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

Université des sciences et technologies de Lille (Lille 1)

Activity Report 2018

Project-Team SEQUEL

Sequential Learning

IN COLLABORATION WITH: Centre de Recherche en Informatique, Signal et Automatique de Lille

RESEARCH CENTER
Lille - Nord Europe

THEME
Optimization, machine learning and statistical methods

Table of contents

Project-Team SEQUEL

Creation of the Project-Team: 2007 July 01

Keywords:

Computer Science and Digital Science:

A3. - Data and knowledge
A3.1. - Data
A3.1.1. - Modeling, representation
A3.1.4. - Uncertain data
A3.3. - Data and knowledge analysis
A3.3.1. - On-line analytical processing
A3.3.2. - Data mining
A3.3.3. - Big data analysis
A3.4. - Machine learning and statistics
A3.4.1. - Supervised learning
A3.4.2. - Unsupervised learning
A3.4.3. - Reinforcement learning
A3.4.4. - Optimization and learning
A3.4.6. - Neural networks
A3.4.8. - Deep learning
A3.5.2. - Recommendation systems
A5.1. - Human-Computer Interaction
A9. - Artificial intelligence
A9.2. - Machine learning
A9.3. - Signal analysis
A9.4. - Natural language processing
A9.7. - AI algorithmics

Other Research Topics and Application Domains:

B5.8. - Learning and training
B6.1. - Software industry
B7.2.1. - Smart vehicles
B9.1.1. - E-learning, MOOC
B9.5. - Sciences
B9.5.6. - Data science

1. Team, Visitors, External Collaborators

Research Scientists

Émilie Kaufmann [CNRS, Researcher]
Odalric Maillard [Inria, Researcher]
Michal Valko [Inria, Researcher, HDR]

Faculty Member

Philippe Preux [Team leader, Univ Charles de Gaulle, Professor, HDR]

Post-Doctoral Fellows

Matteo Pirotta [Inria, until Oct 2018]
Mohammad Sadegh Talebi Mazraeh Shahi [Inria, from Jun 2018]

PhD Students

Sheikh Waqas Akhtar [Inria, until Aug 2018]
Merwan Barlier [Univ de Lille, from Feb 2018 until Nov 2018]
Lilian Besson [Ecole Normale Supérieure Cachan]
Nicolas Carrara [Orange Labs puis Univ de Lille, since 2015]
Omar Darwiche Domingues [Inria, from Oct 2018]
Yannis Flet Berliac [Univ des sciences et technologies de Lille, from Oct 2018]
Ronan Fruit [Inria, since Dec 2015]
Jean-Bastien Grill [Ecole Normale Supérieure Paris, until Dec 2018]
Édouard Leurent [Renault]
Pierre Perrault [Inria]
Hassan Saber [Inria, from Oct 2018]
Mathieu Seurin [Univ des sciences et technologies de Lille]
Julien Seznec [Le Livre Scolaire]
Xuedong Shang [Univ des sciences et technologies de Lille]
Florian Strub [Univ des sciences et technologies de Lille]
Kiewan Villatel [Critéo]
Romain Warlop [55, until Oct 2018]

Technical staff

Guillaume Gautier [CNRS&Inria, since 2017]

Interns

Quentin Burthier [Inria, from Jun 2018 until Aug 2018]
Edouard Dendauw [Univ de Lille, from May 2018 until Jul 2018]
Thibault Felicite [Inria, Jul 2018]
Robert Lindland [Inria, from May 2018 until Aug 2018]
Jian Qian [Inria, from May 2018 until Oct 2018]
Hassan Saber [Inria, from Apr 2018 until Aug 2018]
Benoit Schmitt [Inria, from Mar 2018 until Aug 2018]
Han Shao [Chinese University of Hong-Kong, from Oct 2018 until Nov 2018]
Annie Yun [Inria, from May 2018 until Aug 2018]
Arnaud Fanthomme [Ecole Normale Supérieure Paris, from Apr 2018 until Aug 2018]

Administrative Assistant

Amelie Supervielle [Inria]

Visiting Scientist

Jumpei Komiyama [Univ Tokyo, from Sep 2018 until Nov 2018]

External Collaborators

Remi Bardenet [CNRS]
Jérémie Mary [Critéo, HDR]
Olivier Pietquin [Google, HDR]

2. Overall Objectives

2.1. Presentation

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 [62]). 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. Research Program

3.1. In Short

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 [60].

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 [59]) 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) [57], 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 [58]. 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 ([63]):

- 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 challenges 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 [61], 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.* [56] 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 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 behavior (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 behavior, or than a class of other behaviors.

The problem of hypothesis testing can also be studied in its general formulations: given two (abstract) hypothesis H_0 and H_1 about the unknown measure that generates the data, 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) behavior 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 $x^1 = (x_1^1, \cdots, x_{n_1}^1), \cdots, x^N = (x_N^1, \cdots, x_{n_N}^N)$, 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.

3.3.6. Online Kernel and Graph-Based Methods

Large-scale kernel ridge regression is limited by the need to store a large kernel matrix. Similarly, large-scale graph-based learning is limited by storing the graph Laplacian. Furthermore, if the data come online, at some point no finite storage is sufficient and per step operations become slow.

Our challenge is to design sparsification methods that give guaranteed approximate solutions with a reduced storage requirements.

4. Application Domains

4.1. Sequential decision making under uncertainty and prediction

The spectrum of applications of our research is very wide: it ranges from the core of our research, that is sequential decision making under uncertainty, to the application of components used to solve this decision making problem.

To be more specific, we work on computational advertising and recommendation systems; these problems are considered as a sequential matching problem in which resources available in a limited amount have to be matched to meet some users' expectations. The sequential approach we advocate paves the way to better tackle the cold-start problem, and non stationary environments. More generally, these approaches are applied to the optimization of budgeted resources under uncertainty, in a time-varying environment, including constraints on computational times (typically, a decision has to be made in less than 1 ms in a recommendation system). An other field of applications of our research is related to education which we consider as a sequential matching problem between a student, and educational contents.

The algorithms to solve these tasks heavily rely on tools from machine learning, statistics, and optimization. Henceforth, we also apply our work to more classical supervised learning, and prediction tasks, as well as unsupervised learning tasks. The whole range of methods is used, from decision forests, to kernel methods, to deep learning. For instance, we have recently used deep learning on images. We also have a line of works related to software development studying how machine learning can improve the quality of software being developed. More generally, we apply our research to data science.

5. Highlights of the Year

5.1. Highlights of the Year

- Daniele Calandriello is awarded with the AFIA price for his PhD defended in December 2017. As a side note, this is the 5th time a PhD student of SEQUEL receives this award since our first PhD defense in 2010.

- We organized the 14th European Workshop on Reinforcement Learning in Lille. This event gathered 200 researchers; there were a dozen invited presentations by world research leaders, including Prof. Richard Sutton (U. Alberta), the founder of modern RL, Prof. Tze Leung Lai (Stanford U.), one of the key reference in bandit research, and also Nicolò Cesa-Bianchi (U. Milan), Peter Auer (U. of Leoben), Rémi Munos (Deepmind Paris), and Joelle Pineau (Mc Gill and FAIR).

5.1.1. Awards

- Former 2017 intern M. Asadi got a It was "Best Poster Award" at Transylvania Machine Learning Summer School (TMLSS), July 2018 for the work she did while in SEQUEL
- É. Kaufmann is among the top 10 reviewers at ICML 2018 (out of 1800 reviewers)
- Ph. Preux was among the 24 "level-2 Distinguished Senior Program Committee Members" for IJCAI 2018 (out of 498 SPC members, 115 were distinguished, 23 at level 2, the highest level)
- M. Valko is among the top 10 reviewers at ICML 2018 (out of 1800 reviewers)

6. New Software and Platforms

6.1. BAC

Bayesian Policy Gradient and Actor-Critic Algorithms
KEYWORDS: Machine learning - Incremental learning - Policy Learning
FUNCTIONAL DESCRIPTION: To address this issue, we proceed to supplement our Bayesian policy gradient framework with a new actor-critic learning model in which a Bayesian class of non-parametric critics, based on Gaussian process temporal difference learning, is used. Such critics model the action-value function as a Gaussian process, allowing Bayes' rule to be used in computing the posterior distribution over action-value functions, conditioned on the observed data. Appropriate choices of the policy parameterization and of the prior covariance (kernel) between action-values allow us to obtain closed-form expressions for the posterior distribution of the gradient of the expected return with respect to the policy parameters. We perform detailed experimental comparisons of the proposed Bayesian policy gradient and actor-critic algorithms with classic Monte-Carlo based policy gradient methods, as well as with each other, on a number of reinforcement learning problems.

- Contact: Michal Valko
- URL: https://team.inria.fr/sequel/Software/BAC/

6.2. GuessWhat?!

GuessWhat?! Visual object discovery through multi-modal dialogue
KEYWORDS: Deep learning - Dialogue System
FUNCTIONAL DESCRIPTION: This project train a AI to play the GuessWhat?! game. Thus, you can train an AI to ask questions, to answer questions about images. You can also perform basic visual reasoning. This project is a testbed for future interactive dialogue system.

- Partner: Universite de Montreal
- Contact: Florian Strub
- Publications: GuessWhat?! Visual object discovery through multi-modal dialogue - End-to-end optimization of goal-driven and visually grounded dialogue systems Harm de Vries

6.3. Squeak

Sequential sampling for kernel matrix approximation

KEYWORD: Machine learning

- Contact: Daniele Calandriello
- URL: http://researchers.lille.inria.fr/~valko/hp/serve.php?what=publications/squeak.py

6.4. OOR

Optimistic Optimization in R
KEYWORDS: Black-box optimization - Machine learning

- Contact: Mickael Binois
- URL: https://cran.r-project.org/web/packages/OOR/index.html

6.5. DPPy

Sampling Determinantal Point Processes with Python
KEYWORD: Determinantal point processes
FUNCTIONAL DESCRIPTION: Determinantal point processes (DPPs) are specific probability distributions over clouds of points that are used as models and computational tools across physics, probability, statistics, and more recently machine learning. Sampling from DPPs is nontrivial and therefore we present. DPPy, a Python toolbox that gathers known exact and approximate sampling algorithms. The project is hosted on GitHub and equipped with an extensive documentation.

- Contact: Guillaume Gautier
- URL: https://github.com/guilgautier/DPPy/

6.6. SMPyBandits

Open-Source Python package for Single- and Multi-Players multi-armed Bandits algorithms.
KEYWORD: Machine learning
FUNCTIONAL DESCRIPTION: The library contains the implementation of many single-player mutli-armed bandit algorithms as well as the implementation of all the state-of-the-art multi-player algorithms.

- Contact: Lilian Besson

7. New Results

7.1. Decision-making Under Uncertainty

7.1.1. Reinforcement Learning

A Fitted-Q Algorithm for Budgeted MDPs, [26]

We address the problem of budgeted reinforcement learning, in continuous state-space, using a batch of transitions. To this extend, we introduce a novel algorithm called Budgeted Fitted-Q (BFTQ). Benchmarks show that BFTQ performs as well as a regular Fitted-Q algorithm in a continuous 2-D world but also allows one to choose the right amount of budget that fits to a given task without the need of engineering the rewards. We believe that the general principles used to design BFTQ can be applied to extend others classical reinforcement learning algorithms for budgeted oriented applications.

