprentissage incrémental en ligne sur flux de données*, Université de Lille 3, November 30, 1012, Philippe Preux, [4].
- PhD : *Emmanuel Delande*, “Multi-sensor PHD filtering with application to sensor management”, Jan. 2012, encadrement : E. Duflos and P. Vanheeghe.
- PhD in progress : *Boris Baldassari*, *Apprentissage automatique et développement logiciel*, Sep. 2011, encadrement: Philippe Preux.
- PhD in progress : *Victor Gabillon*, “Active Learning in Classification-based Policy Iteration”, Sep. 2009, encadrement: M. Ghavamzadeh, Ph. Preux.
- PhD in progress : *Azadeh Khaleghi*, “Unsupervised Learning of Sequential Data”, Sep. 2010, encadrement: D. Ryabko, Ph. Preux.
- PhD in progress : *Sami Naamane*, “Filtrage collaboratif adverse et dynamique”, Nov. 2011, encadrement: J. Mary, Ph. Preux.
- PhD in progress : *Olivier Nicol*, “Apprentissage par renforcement sous contrainte de ressources finies, dans un environnement non stationnaire, face à des flux de données massifs”, Nov. 2010, encadrement: J. Mary, Ph. Preux.
- PhD in progress : *Amir Sani*, “Learning under uncertainty”, Oct. 2011, encadrement: R. Munos, A. Lazaric.
- PhD in progress : *Emilie Kaufmann*, “Bayesian Bandits”, Oct. 2011, encadrement: R. Munos, O. Cappé, A. Garivier.
- PhD in progress : *Marta Soare*, “Pure Exploration in Multi-arm Bandit”, Oct. 2012, encadrement: R. Munos, A. Lazaric.
- PhD in progress : *Adrien Hoarau*, “Multi-arm Bandit Theory”, Oct. 2012, encadrement: R. Munos.

9.4. Juries

- Ph. Preux is an examiner of the H.D.R. of Ludovic Denoyer, Paris 6.
- *E. Duflos* is an examiner of the Ph.D. of GU Wei (IRCICA).

9.5. Popularization

- *J. Mary* received a bachelor student for one week to present some research oriented activities in informatics.
- *J. Mary* was involved in different PICOM meeting with private companies to present research on sequential data analysis.

10. Bibliography

Publications of the year

Doctoral Dissertations and Habilitation Theses

[1] A. CARPENTIER. *Toward optimal sampling in low and high dimension*, Université Lille 1, Lille, France, Octobre 2012.

[2] E. DELANDE. *Multi-sensor PHD filtering with application to sensor management*, Ecole Centrale, Lille, France, Octobre 2012, http://www.theses.fr/2012ECLI0001.

[3] J. F. HREN. *Planification optimiste pour systèmes dèterministes*, Université Lille 1, Lille, France, Juin 2012.

[4] C. SALPERWYCK. *Apprentissage incrémental en ligne sur flux de données*, Université de Lille 3, Nov 2012.

Articles in International Peer-Reviewed Journals

[5] M. G. AZAR, R. MUNOS, H. KAPPEN. *Minimax PAC-Bounds on the Sample Complexity of Reinforcement Learning with a Generative Model*, in "Machine Learning Journal", 2012, To appear.

[6] O. CAPPÉ, A. GARIVIER, O.-A. MAILLARD, R. MUNOS, G. STOLTZ. *Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation*, in "Annals of Statistics", 2012, Submitted to.

[7] A. CARPENTIER, A. LAZARIC, M. GHAVAMZADEH, R. MUNOS, P. AUER. *Upper-Confidence-Bound Algorithms for Active Learning in Multi-Armed Bandits*, in "Theoretical Computer Science", 2012, To appear.

[8] A. CARPENTIER, R. MUNOS, A. ANTOS. *Minimax strategy for Stratified Sampling for Monte Carlo*, in "Journal of Machine Learning Research", 2012, Submitted to.

[9] G. DULAC-ARNOLD, L. DENOYER, P. PREUX, P. GALLINARI. *Sequential approaches for learning datum-wise sparse representations*, in "Machine Learning", October 2012, vol. 89, n^{o} 1-2, p. 87-122.

[10] J. FRUITET, A. CARPENTIER, R. MUNOS, M. CLERC. *Automatic motor task selection via a bandit algorithm for a brain-controlled button*, in "Journal of Neural Engineering", 2012, To appear.

[11] S. GIRGIN, J. MARY, P. PREUX, O. NICOL. *Managing advertising campaigns – an approximate planning approach*, in "Frontiers of Computer Science", 2012, vol. 6, n^{o} 2, p. 209-229 [*DOI :* 10.1007/s11704-012-2873-5], http://hal.inria.fr/hal-00747722.