Safe transfer learning for dialogue applications, [27]

In this paper, we formulate the hypothesis that the first dialogues with a new user should be handle in a very conservative way, for two reasons : avoid user dropout; gather more successful dialogues to speedup the learning of the asymptotic strategy. To this extend, we propose to transfer a safe strategy to initiate the first dialogues.

Variance-Aware Regret Bounds for Undiscounted Reinforcement Learning in MDPs, [17]

The problem of reinforcement learning in an unknown and discrete Markov Decision Process (MDP) under the average-reward criterion is considered, when the learner interacts with the system in a single stream of observations, starting from an initial state without any reset. We revisit the minimax lower bound for that problem by making appear the local variance of the bias function in place of the diameter of the MDP. Furthermore, we provide a novel analysis of the KL-Ucrl algorithm establishing a high-probability regret bound scaling as O S s,a V s,a T for this algorithm for ergodic MDPs, where S denotes the number of states and where V s,a is the variance of the bias function with respect to the next-state distribution following action a in state s. The resulting bound improves upon the best previously known regret bound O(DS $\sqrt{}$ AT) for that algorithm, where A and D respectively denote the maximum number of actions (per state) and the diameter of MDP. We finally compare the leading terms of the two bounds in some benchmark MDPs indicating that the derived bound can provide an order of magnitude improvement in some cases. Our analysis leverages novel variations of the transportation lemma combined with Kullback-Leibler concentration inequalities, that we believe to be of independent interest.

Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning, [29]

We introduce SCAL, an algorithm designed to perform efficient exploration-exploitation in any unknown weakly-communicating Markov decision process (MDP) for which an upper bound c on the span of the optimal bias function is known. For an MDP with S states, A actions and $\Gamma \leq S$ possible next states, we prove a regret bound of $\widetilde{O}(c\sqrt{\Gamma SAT})$, which significantly improves over existing algorithms (e.g., UCRL and PSRL), whose regret scales linearly with the MDP diameter D. In fact, the optimal bias span is finite and often much smaller than D (e.g., $D = \infty$ in non-communicating MDPs). A similar result was originally derived by Bartlett and Tewari (2009) for REGAL.C, for which no tractable algorithm is available. In this paper, we relax the optimization problem at the core of REGAL.C, we carefully analyze its properties, and we provide the first computationally efficient algorithm to solve it. Finally, we report numerical simulations supporting our theoretical findings and showing how SCAL significantly outperforms UCRL in MDPs with large diameter and small span.

Near Optimal Exploration-Exploitation in Non-Communicating Markov Decision Processes, [28]

While designing the state space of an MDP, it is common to include states that are transient or not reachable by any policy (e.g., in mountain car, the product space of speed and position contains configurations that are not physically reachable). This leads to defining weakly-communicating or multi-chain MDPs. In this paper, we introduce TUCRL, the first algorithm able to perform efficient exploration-exploitation in any finite Markov Decision Process (MDP) without requiring any form of prior knowledge. In particular, for any MDP with S^{C} communicating states, A actions and $\Gamma^{\mathrm{C}} \leq S^{\mathrm{C}}$ possible communicating next states, we derive a $\widetilde{O}(D^{\mathrm{C}}\sqrt{\Gamma^{\mathrm{C}} S^{\mathrm{C}} AT})$ regret bound, where D^{C} is the diameter (i.e., the longest shortest path) of the communicating part of the MDP. This is in contrast with optimistic algorithms (e.g., UCRL, Optimistic PSRL) that suffer linear regret in weakly-communicating MDPs, as well as posterior sampling or regularized algorithms (e.g., REGAL), which require prior knowledge on the bias span of the optimal policy to bias the exploration to achieve sub-linear regret. We also prove that in weakly-communicating MDPs, no algorithm can ever achieve a logarithmic growth of the regret without first suffering a linear regret for a number of steps that is exponential in the parameters of the MDP. Finally, we report numerical simulations supporting our theoretical findings and showing how TUCRL overcomes the limitations of the state-of-the-art.

Upper Confidence Reinforcement Learning exploiting state-action equivalence, [53]

Stochastic Variance-Reduced Policy Gradient, [34]

In this paper, we propose a novel reinforcement-learning algorithm consisting in a stochastic variance-reduced version of policy gradient for solving Markov Decision Processes (MDPs). Stochastic variance-reduced gradient (SVRG) methods have proven to be very successful in supervised learning. However, their adaptation to policy gradient is not straightforward and needs to account for I) a non-concave objective function; II) approximations in the full gradient computation; and III) a non-stationary sampling process. The result is SVRPG, a stochastic variance-reduced policy gradient algorithm that leverages on importance weights to

preserve the unbiasedness of the gradient estimate. Under standard assumptions on the MDP, we provide convergence guarantees for SVRPG with a convergence rate that is linear under increasing batch sizes. Finally, we suggest practical variants of SVRPG, and we empirically evaluate them on continuous MDPs.

Importance Weighted Transfer of Samples in Reinforcement Learning, [38]

We consider the transfer of experience samples (i.e., tuples < s, a, s', r >) in reinforcement learning (RL), collected from a set of source tasks to improve the learning process in a given target task. Most of the related approaches focus on selecting the most relevant source samples for solving the target task, but then all the transferred samples are used without considering anymore the discrepancies between the task models. In this paper, we propose a model-based technique that automatically estimates the relevance (importance weight) of each source sample for solving the target task. In the proposed approach, all the samples are transferred and used by a batch RL algorithm to solve the target task, but their contribution to the learning process is proportional to their importance weight. By extending the results for importance weighting provided in supervised learning literature, we develop a finite-sample analysis of the proposed batch RL algorithm. Furthermore, we empirically compare the proposed algorithm to state-of-the-art approaches, showing that it achieves better learning performance and is very robust to negative transfer, even when some source tasks are significantly different from the target task.

Training Dialogue Systems With Human Advice, [20]

One major drawback of Reinforcement Learning (RL) Spoken Dialogue Systems is that they inherit from the general exploration requirements of RL which makes them hard to deploy from an industry perspective. On the other hand, industrial systems rely on human expertise and hand written rules so as to avoid irrelevant behavior to happen and maintain acceptable experience from the user point of view. In this paper, we attempt to bridge the gap between those two worlds by providing an easy way to incorporate all kinds of human expertise in the training phase of a Reinforcement Learning Dialogue System. Our approach, based on the TAMER framework, enables safe and efficient policy learning by combining the traditional Reinforcement Learning reward signal with an additional reward, encoding expert advice. Experimental results show that our method leads to substantial improvements over more traditional Reinforcement Learning methods.

7.1.1.1. Deep reinforcement learning

FiLM: Visual Reasoning with a General Conditioning Layer, [35]

We introduce a general-purpose conditioning method for neural networks called FiLM: Feature-wise Linear Modulation. FiLM layers influence neural network computation via a simple , feature-wise affine transformation based on conditioning information. We show that FiLM layers are highly effective for visual reasoning - answering image-related questions which require a multi-step, high-level process - a task which has proven difficult for standard deep learning methods that do not explicitly model reasoning. Specifically, we show on visual reasoning tasks that FiLM layers 1) halve state-of-the-art error for the CLEVR benchmark, 2) modulate features in a coherent manner, 3) are robust to ablations and architectural modifications, and 4) generalize well to challenging, new data from few examples or even zero-shot.

Feature-wise transformations, [13]

Deep Reinforcement Learning and the Deadly Triad, [55]

We know from reinforcement learning theory that temporal difference learning can fail in certain cases. Sutton and Barto (2018) identify a deadly triad of function approximation, bootstrapping, and off-policy learning. When these three properties are combined, learning can diverge with the value estimates becoming unbounded. However, several algorithms successfully combine these three properties, which indicates that there is at least a partial gap in our understanding. In this work, we investigate the impact of the deadly triad in practice, in the context of a family of popular deep reinforcement learning models - deep Q-networks trained with experience replay - analyzing how the components of this system play a role in the emergence of the deadly triad, and in the agent's performance

7.1.2. Multi-armed Bandit Theory

Corrupt Bandits for Preserving Local Privacy, [30]

We study a variant of the stochastic multi-armed bandit (MAB) problem in which the rewards are corrupted. In this framework, motivated by privacy preservation in online recommender systems, the goal is to maximize the sum of the (unobserved) rewards, based on the observation of transformation of these rewards through a stochastic corruption process with known parameters. We provide a lower bound on the expected regret of any bandit algorithm in this corrupted setting. We devise a frequentist algorithm, KLUCB-CF, and a Bayesian algorithm, TS-CF and give upper bounds on their regret. We also provide the appropriate corruption parameters to guarantee a desired level of local privacy and analyze how this impacts the regret. Finally, we present some experimental results that confirm our analysis.

A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption, [21]

We study the problem of optimizing a function under a budgeted number of evaluations. We only assume that the function is locally smooth around one of its global optima. The difficulty of optimization is measured in terms of 1) the amount of noise b of the function evaluation and 2) the local smoothness, d, of the function. A smaller d results in smaller optimization error. We come with a new, simple, and parameter-free approach. First, for all values of b and d, this approach recovers at least the state-of-the-art regret guarantees. Second, our approach additionally obtains these results while being agnostic to the values of both b and d. This leads to the first algorithm that naturally adapts to an unknown range of noise b and leads to significant improvements in a moderate and low-noise regime. Third, our approach also obtains a remarkable improvement over the state-of-the-art SOO algorithm when the noise is very low which includes the case of optimization under deterministic feedback ($b = 0$). There, under our minimal local smoothness assumption, this improvement is of exponential magnitude and holds for a class of functions that covers the vast majority of functions that practitioners optimize ($d = 0$). We show that our algorithmic improvement is also borne out in the numerical experiments, where we empirically show faster convergence on common benchmark functions.

Best of both worlds: Stochastic & adversarial best-arm identification, [18]

We study bandit best-arm identification with arbitrary and potentially adversarial rewards. A simple random uniform learner obtains the optimal rate of error in the adversarial scenario. However, this type of strategy is suboptimal when the rewards are sampled stochastically. Therefore, we ask: Can we design a learner that performs optimally in both the stochastic and adversarial problems while not being aware of the nature of the rewards? First, we show that designing such a learner is impossible in general. In particular, to be robust to adversarial rewards, we can only guarantee optimal rates of error on a subset of the stochastic problems. We give a lower bound that characterizes the optimal rate in stochastic problems if the strategy is constrained to be robust to adversarial rewards. Finally, we design a simple parameter-free algorithm and show that its probability of error matches (up to log factors) the lower bound in stochastic problems, and it is also robust to adversarial ones.

Optimistic optimization of a Brownian, [31]

We address the problem of optimizing a Brownian motion. We consider a (random) realization W of a Brownian motion with input space in $[0, 1]$. Given W, our goal is to return an ϵ-approximation of its maximum using the smallest possible number of function evaluations, the sample complexity of the algorithm. We provide an algorithm with sample complexity of order $\log 2\ (1/\epsilon)$. This improves over previous results of Al-Mharmah and Calvin (1996) and Calvin et al. (2017) which provided only polynomial rates. Our algorithm is adaptive-each query depends on previous values-and is an instance of the optimism-in-the-face-of-uncertainty principle.

Rotting bandits are no harder than stochastic ones, [37]

In bandits, arms' distributions are stationary. This is often violated in practice, where rewards change over time. In applications as recommendation systems, online advertising, and crowdsourcing, the changes may be triggered by the pulls, so that the arms' rewards change as a function of the number of pulls. In this paper, we consider the specific case of non-parametric rotting bandits, where the expected reward of an arm may decrease every time it is pulled. We introduce the filtering on expanding window average (FEWA) algorithm that at each round constructs moving averages of increasing windows to identify arms that are

more likely to return high rewards when pulled once more. We prove that, without any knowledge on the decreasing behavior of the arms, FEWA achieves similar anytime problem-dependent, $\widetilde{\mathcal{O}}(\log(KT))$, and problem-independent, $\widetilde{\mathcal{O}}(\sqrt{KT})$, regret bounds of near-optimal stochastic algorithms as UCB1 of Auer et al. (2002a). This result substantially improves the prior result of Levine et al. (2017) which needed knowledge of the horizon and decaying parameters to achieve problem-independent bound of only $\widetilde{\mathcal{O}}(K^{1/3}T^{2/3})$. Finally, we report simulations confirming the theoretical improvements of FEWA.

Adaptive black-box optimization got easier: HCT only needs local smoothness, [41]

Hierarchical bandits is an approach for global optimization of extremely irregular functions. This paper provides new elements regarding POO, an adaptive meta-algorithm that does not require the knowledge of local smoothness of the target function. We first highlight the fact that the subroutine algorithm used in POO should have a small regret under the assumption of local smoothness with respect to the chosen partitioning, which is unknown if it is satisfied by the standard subroutine HOO. In this work, we establish such regret guarantee for HCT, which is another hierarchical optimistic optimization algorithm that needs to know the smoothness. This confirms the validity of POO. We show that POO can be used with HCT as a subroutine with a regret upper bound that matches the one of best-known algorithms using the knowledge of smoothness up to a $\sqrt{}$ log n factor.

Boundary Crossing Probabilities for General Exponential Families, [16]

Multi-Player Bandits Revisited, [22]

Multi-player Multi-Armed Bandits (MAB) have been extensively studied in the literature, motivated by applications to Cognitive Radio systems. Driven by such applications as well, we motivate the introduction of several levels of feedback for multi-player MAB algorithms. Most existing work assume that sensing information is available to the algorithm. Under this assumption, we improve the state-of-the-art lower bound for the regret of any decentralized algorithms and introduce two algorithms, RandTopM and MCTopM, that are shown to empirically outperform existing algorithms. Moreover, we provide strong theoretical guarantees for these algorithms, including a notion of asymptotic optimality in terms of the number of selections of bad arms. We then introduce a promising heuristic, called Selfish, that can operate without sensing information, which is crucial for emerging applications to Internet of Things networks. We investigate the empirical performance of this algorithm and provide some first theoretical elements for the understanding of its behavior.

Pure Exploration in Infinitely-Armed Bandit Models with Fixed-Confidence, [19]

We consider the problem of near-optimal arm identification in the fixed confidence setting of the infinitely armed bandit problem when nothing is known about the arm reservoir distribution. We (1) introduce a PAC-like framework within which to derive and cast results; (2) derive a sample complexity lower bound for near-optimal arm identification; (3) propose an algorithm that identifies a nearly-optimal arm with high probability and derive an upper bound on its sample complexity which is within a log factor of our lower bound; and (4) discuss whether our $\log^2(1/\delta)$ dependence is inescapable for "two-phase" (select arms first, identify the best later) algorithms in the infinite setting. This work permits the application of bandit models to a broader class of problems where fewer assumptions hold.

Aggregation of Multi-Armed Bandits Learning Algorithms for Opportunistic Spectrum Access, [23]

Multi-armed bandit algorithms have been recently studied and evaluated for Cognitive Radio (CR), especially in the context of Opportunistic Spectrum Access (OSA). Several solutions have been explored based on various models, but it is hard to exactly predict which could be the best for real-world conditions at every instants. Hence, expert aggregation algorithms can be useful to select on the run the best algorithm for a specific situation. Aggregation algorithms, such as Exp4 dating back from 2002, have never been used for OSA learning, and we show that it appears empirically sub-efficient when applied to simple stochastic problems. In this article, we present an improved variant, called Aggregator. For synthetic OSA problems modeled as Multi-Armed Bandit (MAB) problems, simulation results are presented to demonstrate its empirical efficiency. We combine classical algorithms, such as Thompson sampling, Upper-Confidence Bounds algorithms (UCB and variants), and Bayesian or Kullback-Leibler UCB. Our algorithm offers good performance compared to

state-of-the-art algorithms (Exp4, CORRAL or LearnExp), and appears as a robust approach to select on the run the best algorithm for any stochastic MAB problem, being more realistic to real-world radio settings than any tuning-based approach.

What Doubling Tricks Can and Can't Do for Multi-Armed Bandits, [47]

An online reinforcement learning algorithm is anytime if it does not need to know in advance the horizon T of the experiment. A well-known technique to obtain an anytime algorithm from any non-anytime algorithm is the "Doubling Trick". In the context of adversarial or stochastic multi-armed bandits, the performance of an algorithm is measured by its regret, and we study two families of sequences of growing horizons (geometric and exponential) to generalize previously known results that certain doubling tricks can be used to conserve certain regret bounds. In a broad setting, we prove that a geometric doubling trick can be used to conserve (minimax) bounds in $R\_T = O(\sqrt{T})$ but cannot conserve (distribution-dependent) bounds in $R\_T = O(\log T)$. We give insights as to why exponential doubling tricks may be better, as they conserve bounds in $R\_T = O(\log T)$, and are close to conserving bounds in $R\_T = O(\sqrt{T})$.

Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals, [50]

This paper presents new deviation inequalities that are valid uniformly in time under adaptive sampling in a multi-armed bandit model. The deviations are measured using the Kullback-Leibler divergence in a given one-dimensional exponential family, and may take into account several arms at a time. They are obtained by constructing for each arm a mixture martingale based on a hierarchical prior, and by multiplying those martingales. Our deviation inequalities allow us to analyze stopping rules based on generalized likelihood ratios for a large class of sequential identification problems, and to construct tight confidence intervals for some functions of the means of the arms.

7.1.3. Stochastic Games

Actor-Critic Fictitious Play in Simultaneous Move Multistage Games, [36]

Fictitious play is a game theoretic iterative procedure meant to learn an equilibrium in normal form games. However, this algorithm requires that each player has full knowledge of other players' strategies. Using an architecture inspired by actor-critic algorithms, we build a stochastic approximation of the fictitious play process. This procedure is on-line, decentralized (an agent has no information of others' strategies and rewards) and applies to multistage games (a generalization of normal form games). In addition, we prove convergence of our method towards a Nash equilibrium in both the cases of zero-sum two-player multistage games and cooperative multistage games. We also provide empirical evidence of the soundness of our approach on the game of Alesia with and without function approximation.

Sequential Test for the Lowest Mean: From Thompson to Murphy Sampling, [39]

Learning the minimum/maximum mean among a finite set of distributions is a fundamental sub-task in planning, game tree search and reinforcement learning. We formalize this learning task as the problem of sequentially testing how the minimum mean among a finite set of distributions compares to a given threshold. We develop refined non-asymptotic lower bounds, which show that optimality mandates very different sampling behavior for a low vs high true minimum. We show that Thompson Sampling and the intuitive Lower Confidence Bounds policy each nail only one of these cases. We develop a novel approach that we call Murphy Sampling. Even though it entertains exclusively low true minima, we prove that MS is optimal for both possibilities. We then design advanced self-normalized deviation inequalities, fueling more aggressive stopping rules. We complement our theoretical guarantees by experiments showing that MS works best in practice.

7.1.4. Online Kernel and Graph-Based Methods

Improved large-scale graph learning through ridge spectral sparsification, [25]

The representation and learning benefits of methods based on graph Laplacians, such as Laplacian smoothing or harmonic function solution for semi-supervised learning (SSL), are empirically and theoretically well supported. Nonetheless , the exact versions of these methods scale poorly with the number of nodes n of the graph. In this paper, we combine a spectral sparsification routine with Laplacian learning. Given a graph G as input, our algorithm computes a sparsifier in a distributed way in $O(n \log 3(n))$ time, $O(m \log 3(n))$ work and $O(n \log(n))$ memory, using only $\log(n)$ rounds of communication. Furthermore , motivated by the regularization often employed in learning algorithms, we show that constructing sparsifiers that preserve the spectrum of the Laplacian only up to the regularization level may drastically reduce the size of the final graph. By constructing a spectrally-similar graph, we are able to bound the error induced by the sparsification for a variety of downstream tasks (e.g., SSL). We empirically validate the theoretical guarantees on Amazon co-purchase graph and compare to the state-of-the-art heuristics.

DPPy: Sampling Determinantal Point Processes with Python, [49]

Determinantal point processes (DPPs) are specific probability distributions over clouds of points that are used as models and computational tools across physics, probability, statistics, and more recently machine learning. Sampling from DPPs is a challenge and therefore we present DPPy, a Python toolbox that gathers known exact and approximate sampling algorithms. The project is hosted on GitHub and equipped with an extensive documentation. This documentation takes the form of a short survey of DPPs and relates each mathematical property with DPPy objects.

Streaming kernel regression with provably adaptive mean, variance, and regularization, [14]

We consider the problem of streaming kernel regression, when the observations arrive sequentially and the goal is to recover the underlying mean function, assumed to belong to an RKHS. The variance of the noise is not assumed to be known. In this context, we tackle the problem of tuning the regularization parameter adaptively at each time step, while maintaining tight confidence bounds estimates on the value of the mean function at each point. To this end, we first generalize existing results for finite-dimensional linear regression with fixed regularization and known variance to the kernel setup with a regularization parameter allowed to be a measurable function of past observations. Then, using appropriate self-normalized inequalities we build upper and lower bound estimates for the variance, leading to Bersntein-like concentration bounds. The later is used in order to define the adaptive regularization. The bounds resulting from our technique are valid uniformly over all observation points and all time steps, and are compared against the literature with numerical experiments. Finally, the potential of these tools is illustrated by an application to kernelized bandits, where we revisit the Kernel UCB and Kernel Thompson Sampling procedures, and show the benefits of the novel adaptive kernel tuning strategy.

7.2. Applications

7.2.1. Dialogue Systems and Natural Language

FiLM: Visual Reasoning with a General Conditioning Layer, [35]

We introduce a general-purpose conditioning method for neural networks called FiLM: Feature-wise Linear Modulation. FiLM layers influence neural network computation via a simple , feature-wise affine transformation based on conditioning information. We show that FiLM layers are highly effective for visual reasoning - answering image-related questions which require a multi-step, high-level process - a task which has proven difficult for standard deep learning methods that do not explicitly model reasoning. Specifically, we show on visual reasoning tasks that FiLM layers 1) halve state-of-the-art error for the CLEVR benchmark, 2) modulate features in a coherent manner, 3) are robust to ablations and architectural modifications, and 4) generalize well to challenging, new data from few examples or even zero-shot.