[12] M. HAUSKRECHT, I. BATAL, M. VALKO, S. VISWESWARAN, G. F. COOPER, G. CLERMONT. *Outlier detection for patient monitoring and alerting.*, in "Journal of Biomedical Informatics", August 2012 [*DOI :* 10.1016/J.JBI.2012.08.004], http://hal.inria.fr/hal-00742097.

[13] A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Analysis of Classification-based Policy Iteration Algorithms*, in "Journal of Machine learning Research", 2012, Submitted to.

[14] A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Finite-Sample Analysis of Least-Squares Policy Iteration*, in "Journal of Machine Learning Research", 2012, vol. 13, p. 3041-3074.

[15] A. LAZARIC, R. MUNOS. *Learning with stochastic inputs and adversarial outputs*, in "Journal of Computer and System Sciences (JCSS)", 2012, vol. 78, n^{o} 5, p. 1516–1537 [*DOI :* 10.1016/J.JCSS.2011.12.027], http://www.sciencedirect.com/science/article/pii/S002200001200027X.

[16] O.-A. MAILLARD, R. MUNOS. *Linear Regression with Random Projections*, in "Journal of Machine learning Research", 2012, vol. 13, p. 2735-2772.

[17] R. MUNOS. *The Optimistic Principle applied to Games, Optimization and Planning: Towards Foundations of Monte-Carlo Tree Search*, in "Foundations and Trends in Machine Learning", 2012, Submitted to, http://hal.archives-ouvertes.fr/hal-00747575.

[18] O. NICOL, J. MARY, P. PREUX. *ICML Exploration & Exploitation challenge: Keep it simple!*, in "Journal of Machine Learning research Workshop and Conference Proceedings", 2012, vol. 26, p. 62-85, http://hal.inria.fr/hal-00747725.

[19] A. RABAOUI, N. VIANDIER, J. MARAIS, E. DUFLOS, P. VANHEEGHE. *Dirichlet Process Mixtures for Density Estimation in Dynamic Nonlinear Modeling: Application to GPS Positioning in Urban Canyons*, in "IEEE Transactions on Signal Processing", April 2012, vol. 60, n° 4, p. 1638 - 1655 [*DOI :* 10.1109/TSP.2011.2180901], http://hal.inria.fr/hal-00712718.

[20] S. RAZAVI, E. DUFLOS, C. HAAS, P. VANHEEGHE. *Dislocation detection in field environments: A belief functions contribution*, in "Expert Systems with Applications", August 2012, vol. 39, n° 10, p. 8505-8513 [*DOI :* 10.1016/J.ESWA.2011.12.014], http://hal.inria.fr/hal-00712720.

[21] D. RYABKO. *Testing composite hypotheses about discrete ergodic processes*, in "Test", 2012, vol. 21, n° 2, p. 317-329.

[22] D. RYABKO. *Uniform hypothesis testing for finite-valued stationary processes*, in "Statistics", 2013.

[23] M. VALKO, M. GHAVAMZADEH, A. LAZARIC. *Semi-Supervised Apprenticeship Learning*, in "Journal of Machine Learning Research: Workshop and Conference Proceedings", November 2012, vol. 24, http://hal.inria.fr/hal-00747921.

International Conferences with Proceedings

[24] M. G. AZAR, R. MUNOS, H. KAPPEN. *On the Sample Complexity of Reinforcement Learning with a Generative Model*, in "International Conference on Machine Learning", 2012.

[25] L. BUSONIU, R. MUNOS. *Optimistic planning in Markov decision processes*, in "International conference on Artificial Intelligence and Statistics", 2012.

[26] A. CARPENTIER, R. MUNOS. *Adaptive Stratified Sampling for Monte-Carlo integration of Differentiable functions*, in "Advances in Neural Information Processing Systems", 2012.

[27] A. CARPENTIER, R. MUNOS. *Bandit Theory meets Compressed Senseing for high dimensional Stochastic Linear Bandit*, in "International conference on Artificial Intelligence and Statistics", 2012.

[28] A. CARPENTIER, R. MUNOS. *Minimax number of strata for online Stratified Sampling given Noisy Samples*, in "International Conference on Algorithmic Learning Theory", 2012.

[29] P. CHAINAIS. *Towards dictionary learning from images with non Gaussian noise*, in "IEEE Int. Workshop on Machine Learning for Signal Processing", Santander, Spain, September 2012, 0000, http://hal.inria.fr/hal-00749035.