End-to-End Automatic Speech Translation of Audiobooks, [24]

We investigate end-to-end speech-to-text translation on a corpus of audiobooks specifically augmented for this task. Previous works investigated the extreme case where source language transcription is not available during learning nor decoding , but we also study a midway case where source language transcription is available at training time only. In this case, a single model is trained to decode source speech into target text in a single pass. Experimental results show that it is possible to train compact and efficient end-to-end speech translation models in this setup. We also distribute the corpus and hope that our speech translation baseline on this corpus will be challenged in the future.

Visual Reasoning with Multi-hop Feature Modulation, [42]

Recent breakthroughs in computer vision and natural language processing have spurred interest in challenging multi-modal tasks such as visual question-answering and visual dialogue. For such tasks, one successful approach is to condition image-based convolutional network computation on language via Feature-wise Linear Modulation (FiLM) layers, i.e., per-channel scaling and shifting. We propose to generate the parameters of FiLM layers going up the hierarchy of a convolutional network in a multi-hop fashion rather than all at once, as in prior work. By alternating between attending to the language input and generating FiLM layer parameters, this approach is better able to scale to settings with longer input sequences such as dialogue. We demonstrate that multi-hop FiLM generation achieves state-of-the-art for the short input sequence task ReferIt-on-par with single-hop FiLM generation-while also significantly outperforming prior state-of-the-art and single-hop FiLM generation on the GuessWhat?! visual dialogue task.

7.2.2. Recommendation systems

Recurrent Neural Networks for Long and Short-Term Sequential Recommendation, [54]

Recommender systems objectives can be broadly characterized as modeling user preferences over short- or long-term time horizon. A large body of previous research studied long-term recommendation through dimensionality reduction techniques applied to the historical user-item interactions. A recently introduced session-based recommendation setting highlighted the importance of modeling short-term user preferences. In this task, Recurrent Neural Networks (RNN) have shown to be successful at capturing the nuances of user's interactions within a short time window. In this paper, we evaluate RNN-based models on both short-term and long-term recommendation tasks. Our experimental results suggest that RNNs are capable of predicting immediate as well as distant user interactions. We also find the best performing configuration to be a stacked RNN with layer normalization and tied item embeddings.

Fighting Boredom in Recommender Systems with Linear Reinforcement Learning, [43]

A common assumption in recommender systems (RS) is the existence of a best fixed recommendation strategy. Such strategy may be simple and work at the item level (e.g., in multi-armed bandit it is assumed one best fixed arm/item exists) or implement more sophisticated RS (e.g., the objective of A/B testing is to find the best fixed RS and execute it thereafter). We argue that this assumption is rarely verified in practice, as the recommendation process itself may impact the user's preferences. For instance, a user may get bored by a strategy, while she may gain interest again, if enough time passed since the last time that strategy was used. In this case, a better approach consists in alternating different solutions at the right frequency to fully exploit their potential. In this paper, we first cast the problem as a Markov decision process, where the rewards are a linear function of the recent history of actions, and we show that a policy considering the long-term influence of the recommendations may outperform both fixed-action and contextual greedy policies. We then introduce an extension of the UCRL algorithm (LINUCRL) to effectively balance exploration and exploitation in an unknown environment, and we derive a regret bound that is independent of the number of states. Finally, we empirically validate the model assumptions and the algorithm in a number of realistic scenarios.

7.2.3. Autonomous car

A Survey of State-Action Representations for Autonomous Driving, [51]

Approximate Robust Control of Uncertain Dynamical Systems, [40]

This work studies the design of safe control policies for large-scale non-linear systems operating in uncertain environments. In such a case, the robust control framework is a principled approach to safety that aims to maximize the worst-case performance of a system. However, the resulting optimization problem is generally intractable for non-linear systems with continuous states. To overcome this issue, we introduce two tractable methods that are based either on sampling or on a conservative approximation of the robust objective. The proposed approaches are applied to the problem of autonomous driving.

7.2.4. Software development

Correctness Attraction: A Study of Stability of Software Behavior Under Runtime Perturbation, [12]

Can the execution of a software be perturbed without breaking the correctness of the output? In this paper, we devise a novel protocol to answer this rarely investigated question. In an experimental study, we observe that many perturbations do not break the correctness in ten subject programs. We call this phenomenon "correctness attraction". The uniqueness of this protocol is that it considers a systematic exploration of the perturbation space as well as perfect oracles to determine the correctness of the output. To this extent, our findings on the stability of software under execution perturbations have a level of validity that has never been reported before in the scarce related work. A qualitative manual analysis enables us to set up the first taxonomy ever of the reasons behind correctness attraction.

This paper has attracted a significant interest in the SE community. This work has been invited for an oral presentation (along a 1 page summary) at the 40th International Conference on Software Engineering, the main conference in software engineering. It has then been invited on the IEEE Software review blog.

SMPyBandits: an Experimental Framework for Single and Multi-Players Multi-Arms Bandits Algorithms in Python, [46]

SMPyBandits is a package for numerical simulations on single-player and multi-players Multi-Armed Bandits (MAB) algorithms, written in Python (2 or 3). This library is the most complete open-source implementation of state-of-the-art algorithms tackling various kinds of sequential learning problems referred to as Multi-Armed Bandits. It is extensive, simple to use and maintain, with a clean and well documented codebase. It allows fast prototyping of experiments, with an easy configuration system and command-line options to customize experiments.

Lilian Besson developed a library for multi-armed bandit algorithms in Python for single and multi-player bandits.

7.2.5. Deep Learning

FiLM: Visual Reasoning with a General Conditioning Layer, [35]

We introduce a general-purpose conditioning method for neural networks called FiLM: Feature-wise Linear Modulation. FiLM layers influence neural network computation via a simple , feature-wise affine transformation based on conditioning information. We show that FiLM layers are highly effective for visual reasoning - answering image-related questions which require a multi-step, high-level process - a task which has proven difficult for standard deep learning methods that do not explicitly model reasoning. Specifically, we show on visual reasoning tasks that FiLM layers 1) halve state-of-the-art error for the CLEVR benchmark, 2) modulate features in a coherent manner, 3) are robust to ablations and architectural modifications, and 4) generalize well to challenging, new data from few examples or even zero-shot.

Feature-wise transformations, [13]

i-RevNet: Deep Invertible Networks, [32]

It is widely believed that the success of deep convolutional networks is based on progressively discarding uninformative variability about the input with respect to the problem at hand. This is supported empirically by the difficulty of recovering images from their hidden representations, in most commonly used network architectures. In this paper we show via a one-to-one mapping that this loss of information is not a necessary condition to learn representations that generalize well on complicated problems, such as ImageNet. Via a cascade of homeomorphic layers, we build the i-RevNet, a network that can be fully inverted up to the final

projection onto the classes, i.e. no information is discarded. Building an invertible architecture is difficult, for one, because the local inversion is ill-conditioned, we overcome this by providing an explicit inverse. An analysis of i-RevNets learned representations suggests an alternative explanation for the success of deep networks by a progressive contraction and linear separation with depth. To shed light on the nature of the model learned by the i-RevNet we reconstruct linear interpolations between natural image representations.

Compressing the Input for CNNs with the First-Order Scattering Transform, [33]

We study the first-order scattering transform as a candidate for reducing the signal processed by a convolutional neural network (CNN). We study this transformation and show theoretical and empirical evidence that in the case of natural images and sufficiently small translation invariance, this transform preserves most of the signal information needed for classification while substantially reducing the spatial resolution and total signal size. We show that cascading a CNN with this representation performs on par with ImageNet classification models commonly used in downstream tasks such as the ResNet-50. We subsequently apply our trained hybrid ImageNet model as a base model on a detection system, which has typically larger image inputs. On Pascal VOC and COCO detection tasks we deliver substantial improvements in the inference speed and training memory consumption compared to models trained directly on the input image.

Visual Reasoning with Multi-hop Feature Modulation, [42]

Recent breakthroughs in computer vision and natural language processing have spurred interest in challenging multi-modal tasks such as visual question-answering and visual dialogue. For such tasks, one successful approach is to condition image-based convolutional network computation on language via Feature-wise Linear Modulation (FiLM) layers, i.e., per-channel scaling and shifting. We propose to generate the parameters of FiLM layers going up the hierarchy of a convolutional network in a multi-hop fashion rather than all at once, as in prior work. By alternating between attending to the language input and generating FiLM layer parameters, this approach is better able to scale to settings with longer input sequences such as dialogue. We demonstrate that multi-hop FiLM generation achieves state-of-the-art for the short input sequence task ReferIt-on-par with single-hop FiLM generation-while also significantly outperforming prior state-of-the-art and single-hop FiLM generation on the GuessWhat?! visual dialogue task.

8. Bilateral Contracts and Grants with Industry

8.1. Bilateral Contracts with Industry

8.1.1. Lelivrescolaire.fr

- contract with http://Lelivrescolaire.fr; PI: Michal Valko

  Title: Sequential Machine Learning for Adaptive Educational Systems

  Duration: Mar. 2018 – Feb. 2021

  Abstract: Adaptive educational content are technologies which adapt to the difficulties encountered by students. With the rise of digital content in schools, the mass of data coming from education enables but also ask for machine learning methods. Since 2010, Lelivrescolaire.fr has been developing some learning materials for teachers and students through collaborative creation process. For instance, during the school year 2015/2016, students has achieved more than 8 000 000 exercises on its homework platform Afterclasse.fr. Our approach would be based on sequential machine learning: the algorithm learns to recommend some exercises which adapt to students gradually as they answer. **Participants:** Julien Seznec, Michal Valko.

8.1.2. Sidexa

- contract with “Sidexa”; PI: Philippe Preux

  Title: vision applied to the segmentation and recognition of cars and car related documents.

Duration: 6 months

Abstract: this is a follow-up to the successful contract realized in 2017 with Sidexa. We studied multi-class supervised classification problems in order to classify documents related to a car, and also to identify various characteristics of a car, such as its color, its make, its type.

This work is done with an InriaTech engineer.
Participant: Philippe Preux.

8.1.3. Renault

- contract with Renault; PI: Philippe Preux

  Title: Control of an autonomous vehicle

  Duration: 3 years (12/2017–11/2020)

  Abstract: This contract comes along the CIFRE grant on the same topic. This work is done in collaboration with the NON-A team-project.
  Participants: Édouard Leurent, Odalric Maillard, Philippe Preux.

8.1.4. Critéo

- contract with "Criteo"; PI: Philippe Preux

  Title: Computational advertizing

  Duration: 3 years (12/2017–11/2020)

  Abstract: This contract comes along the CIFRE grant on the same topic. The goal is to investigate reinforcement learning and deep learning on the problem of ad selection on the Internet.
  Participants: Philippe Preux, Kiewan Villatel.

8.1.5. Orange Labs

- contract with "Orange Labs"; PI: Olivier Pietquin

  Title: Inter User Transfer in dialogue systems

  Duration: 3 years

  Abstract: This contract comes along the CIFRE grant on the same topic. The research aims at developing new algorithms to learn fast adaptation strategies for dialogue systems when a new user starts using them while we collected data from previous interactions with other users. Especially, it addresses the cold-start problem encountered when a new user faces the system, before samples can be collected to optimize the interaction strategy.
  Participants: Merwan Barlier, Nicolas Carrara, Olivier Pietquin.

8.1.6. 55

- contract with "55"; PI: Jérémie Mary

  Title: Novel Learning and Exploration-Exploitation Methods for Effective Recommender Systems

  Duration: Oct. 2015 – Sep. 2018

  Abstract: This contract comes along the CIFRE grant on the same topic. In this Ph.D. thesis we intend to deal with this problem by developing novel and more sophisticated recommendation strategies in which the collection of data and the improvement of the performance are considered as a unique process, where the trade-off between the quality of the data and the performance of the recommendation strategy is optimized over time. This work also consider tensor methods (one layer of the tensor can be the time) with the goal to scale them at RS level.

  The PhD was defended in Fall 2018.
  Participants: Jérémie Mary, Romain Warlop.

8.1.7. AB-Tasty

- Thompson Sampling for A/B/C Testing with Delayed Conversions; PI: Émilie Kaufmann

  Duration: 1 month

  Abstract: We investigated the use of Thompson Sampling as well as other state-of-the-art methods for the stochastic MAB problem in the context of delayed feedback. We provided theoretical justification for a method developed by AB Tasty, and proposed some variants of it, as well as a comparison with existing methods from the literature.
  Participant: Émilie Kaufmann.

9. Partnerships and Cooperations

9.1. National Initiatives

9.1.1. ANR BoB

Participant: Michal Valko.

- *Title*: Bayesian statistics for expensive models and tall data
- *Type*: National Research Agency
- *Coordinator*: CNRS (Rémi Bardenet)
- *Duration*: 2016-2020
- *Abstract*:

  Bayesian methods are a popular class of statistical algorithms for updating scientific beliefs. They turn data into decisions and models, taking into account uncertainty about models and their parameters. This makes Bayesian methods popular among applied scientists such as biologists, physicists, or engineers. However, at the heart of Bayesian analysis lie 1) repeated sweeps over the full dataset considered, and 2) repeated evaluations of the model that describes the observed physical process. The current trends to large-scale data collection and complex models thus raises two main issues. Experiments, observations, and numerical simulations in many areas of science nowadays generate terabytes of data, as does the LHC in particle physics for instance. Simultaneously, knowledge creation is becoming more and more data-driven, which requires new paradigms addressing how data are captured, processed, discovered, exchanged, distributed, and analyzed. For statistical algorithms to scale up, reaching a given performance must require as few iterations and as little access to data as possible. It is not only experimental measurements that are growing at a rapid pace. Cell biologists tend to have scarce data but large-scale models of tens of nonlinear differential equations to describe complex dynamics. In such settings, evaluating the model once requires numerically solving a large system of differential equations, which may take minutes for some tens of differential equations on today's hardware. Iterative statistical processing that requires a million sequential runs of the model is thus out of the question. In this project, we tackle the fundamental cost-accuracy trade-off for Bayesian methods, in order to produce generic inference algorithms that scale favorably with the number of measurements in an experiment and the number of runs of a statistical model. We propose a collection of objectives with different risk-reward trade-offs to tackle these two goals. In particular, for experiments with large numbers of measurements, we further develop existing subsampling-based Monte Carlo methods, while developing a novel decision theory framework that includes data constraints. For expensive models, we build an ambitious programme around Monte Carlo methods that leverage determinantal processes, a rich class of probabilistic tools that lead to accurate inference with limited model evaluations. In short, using innovative techniques such as subsampling-based Monte Carlo and determinantal point processes, we propose in this project to push the boundaries of the applicability of Bayesian inference.

9.1.2. ANR Badass

Participants: Odalric Maillard, Émilie Kaufmann.

- *Title*: BAnDits for non-Stationarity and Structure
- *Type*: National Research Agency
- *Coordinator*: Inria Lille (O. Maillard)
- *Duration*: 2016-2020
- *Abstract*: Motivated by the fact that a number of modern applications of sequential decision making require developing strategies that are especially robust to change in the stationarity of the signal, and in order to anticipate and impact the next generation of applications of the field, the BADASS project intends to push theory and application of MAB to the next level by incorporating non-stationary observations while retaining near optimality against the best not necessarily constant decision strategy. Since a non-stationary process typically decomposes into chunks associated with some possibly hidden variables (states), each corresponding to a stationary process, handling non-stationarity crucially requires exploiting the (possibly hidden) structure of the decision problem. For the same reason, a MAB for which arms can be arbitrary non-stationary processes is powerful enough to capture MDPs and even partially observable MDPs as special cases, and it is thus important to jointly address the issue of non-stationarity together with that of structure. In order to advance these two nested challenges from a solid theoretical standpoint, we intend to focus on the following objectives: *(i)* To broaden the range of optimal strategies for stationary MABs: current strategies are only known to be provably optimal in a limited range of scenarios for which the class of distribution (structure) is perfectly known; also, recent heuristics possibly adaptive to the class need to be further analyzed. *(ii)* To strengthen the literature on pure sequential prediction (focusing on a single arm) for non-stationary signals via the construction of adaptive confidence sets and a novel measure of complexity: traditional approaches consider a worst-case scenario and are thus overly conservative and non-adaptive to simpler signals. *(iii)* To embed the low-rank matrix completion and spectral methods in the context of reinforcement learning, and further study models of structured environments: promising heuristics in the context of e.g. contextual MABs or Predictive State Representations require stronger theoretical guarantees.

  This project will result in the development of a novel generation of strategies to handle non-stationarity and structure that will be evaluated in a number of test beds and validated by a rigorous theoretical analysis. Beyond the significant advancement of the state of the art in MAB and RL theory and the mathematical value of the program, this JCJC BADASS is expected to strategically impact societal and industrial applications, ranging from personalized health-care and e-learning to computational sustainability or rain-adaptive river-bank management to cite a few.

9.1.3. ANR ExTra-Learn

Participants: Jérémie Mary, Michal Valko.

- *Title*: Extraction and Transfer of Knowledge in Reinforcement Learning
- *Type*: National Research Agency (ANR-9011)
- *PI*: M. Valko
- *Duration*: 2014-2018
- *Abstract*: ExTra-Learn is directly motivated by the evidence that one of the key features that allows humans to accomplish complicated tasks is their ability of building knowledge from past experience and transfer it while learning new tasks. We believe that integrating transfer of learning in machine learning algorithms will dramatically improve their learning performance and enable them to solve complex tasks. We identify in the reinforcement learning (RL) framework the most suitable candidate for this integration. RL formalizes the problem of learning an optimal control policy from the experience directly collected from an unknown environment. Nonetheless, practical limitations of current algorithms encouraged research to focus on how to integrate prior knowledge

into the learning process. Although this improves the performance of RL algorithms, it dramatically reduces their autonomy. In this project we pursue a paradigm shift from designing RL algorithms incorporating prior knowledge, to methods able to incrementally discover, construct, and transfer "prior" knowledge in a fully automatic way. More in detail, three main elements of RL algorithms would significantly benefit from transfer of knowledge. *(i)* For every new task, RL algorithms need exploring the environment for a long time, and this corresponds to slow learning processes for large environments. Transfer learning would enable RL algorithms to dramatically reduce the exploration of each new task by exploiting its resemblance with tasks solved in the past. *(ii)* RL algorithms evaluate the quality of a policy by computing its state-value function. Whenever the number of states is too large, approximation is needed. Since approximation may cause instability, designing suitable approximation schemes is particularly critical. While this is currently done by a domain expert, we propose to perform this step automatically by constructing features that incrementally adapt to the tasks encountered over time. This would significantly reduce human supervision and increase the accuracy and stability of RL algorithms across different tasks. *(iii)* In order to deal with complex environments, hierarchical RL solutions have been proposed, where state representations and policies are organized over a hierarchy of subtasks. This requires a careful definition of the hierarchy, which, if not properly constructed, may lead to very poor learning performance. The ambitious goal of transfer learning is to automatically construct a hierarchy of skills, which can be effectively reused over a wide range of similar tasks.

9.1.4. Grant of Fondation Mathématique Jacques Hadamard

Participants: Michal Valko, Matteo Pirotta, Alessandro Lazaric, Ronan Fruit.

- *Title*: Theoretically grounded efficient algorithms for high-dimensional and continuous reinforcement learning
- *Type*: PGMO-IRMO, funded by Criteo
- *PI*: M. Valko
- *Criteo contact*: Marc Abeille
- *Duration*: 2018-2020
- *Abstract*: While learning how to behave optimally in an unknown environment, a reinforcement learning (RL) agent must trade off the exploration needed to collect new information about the dynamics and reward of the environment, and the exploitation of the experience gathered so far to gain as much reward as possible. A good measure of the agent's performance is the regret, which measures the difference between the performance of optimal policy and the actual rewards accumulated by the agent. Two common approaches to the exploration-exploitation dilemma with provably good regret guarantees are the optimism in the face of uncertainty principle and Thompson Sampling. While these approaches have been successfully applied to small environments with a finite number of states and action (tabular scenario), existing approach for large or continuous environments either rely on heuristics and come with no regret guarantees, or can be proved to achieve small regret but cannot be implemented efficiently. In this project, we propose to make a significant contribution in the understanding of large and/or continuous RL problems by developing and analyzing new algorithms that perform well both in theory and practice. This research line can have a practical impact in all the applications requiring continuous interaction with an unknown environment. Recommendation systems belong to this category and, by definition, they can be modeled has a sequence of repeated interaction between a learning agent and a large (possibly continuous) environment.

9.1.5. National Partners

- ENS Paris-Saclay
  - M. Valko collaborated with V. Perchet on structured bandit problem. They co-supervise a PhD student (P. Perrault) together.

- Institut de Mathématiques de Toulouse, then Ecole Normale Supérieure de Lyon
  - E. Kaufmann collaborated with Aurélien Garivier on sequential testing and structured bandit problems.
- Centrale-Supélec Rennes
  - E. Kaufmann co-advises Lilian Besson, who works at CentraleSupélec with Christophe Moy on MAB for cognitive radio and Internet-of-Things communications.
- Participation to the Inria Project Lab (IPL) "HPC – Big Data". Started in 2018, this IPL gathers a dozen Inria team-projects, mixing researchers in HPC with researchers in machine learning and data science. SEQUEL contribution in this project is about how we can take advantage of HPC for our computational needs regarding deep learning and deep reinforcement learning, and also how such learning algorithms might be redesigned or re-implemented in order to take advantage of HPC architectures.