[30] R. COULOM. *CLOP: Confident Local Optimization for Noisy Black-Box Parameter Tuning*, in "Advances in Computer Games - 13th International Conference", Tilburg, Pays-Bas, H. J. VAN DEN HERIK, A. PLAAT

(editors), Lecture Notes in Computer Science, Springer, 2012, vol. 7168, p. 146-157 [*DOI :* 10.1007/978-3-642-31866-5_13], http://hal.inria.fr/hal-00750326.

[31] G. DULAC-ARNOLD, L. DENOYER, P. PREUX, P. GALLINARI. *Fast Reinforcement Learning with Large Action Sets Using Error-Correcting Output Codes for MDP Factorization*, in "European Conference on Machine Learning", Bristol, United Kingdom, Springer, 2012, vol. 2, p. 180-194 [*DOI :* 10.1007/978-3-642-33486-3_12], http://hal.inria.fr/hal-00747729.

[32] J. FRUITET, A. CARPENTIER, R. MUNOS, M. CLERC. *Bandit Algorithms boost motor-task selection for Brain Computer Interfaces*, in "Advances in Neural Information Processing Systems", 2012.

[33] V. GABILLON, M. GHAVAMZADEH, A. LAZARIC. *Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence*, in "Proceedings of Advances in Neural Information Processing Systems 25", MIT Press, 2012.

[34] N. GATTI, A. LAZARIC, F. TROVÒ. *A Truthful Learning Mechanism for Multi-Slot Sponsored Search Auctions with Externalities (Extended Abstract)*, in "AAMAS", 2012.

[35] N. GATTI, A. LAZARIC, F. TROVÒ. *A Truthful Learning Mechanism for Multi-Slot Sponsored Search Auctions with Externalities*, in "Proceedings of the 13th ACM Conference on Electronic Commerce (EC'12)", 2012.

[36] M. GEIST, B. SCHERRER, A. LAZARIC, M. GHAVAMZADEH. *A Dantzig Selector Approach to Temporal Difference Learning*, in "Proceedings of the Twenty-Ninth International Conference on Machine Learning", 2012, p. 1399-1406.

[37] M. GHAVAMZADEH, A. LAZARIC. *Conservative and Greedy Approaches to Classification-based Policy Iteration*, in "Proceedings of the Twenty-Sixth Conference on Artificial Intelligence", 2012, p. 914-920.

[38] E. KAUFFMANN, N. KORDA, R. MUNOS. *Thompson Sampling: an Asymptotically Optimal Finite Time Analysis*, in "International Conference on Algorithmic Learning Theory", 2012.

[39] A. KHALEGHI, D. RYABKO. *Locating Changes in Highly Dependent Data with Unknown Number of Change Points*, in "NIPS", Lake Tahoe, USA, 2012.

[40] A. KHALEGHI, D. RYABKO, J. MARY, P. PREUX. *Online Clustering of Processes*, in "AISTATS", JMLR W&CP 22, 2012, p. 601-609.

[41] B. KVETON, M. VALKO. *Learning from a Single Labeled Face and a Stream of Unlabeled Data*, in "10th IEEE International Conference on Automatic Face and Gesture Recognition", Shanghai, China, November 2012, http://hal.inria.fr/hal-00749197.

[42] O.-A. MAILLARD, A. CARPENTIER. *Online allocation and homogeneous partitioning for piecewise constant mean approximation*, in "Advances in Neural Information Processing Systems", 2012.

[43] R. ORTNER, D. RYABKO, P. AUER, R. MUNOS. *Regret Bounds for Restless Markov Bandits*, in "Proc. 23th International Conf. on Algorithmic Learning Theory (ALT'12)", Lyon, France, LNCS 7568, Springer, Berlin, 2012, p. 214–228.

[44] R. ORTNER, D. RYABKO. *Online Regret Bounds for Undiscounted Continuous Reinforcement Learning*, in "NIPS", Lake Tahoe, USA, 2012.

[45] D. RYABKO, J. MARY. *Reducing statistical time-series problems to binary classification*, in "NIPS", Lake Tahoe, USA, 2012.

[46] A. SANI, A. LAZARIC, R. MUNOS. *Risk-Aversion in Multi-Armed Bandits*, in "Advances in Neural Information Processing Systems", 2012.

[47] B. SCHERRER, M. GHAVAMZADEH, V. GABILLON, M. GEIST. *Approximate Modified Policy Iteration*, in "Proceedings of the Twenty-Ninth International Conference on Machine Learning", 2012, p. 1207-1214.