9.2. European Initiatives

9.2.1. FP7 & H2020 Projects

9.2.1.1. H2020 BabyRobot

Program: H2020

Project acronym: BabyRobot

Project title: Child-Robot Communication and Collaboration

Duration: 01/2016 - 12/2018

Coordinator: Alexandros Potamianos (Athena Research and Innovation Center in Information Communication and Knowledge Technologies, Greece)

Other partners: Institute of Communication and Computer Systems (Greece), The University of Hertfordshire Higher Education Corporation (UK), Universitaet Bielefeld (Germany), Kunlgliga Tekniska Hoegskolan (Sweden), Blue Ocean Robotics ApS (Denmark), Univ. Lille (France), Furhat Robotics AB (Sweden)

Abstract: The crowning achievement of human communication is our unique ability to share intentionality, create and execute on joint plans. Using this paradigm we model human-robot communication as a three step process: sharing attention, establishing common ground and forming shared goals. Prerequisites for successful communication are being able to decode the cognitive state of people around us (mind reading) and building trust. Our main goal is to create robots that analyze and track human behavior over time in the context of their surroundings (situational) using audio-visual monitoring in order to establish common ground and mind-reading capabilities. On BabyRobot we focus on the typically developing and autistic spectrum children user population. Children have unique communication skills, are quick and adaptive learners, eager to embrace new robotic technologies. This is especially relevant for special education where the development of social skills is delayed or never fully develops without intervention or therapy. Thus our second goal is to define, implement and evaluate child-robot interaction application scenarios for developing specific socio-affective, communication and collaboration skills in typically developing and autistic spectrum children. We will support not supplant the therapist or educator, working hand-in hand to create a low risk environment for learning and cognitive development. Breakthroughs in core robotic technologies are needed to support this research mainly in the areas of motion planning and control in constrained spaces, gestural kinematics, sensorimotor learning and adaptation. Our third goal is to push beyond the state-of-the-art in core robotic technologies to support natural human-robot interaction and collaboration for edutainment and healthcare applications. Creating robots that can establish communication protocols and form collaboration plans on the fly will have impact beyond the application scenarios investigated here.

9.2.1.2. CHIST-ERA DELTA

Participants: Michal Valko, Émilie Kaufmann.

Program: CHIST-ERA

Project acronym: DELTA

Project title: Dynamically Evolving Long-Term Autonomy

Duration: October 2017 - December 2021

Coordinator: Anders Jonsson (PI)

Inria Coordinator: Michal Valko

Other partners: UPF Spain, MUL Austria, ULG Belgium

Abstract: Many complex autonomous systems (e.g., electrical distribution networks) repeatedly select actions with the aim of achieving a given objective. Reinforcement learning (RL) offers a powerful framework for acquiring adaptive behavior in this setting, associating a scalar reward with each action and learning from experience which action to select to maximise long-term reward. Although RL has produced impressive results recently (e.g., achieving human-level play in Atari games and beating the human world champion in the board game Go), most existing solutions only work under strong assumptions: the environment model is stationary, the objective is fixed, and trials end once the objective is met. The aim of this project is to advance the state of the art of fundamental research in lifelong RL by developing several novel RL algorithms that relax the above assumptions. The new algorithms should be robust to environmental changes, both in terms of the observations that the system can make and the actions that the system can perform. Moreover, the algorithms should be able to operate over long periods of time while achieving different objectives. The proposed algorithms will address three key problems related to lifelong RL: planning, exploration, and task decomposition. Planning is the problem of computing an action selection strategy given a (possibly partial) model of the task at hand. Exploration is the problem of selecting actions with the aim of mapping out the environment rather than achieving a particular objective. Task decomposition is the problem of defining different objectives and assigning a separate action selection strategy to each. The algorithms will be evaluated in two realistic scenarios: active network management for electrical distribution networks, and microgrid management. A test protocol will be developed to evaluate each individual algorithm, as well as their combinations.

9.2.1.3. CHIST-ERA IGLU

Program: CHIST-ERA

Project acronym: IGLU

Project title: Interactively Grounded Language Understanding

Duration: 11/2015 - 10/2018

Coordinator: Jean Rouat (Université de Sherbrooke, Canada)

Other partners: UMONS (Belgique), Inria (France), Univ-Lille (France), KTH (sweden), Universidad de Zaragoza (Spain)

Abstract: Language is an ability that develops in young children through joint interaction with their caretakers and their physical environment. At this level, human language understanding could be referred as interpreting and expressing semantic concepts (e.g. objects, actions and relations) through what can be perceived (or inferred) from current context in the environment. Previous work in the field of artificial intelligence has failed to address the acquisition of such perceptually-grounded knowledge in virtual agents (avatars), mainly because of the lack of physical embodiment (ability to interact physically) and dialogue, communication skills (ability to interact verbally). We believe that robotic agents are more appropriate for this task, and that interaction is a so important aspect of human language learning and understanding that pragmatic knowledge (identifying or conveying intention) must be present to complement semantic knowledge. Through a developmental approach

where knowledge grows in complexity while driven by multimodal experience and language interaction with a human, we propose an agent that will incorporate models of dialogues, human emotions and intentions as part of its decision-making process. This will lead anticipation and reaction not only based on its internal state (own goal and intention, perception of the environment), but also on the perceived state and intention of the human interactant. This will be possible through the development of advanced machine learning methods (combining developmental, deep and reinforcement learning) to handle large-scale multimodal inputs, besides leveraging state-of-the-art technological components involved in a language-based dialog system available within the consortium. Evaluations of learned skills and knowledge will be performed using an integrated architecture in a culinary use-case, and novel databases enabling research in grounded human language understanding will be released.

9.3. International Initiatives

9.3.1. Inria Associate Teams Not Involved in an Inria International Labs

9.3.1.1. Allocate

Participants: Pierre Perrault, Julien Seznec, Michal Valko, Émilie Kaufmann, Odalric Maillard.

Title: Adaptive allocation of resources for recommender systems

Inria contact: Michal Valko

International Partner (Institution - Laboratory - Researcher):

Otto-von-Guericke-Universität Magdeburg A. Carpentier

Start year: 2017

We plan to improve a practical scenario of *resource allocation in market surveys*, such as product appraisals and music recommendation. In practice, the market is typically divided into segments: geographic regions, age groups, ...These groups are then queried for preference with some fixed rule of a number of queries per group. This testing is *costly and non-adaptive*. The reason is some groups are easier to estimate than others, but this is impossible to know a priori. Our challenge is **adaptively allocate the optimal number of samples** to each group and improve the efficient of market studies, by providing *sample-efficient* solutions. In 2018 we made big advances that resulted in two new research results, currently under review.

9.3.2. Inria International Partners

9.3.2.1. Declared Inria International Partners

SequeL

Title: The multi-armed bandit problem

International Partner (Institution - Laboratory - Researcher):

University of Leoben (Austria) Peter Auer

Duration: 2014 - 2018

Start year: 2014

In a nutshell, the collaboration is focusing on nonparametric algorithms for active learning problems, mainly involving theoretical analysis of reinforcement learning and bandits problems beyond the traditional settings of finite-state MDPs (for RL) or i.i.d. rewards (for bandits). Peter Auer from University of Leoben is a worldwide leader in the field, having introduced the UCB approach around 2000, along with its finite-time analysis. Today, SequeL is likely to be the largest research group working in this field in the world, enjoying worldwide recognition. SequeL and P. Auer's group have been collaborating for a couple of years now; they have co-authored papers, visited each other (sabbatical stay, post-doc), coorganized workshops; the STREP Complacs partially funds this very active collaboration.

9.3.2.2. CWI

We also collaborate with P. Grunwald, and W. Koolen through the associate team headed by Benjamin Guedj from Modal.

9.3.3. Participation in Other International Programs

In 2017, we mentioned many collaborations with: Adobe, MIT, Stanford, Leoben, ...

Massachusetts Institute of Technology

Victor-Emmanuel Brunel *Collaborator*

M. Valko collaborated with V.-E. Brunel on the estimation of low rank determinantal point processes useful for diverse recommender systems.

Otto-von-Guericke-Universität Magdeburg

Alexandra Carpentier *Collaborator*

M. Valko collaborated with A. Carpentier on adaptive estimation of the block-diagonal matrices with application to market segmentations. This collaboration formalized in September 2017 by creating a north-european associate team. which results in two finished results.

Adobe Research

Y. Abbasi-Yadkori *Collaborator*

M. Valko collaborated on learning in unpredictable but potentially easy environment. This led to a publication in COLT 2018.

University of California, Berkeley

Peter Bartlett *Collaborator*

Victor Gabillon *Collaborator*

Alain Malek *Collaborator*

M. Valko collaborated with P.Barlett, V. Gabillon, and A. Malek on the sample complexities in unknown type of environments.

DeepMind London

Rémi Munos *Collaborator*

M. Valko collaborated with R. Munos on Brownian motion maximization, important for stock value predictions. This led to a publication in NIPS 2018.

Mila, Université de Montréal

A. Courville *Collaborator*

A. Touati *Collaborator*

F. Strub and O. Pietquin collaborate on deep reinforcement learning for language acquisition. This led to several papers at IJCAI, CVPR, and NIPS, as well as the Guesswhat?! dataset and protocol, and the HOME dataset.

M. Valko collaborates on faster learning in submodular learning with limited feedback. This setting has application in marketing when we want to select the inventory while maximizing the profit.

McGill University, Montreal

A. Durand, J. Pineau *Collaborator*

A. Durand and OA. Maillard collaborate on a project of structured bandits, with application in physics (calibration).

Northeastearn University, Boston

M. Aziz, J. Anderton, J. Aslam *Collaborator*

E. Kaufmann collaborate with M. Aziz, J. Anderton and J. Aslam on a project on infinite bandits, which led to an ALT 2018 publication. E. Kaufmann also collaborates with M. Aziz on bandits for phase I clinical trials. This led to the submission of a paper to the Biometrics journal.

9.4. International Research Visitors

9.4.1. Visits of International Scientists

- Xiotian Yu, 1 week, the Chinese University of Hong-Kong
- Junpei Komiyama, 6 weeks, Tokyo University
- Abbas Mehrabian, 1 week, McGill University
- Audrey Durand, 2 weeks, McGill University
- Andrea Locatelli, 2 weeks, Otto-von-Guericke-Universität Magdeburg, Germany
- Jill-Jênn Vie, 1 week, RIKEN AIP, Tokyo, Japan
- Peter Grünwald, 2 times two days (8 hour lectures), CWI and Leiden University, Amsterdam, Netherlands
- Wouter Koolen, 1 week, CWI, Amsterdam, Netherlands

9.4.1.1. Internships

- Quentin Burthier, ENSTA ParisTech, from Jun 2018 until Aug 2018
- Edouard Dendauw, from May 2018 until Jul 2018
- Thibault Felicite, Jul 2018
- Robert Lindland, MIT, from May 2018 until Aug 2018
- Jian Qian, ENS, from May 2018 until Oct 2018
- Hassan Saber, Centrale Paris, from Apr 2018 until Aug 2018
- Benoit Schmitt, Centrale Nantes, from Mar 2018 until Aug 2018
- Han Shao, PhD student from the Chinese University in Hong-Kong, from Oct 2018 until Nov 2018
- Annie Yun, MIT, from May 2018 until Aug 2018
- Arnaud Fanthomme, ENS, from Apr 2018 until Aug 2018

9.4.2. Visits to International Teams

9.4.2.1. Other visits

- OA. Maillard: August, Visit of Aufrey Durand at Mc Gill University (2 weeks)
- OA. Maillard: Novembre, Invited visit of Junya Honda at Tokyo University (4 days)
- E. Kaufmann: March, April, Visit of Wouter Koolen at CWI, Amsterdam (2 times 1 week)

10. Dissemination

10.1. Promoting Scientific Activities

10.1.1. Scientific Events Organisation

10.1.1.1. EWRL 2018

We organized the 14th European Workshop on Reinforcement Learning (EWRL) in October 2018 in Lille. 183 people registered. Despite its name, the audience goes really beyond Europe with 42 from North Amercia, 38 from France, 19 from Germany, 16 from the UK, 12 from Italy, 12 from Israel, 9 from Belgium, ... 40% of participants were students, mostly PhD students, but also some Master students. Among non students, 40% came from industry, the other 60% being academics. We had a quite unique panel of invited speakers highlighted by historical figures of reinforcement learning with Prof. Richard Sutton (U. Alberta and Deepmind), bandit theory with Tze Leung Lai (Stanford U.). EWRL is really the main scientific event on reinforcement learning in the world today.

After 2008 and 2015, this is the third time EWRL is organized in Lille.

10.1.1.2. General Chair, Scientific Chair

M. Valko was a program co-chair for CNRS Summer school on Networks, Graphs, and Machine Learning (RESCOM 2018)

10.1.1.3. Member of the Organizing Committees

- F. Strub, co-organizer of the workshop "Visually Grounded Interaction and Language (ViGIL)" at NIPS 2018
- M. Valko was an organizing co-chair of the ITS Workshop: Optimizing Human Learning (ITS 2018)
- R. Fruit, M. Seurin, M. Pirotta, F. Strub organized the 14th European Workshop on Reinforcement Learning

10.1.2. Scientific Events Selection

10.1.2.1. Member of the Conference Program Committees

- Philippe Preux: SPC IJCAI 2018; PC of ICML, ECML, LOD, EWRL, and French speaking conferences: EGC, SFC JFPDA
- Michal Valko: Area Chair of NIPS 2018, Top 10 reviewer recognition for reviewing at ICML
- Emilie Kaufmann: PC Chair for WiML 2018, Top 10 reviewer recognition for reviewing at ICML
- Odalric-ambrym Maillard: PC chair for ALT

10.1.2.2. Reviewer

Members of SEQUEL have been involved in the following reviewing activities for conferences in 2018:

- AI&Stats, NIPS, ALT, ICML, COLT, IJCAI

10.1.3. Journal

10.1.3.1. Reviewer - Reviewing Activities

- The Annals of Statistics
- Journal of Machine Learning Research
- Machine Learning Research
- Bernoulli
- Annual Reviews in Control
- European Journal of Operation Research
- Information and Inference: a Journal of the IMA (Institute of Mathematics and its Application)
- Operations Research
- IEEE Transactions on Signal Processing

10.1.4. Invited Talks

- Odalric-Ambrym Maillard: invited Opponent for the PhD defence of Stefan Magureanu (Stockholm, Sweden), February 2018
- Odalric-Ambrym Maillard: invited speaker at LTCI, Telecom ParisTech, February 2018
- Odalric-Ambrym Maillard: invited speaker at Journées Probabilités et statistiques de Lille, June 22, 2018
- Odalric-Ambrym Maillard: invited speaker at RL Lab, McGill University, August 23, 2018
- Odalric-Ambrym Maillard: invited speaker at the 21st IBIS conference (Sapporo, Japan), November 06, 2018
- Odalric-Ambrym Maillard: invited speaker at the RIKEN Institute (Tokyo, Japan), November 07, 2018

- Michal Valko: *The power of graphs in speeding up online learning and decision making* Presented on October 23rd, DeepMind, London, UK (*DeepMind 2018*)
- Michal Valko: *Active block-matrix completion with adaptive confidence sets*, Presented on September 10–13th, 2018, International Workshop on Optimization and Machine Learning, CIMI, Toulouse (*CIMI 2018*)
- Michal Valko: *Online influence maximization*, Presented on May 14th, 2018, Workshop on Graph Learning, LINCS, Paris (*LINCS 2018*)
- Michal Valko: *Recommender systems*, Presented on March 22nd, 2018, Journée Big data, Polytech'Lille (*Polytech'Lille 2018*)
- Michal Valko: *Pliable rejection sampling*, Presented on February 8th, 2018 at GDR Isis, Télécom ParisTech in Paris (*ISIS 2018*)
- Michal Valko: *Graph Bandits*, Presented on January 7th, 2018 at MIST conference in Rajecká Lesná (*MIST 2018*)
- Pierre Perrault: *Stochastic multi-arm bandit problem and some extensions*, Presented on November, 23rd, 2018 at Lambda seminar at Université de Bordeaux (*Lambda 2018*)
- Emilie Kaufmann: *(Optimal) Best Arm Identification and applications to Monte-Carlo Tree Search*, presented on January, 18th, 2018 at the Probability and Statistics seminar of IECL, Nancy
- Emilie Kaufmann: *Bandits (for) Games*, presented on March 26th, 2018 at Amazon Research, Berlin
- Emilie Kaufmann, *Bandits (for) Games*, presented on April 25th, 2018 as an invited talk to the Workshop on Modern Challenges on Learning Theory at Univerisé de Montréal
- Emilie Kaufmann, *Bandits (for) Games*, presented on June 13th, 2018 as an invited talk to the Paris Symposium on Game Theory, Paris
- Emilie Kaufmann *New tools for Adaptive Testing and Applications to Bandit Problems*, presented on December 3rd, 2018 at the Probability and Statistics seminar of IRMA, Strasbourg

10.1.5. Scientific Expertise

- Philippe Preux was a member of the hiring committee for CR at Inria Nancy
- Philippe Preux was a member of the hiring committee for an associate professor at Université de Lille
- Philippe Preux evaluated submissions to ANRT (he also declined many such invitations due to lack of time)
- Philippe Preux was a member of an auditing committee of an international company which can not be named (NDA)
- Philippe Preux participates to a "AI mission" with an (other) international company which can not be named (NDA)
- Odalric-Ambrym Maillard evaluated a submission to OTKA (Hungarian ANR), and to ANR.
- M. Valko is an elected member of the evaluation committee and participates in the hiring, promotion, and evaluation juries of Inria, notably
  - Selection committee for Inria award for scientific excellence of confirmed researchers
  - Inria working group for the creation of team RandOpt
  - National committee for the secondments at Inria
- Michal Valko participates in a collaboration with an international company which can not be named (NDA)
- Emilie Kaufmann was a member of the hiring committee for an associate professor position at Université de Lille

- Emilie Kaufmann was a member of the hiring committee for an associate professor at ENS Paris (Computer Science departement)

10.1.6. Research Administration

- Philippe Preux is:
  - "délégué scientifique adjoint" of the Inria center in Lille
  - member of the Inria evaluation committee (CE)
  - member of the Inria internal scientific committee (COSI)
  - member of the scientific committee of CRIStAL
  - the head of the "Data Intelligence" thematic group at CRIStAL
- Michal Valko is a member of the Inria evaluation committee (CE)

10.2. Teaching - Supervision - Juries

10.2.1. Teaching

Master: E. Kaufmann, Spring 2018, Data Mining, M1 Maths/Finances, Université de Lille (36 hours)

Master: E. Kaufmann, Spring 2018, Machine Learning, M2 Maths/Finances, Université de Lille (18 hours)

Master: M. Valko, 2018/2019: Graphs in Machine Learning, 36h eqTD, M2, ENS Cachan

Master: O. Maillard, Spring 2018: Sequential Learning course, parcours DAD, 30h eqTD, Ecole Centrale Lille.

Master: O. Maillard, January 2018: Sequential Learning tutorial, Technicolor, 6h eqTD, Rennes

10.2.2. Supervision

PhD completion: Merwan Barlier, Human-in-the loop reinforcement learning for dialogue systems, started Oct. 2014, advisor: Olivier Pietquin

PhD completion: Alexandre Bérard, Deep learning for post-editing and automatic translation, started Oct. 2014, advisor: Olivier Pietquin

PhD in progress: Lilian Besson, Bandit approach to improve Internet Of Things Communications, started Oct. 2016, advisor: Émilie Kaufmann, Christophe Moy (CentraleSupélec Rennes)

PhD in progress: Ronan Fruit, Exploration-exploitation in hierarchical reinforcement learning, Inria, started Dec. 2015, advisor: Daniil Ryabko, Alessandro Lazaric

PhD in progress: Guillaume Gautier, DPPs in ML, started Oct. 2016, advisor: Michal Valko; Rémi Bardenet

PhD in progress: Jean-Bastien Grill, Création et analyse d'algorithmes efficaces pour la prise de décision dans un environnement inconnu et incertain, Inria/ENS Paris/Lille 1, started Oct. 2014, advisor: Rémi Munos, Michal Valko

PhD in progress: Édouard Leurent, Autonomous vehicle control: application of machine learning to contextualized path planning, started Oct. 2017, advisor: Odalric Maillard, Philippe Preux, Denis Effimov (NON-A), Wilfrid Perruquetti (NON-A)

PhD aborted: Sheikh Waqas Akhtar, Bandits for non-stationarity and structure, started Oct. 2017, advisor: Odalric Maillard, Daniil Ryabko.

PhD in progress: Pierre Perrault, Online Learning on Streaming Graphs, started Sep. 2017, advisor: Michal Valko; Vianney Perchet

PhD in progress: Mathieu Seurin, Multi-scale rewards in reinforcement learning, started Oct. 2017, advisor: Olivier Pietquin, Philippe Preux

PhD in progress: Julien Seznec, Sequential Learning for Educational Systems, started Mar. 2017, advisor: Michal Valko; Alessandro Lazaric, Jonathan Banon

PhD in progress: Xuedong Shang, Adaptive methods for optimization in stochastic environments, started Oct. 2017, advisor: Émilie Kaufmann, Michal Valko

PhD in progress: Florian Strub, Reinforcement Learning for visually grounded interaction, started Jan. 2016, advisors: Olivier Pietquin and Jeremie Mary

PhD in progress: Kiewan Villatel, Deep Learning for Conversion Rate Prediction in Online Advertising, started Oct. 2017, advisor: Philippe Preux

PhD in progress: Yannis Flet-Berliac, start Oct. 2018

PhD in progress: Hassan Saber, start Oct. 2018, Structured Multi-armed bandits, advisor: Odalric Maillard, Philippe Preux.

PhD in progress: Omar Darwiche, start Oct. 2018, Sequential Learning in Dynamic Environments, advisor: Émilie Kaufmann, Michal Valko

10.2.3. Juries

PhD and HDR juries:

- É. Kaufmann:
  - Stefan Magureanu, KTH Stockholm, February 20th, 2018
  - Valentin Reis, LIG, Grenoble, September 28th, 2018
  - Maryam Aziz, Northestearn University (Boston), December 6th, 2018
- O. Maillard: Stefan Magureanu, February 20th, 2018
- Ph. Preux:
  - Saeed Varasteh Yazdi, LIG, Grenoble
  - Fabien Vilar, Marseille
  - Merwan Barlier, Lille
- M. Valko:
  - *Pierre Ménard*, Université Toulouse 3 Paul Sabatier, June 2018, Sur la notion d'optimalité dans les problèmes de bandits stochastiques. *Reviewer*
  - *Mariana Vargas Vieyra*, Université Lille, September 2017, Adaptive graph learning with application to natural language processing. *Ph.D. mid-term evaluation reviewer*

10.3. Popularization

10.3.1. Internal or external Inria responsibilities

Philippe Preux chaired the Inria evaluation seminar of theme "Optimization, machine learning and statistical methods" in March 2018.