National Conferences with Proceeding

[48] G. DULAC-ARNOLD, L. DENOYER, P. PREUX, P. GALLINARI. *Apprentissage par renforcement rapide pour des grands ensembles d'actions en utilisant des codes correcteurs d'erreur*, in "Journées Francophones sur la planification, la décision et l'apprentissage pour le contrôle des systèmes - JFPDA 2012", Villers-lès-Nancy, France, O. BUFFET (editor), 2012, 12 p, http://hal.inria.fr/hal-00736322.

[49] M. GEIST, B. SCHERRER, A. LAZARIC, M. GHAVAMZADEH. *Un sélecteur de Dantzig pour l'apprentissage par différences temporelles*, in "Journées Francophones sur la planification, la décision et l'apprentissage pour le contrôle des systèmes - JFPDA 2012", Villers-lès-Nancy, France, O. BUFFET (editor), 2012, 13 p, http://hal.inria.fr/hal-00736229.

[50] N. JAOUA, E. DUFLOS, P. VANHEEGHE. *DPM pour l'inférence dans les modèles dynamiques non linéaires avec des bruits de mesure alpha-stable*, in "44ème Journées de Statistique", Bruxelles, Belgium, May 2012, p. 1-4, http://hal.inria.fr/hal-00713857.

[51] B. SCHERRER, V. GABILLON, M. GHAVAMZADEH, M. GEIST. *Approximations de l'Algorithme Itérations sur les Politiques Modifié*, in "Journées Francophones sur la planification, la décision et l'apprentissage pour le contrôle des systèmes - JFPDA 2012", Villers-lès-Nancy, France, O. BUFFET (editor), 2012, 1 p, Le corps de cet article est paru, en langue anglaise, dans ICML'2012 (Proceedings of the International Conference on Machine Learning), http://hal.inria.fr/hal-00736226.

Conferences without Proceedings

[52] C. DHANJAL, R. GAUDEL, S. CLÉMENÇON. *Incremental Spectral Clustering with the Normalised Laplacian*, in "DISCML - 3rd NIPS Workshop on Discrete Optimization in Machine Learning - 2011", Sierra Nevada, Espagne, 2012, http://hal.inria.fr/hal-00745666.

[53] A. FARAHMAND, D. PRECUP, M. GHAVAMZADEH. *On Classification-based Approximate Policy Iteration*, in "Thirtieth International Conference on Machine Learning", 2012, submitted.

[54] D. RYABKO. *Asymptotic statistics of stationary ergodic time series*, in "WITMSE", Amsterdam, 2012.

Scientific Books (or Scientific Book chapters)

[55] L. BUSONIU, A. LAZARIC, M. GHAVAMZADEH, R. MUNOS, R. BABUSKA, B. DE SCHUTTER. *Least-Squares Methods for Policy Iteration*, in "Reinforcement Learning: State of the Art", M. WIERING, M. VAN OTTERLO (editors), Springer Verlag, 2012, p. 75-110.

[56] A. LAZARIC. *Transfer in Reinforcement Learning: a Framework and a Survey*, in "Reinforcement Learning: State of the Art", M. WIERING, M. VAN OTTERLO (editors), Springer, 2012.

[57] N. VLASSIS, M. GHAVAMZADEH, S. MANNOR, P. POUPART. *Bayesian Reinforcement Learning*, in "Reinforcement Learning: State of the Art", M. WIERING, M. VAN OTTERLO (editors), Springer Verlag, 2012, p. 359-386.

Research Reports

[58] V. GABILLON, M. GHAVAMZADEH, A. LAZARIC. *Best Arm Identification: A Unified Approach to Fixed Budget and Fixed Confidence*, Inria, 2012, n° inria-00747005.

[59] H. KADRI, M. GHAVAMZADEH, P. PREUX. *A Generalized Kernel Approach to Structured Output Learning*, Inria, May 2012, n° RR-7956, http://hal.inria.fr/hal-00695631.

[60] H. KADRI, A. RAKOTOMAMONJY, F. BACH, P. PREUX. *Multiple Operator-valued Kernel Learning*, Inria, March 2012, n° RR-7900, http://hal.inria.fr/hal-00677012.

[61] B. PIRES, M. GHAVAMZADEH, CS. SZEPESVÁARI. *Risk Bounds in Cost-sensitive Multiclass Classification: an Application to Reinforcement Learning*, Inria, 2012, in preparation.

[62] B. SCHERRER, V. GABILLON, M. GHAVAMZADEH, M. GEIST. *Approximate Modified Policy Iteration*, Inria, May 2012, http://hal.inria.fr/hal-00697169.

References in notes

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

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

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

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

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

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

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

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

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

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

[73] J. RUST. *On the Optimal Lifetime of Nuclear Power Plants*, in "Journal of Business & Economic Statistics", 1997, vol. 15, n^{o} 2, p. 195–208.

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

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

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