10.3.2. Articles and contents

- Ph. Preux interviewed for various journals ("Les échos", ...).
- Adobe research highlights M. Valko's work on online influence maximization presented (January 2018)
- Daniele Calandriello (supervised by A. Lazaric and M. Valko) wins the prize for the Best AI Thesis in France in 2018. Articles in:
  - La Voix du Nord
  - CNRS journal
  - Newstank

- Lille1
- Actu

10.3.3. Education

- Ph. Preux presented and animated 3 sessions on AI at the “congrès annuel du réseau national professionnel des cultures scientifique technique et industrielle” (Amcsti)

10.3.4. Interventions

- Philippe Preux:
  - presented AI related to health industry at the yearly general assembly of Eurasanté
  - presented and animated 3 sessions on AI at the “congrès annuel du réseau national professionnel des cultures scientifique technique et industrielle” (Amcsti)
  - participated to a panel at Conext forum (Lille)
- the work on Guesswhat?!:
  - has been invited to be presented on the Inria booth during The Web Conf in Lyon
  - is presented at the Inria showroom inaugurated in Dec. 2018 in Lille

10.3.5. Creation of media or tools for science outreach

- Ph. Preux was interviewed for a video about robots and AI

11. Bibliography

Major publications by the team in recent years

[1] O. CAPPÉ, A. GARIVIER, O.-A. MAILLARD, R. MUNOS, G. STOLTZ. *Kullback-Leibler Upper Confidence Bounds for Optimal Sequential Allocation*, in "Annals of Statistics", 2013, vol. 41, n^{o} 3, pp. 1516-1541, Accepted, to appear in Annals of Statistics, https://hal.archives-ouvertes.fr/hal-00738209

[2] A. CARPENTIER, M. VALKO. *Revealing graph bandits for maximizing local influence*, in "International Conference on Artificial Intelligence and Statistics", Seville, Spain, May 2016, https://hal.inria.fr/hal-01304020

[3] H. DE VRIES, F. STRUB, J. MARY, H. LAROCHELLE, O. PIETQUIN, A. COURVILLE. *Modulating early visual processing by language*, in "Conference on Neural Information Processing Systems", Long Beach, United States, December 2017, https://hal.inria.fr/hal-01648683

[4] N. GATTI, A. LAZARIC, M. ROCCO, F. TROVÒ. *Truthful Learning Mechanisms for Multi–Slot Sponsored Search Auctions with Externalities*, in "Artificial Intelligence", October 2015, vol. 227, pp. 93-139, https://hal.inria.fr/hal-01237670

[5] M. GHAVAMZADEH, Y. ENGEL, M. VALKO. *Bayesian Policy Gradient and Actor-Critic Algorithms*, in "Journal of Machine Learning Research", January 2016, vol. 17, n^{o} 66, pp. 1-53, https://hal.inria.fr/hal-00776608

[6] H. KADRI, E. DUFLOS, P. PREUX, S. CANU, A. RAKOTOMAMONJY, J. AUDIFFREN. *Operator-valued Kernels for Learning from Functional Response Data*, in "Journal of Machine Learning Research (JMLR)", 2016, https://hal.archives-ouvertes.fr/hal-01221329

[7] E. KAUFMANN, O. CAPPÉ, A. GARIVIER. *On the Complexity of Best Arm Identification in Multi-Armed Bandit Models*, in "Journal of Machine Learning Research", January 2016, vol. 17, pp. 1-42, https://hal.archives-ouvertes.fr/hal-01024894

[8] A. LAZARIC, M. GHAVAMZADEH, R. MUNOS. *Analysis of Classification-based Policy Iteration Algorithms*, in "Journal of Machine Learning Research", 2016, vol. 17, pp. 1 - 30, https://hal.inria.fr/hal-01401513

[9] R. MUNOS. *From Bandits to Monte-Carlo Tree Search: The Optimistic Principle Applied to Optimization and Planning*, in "Foundations and Trends in Machine Learning", 2014, vol. 7, n° 1, pp. 1-129, http://dx.doi.org/10.1561/2200000038

[10] R. ORTNER, D. RYABKO, P. AUER, R. MUNOS. *Regret bounds for restless Markov bandits*, in "Journal of Theoretical Computer Science (TCS)", 2014, vol. 558, pp. 62-76 [*DOI :* 10.1016/J.TCS.2014.09.026], https://hal.inria.fr/hal-01074077

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] R. WARLOP. *Novel Learning and Exploration-Exploitation Methods for Effective Recommender Systems*, Lille1, October 2018, https://hal.inria.fr/tel-01915499

Articles in International Peer-Reviewed Journals

[12] B. DANGLOT, P. PREUX, B. BAUDRY, M. MONPERRUS. *Correctness Attraction: A Study of Stability of Software Behavior Under Runtime Perturbation*, in "Empirical Software Engineering", August 2018, vol. 23, n° 4, pp. 2086–2119, https://arxiv.org/abs/1611.09187 [*DOI :* 10.1007/S10664-017-9571-8], https://hal.archives-ouvertes.fr/hal-01378523

[13] V. DUMOULIN, E. PEREZ, H. VRIES, F. STRUB, N. SCHUCHER, A. COURVILLE, Y. BENGIO. *Feature-wise transformations: A simple and surprisingly effective family of conditioning mechanisms*, in "Distill", July 2018, vol. 3, n° 7 [*DOI :* 10.23915/DISTILL.00011], https://hal.inria.fr/hal-01841985

[14] A. DURAND, O.-A. MAILLARD, J. PINEAU. *Streaming kernel regression with provably adaptive mean, variance, and regularization*, in "Journal of Machine Learning Research", 2018, vol. 1, pp. 1 - 48, https://arxiv.org/abs/1708.00768 , https://hal.archives-ouvertes.fr/hal-01927007

[15] E. KAUFMANN, T. BONALD, M. LELARGE. *A spectral algorithm with additive clustering for the recovery of overlapping communities in networks*, in "Theoretical Computer Science", September 2018, vol. 742, pp. 3-26, https://hal.archives-ouvertes.fr/hal-01963868

[16] O.-A. MAILLARD. *Boundary Crossing Probabilities for General Exponential Families*, in "Mathematical Methods of Statistics", 2018, vol. 27, https://hal.archives-ouvertes.fr/hal-01737150

[17] M. S. TALEBI, O.-A. MAILLARD. *Variance-Aware Regret Bounds for Undiscounted Reinforcement Learning in MDPs*, in "Journal of Machine Learning Research", April 2018, pp. 1-36, https://hal.archives-ouvertes.fr/hal-01737142

International Conferences with Proceedings

[18] Y. ABBASI-YADKORI, P. BARTLETT, V. GABILLON, A. MALEK, M. VALKO. *Best of both worlds: Stochastic & adversarial best-arm identification*, in "Conference on Learning Theory", Stockholm, Sweden, 2018, https://hal.inria.fr/hal-01808948

[19] M. AZIZ, J. ANDERTON, E. KAUFMANN, J. ASLAM. *Pure Exploration in Infinitely-Armed Bandit Models with Fixed-Confidence*, in "ALT 2018 - Algorithmic Learning Theory", Lanzarote, Spain, JMLR Workshop and Conference Proceedings, April 2018, https://arxiv.org/abs/1803.04665 , https://hal.archives-ouvertes.fr/hal-01729969

[20] M. BARLIER, R. LAROCHE, O. PIETQUIN. *Training Dialogue Systems With Human Advice*, in "AAMAS 2018 - the 17th International Conference on Autonomous Agents and Multiagent Systems", Stockholm, Sweden, International Foundation for Autonomous Agents and MultiAgent Systems (IFAAMAS), July 2018, 9 p. , https://hal.archives-ouvertes.fr/hal-01945831

[21] P. BARTLETT, V. GABILLON, M. VALKO. *A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption*, in "Algorithmic Learning Theory", Chicago, United States, 2019, https://hal.inria.fr/hal-01885368

[22] L. BESSON, E. KAUFMANN. *Multi-Player Bandits Revisited*, in "Algorithmic Learning Theory", Lanzarote, Spain, Mehryar Mohri and Karthik Sridharan, April 2018, https://arxiv.org/abs/1711.02317 , https://hal.inria.fr/hal-01629733

[23] L. BESSON, E. KAUFMANN, C. MOY. *Aggregation of Multi-Armed Bandits Learning Algorithms for Opportunistic Spectrum Access*, in "IEEE WCNC - IEEE Wireless Communications and Networking Conference", Barcelona, Spain, April 2018 [*DOI :* 10.1109/WCNC.2018.8377070], https://hal.inria.fr/hal-01705292

[24] A. BÉRARD, L. BESACIER, A. C. KOCABIYIKOGLU, O. PIETQUIN. *End-to-End Automatic Speech Translation of Audiobooks*, in "ICASSP 2018 - IEEE International Conference on Acoustics, Speech and Signal Processing", Calgary, Alberta, Canada, April 2018, https://hal.archives-ouvertes.fr/hal-01709586

[25] D. CALANDRIELLO, I. KOUTIS, A. LAZARIC, M. VALKO. *Improved large-scale graph learning through ridge spectral sparsification*, in "International Conference on Machine Learning", Stockholm, Sweden, ICML 2018 - Thirty-fifth International Conference on Machine Learning, July 2018, https://hal.inria.fr/hal-01810980

[26] N. CARRARA, R. LAROCHE, J.-L. BOURAOUI, T. URVOY, O. PIETQUIN. *A Fitted-Q Algorithm for Budgeted MDPs*, in "EWRL 2018 - 14th European workshop on Reinforcement Learning", Lille, France, October 2018, https://hal.archives-ouvertes.fr/hal-01928092

[27] N. CARRARA, R. LAROCHE, J.-L. BOURAOUI, T. URVOY, O. PIETQUIN. *Safe transfer learning for dialogue applications*, in "SLSP 2018 - 6th International Conference on Statistical Language and Speech Processing", Mons, Belgium, October 2018, https://hal.archives-ouvertes.fr/hal-01928102

[28] R. FRUIT, M. PIROTTA, A. LAZARIC. *Near Optimal Exploration-Exploitation in Non-Communicating Markov Decision Processes*, in "32nd Conference on Neural Information Processing Systems", Montréal, Canada, December 2018, https://hal.inria.fr/hal-01941220

[29] R. FRUIT, M. PIROTTA, A. LAZARIC, R. ORTNER. *Efficient Bias-Span-Constrained Exploration-Exploitation in Reinforcement Learning*, in "ICML 2018 - The 35th International Conference on Machine

Learning", Stockholm, Sweden, Proceedings of Machine Learning Research, July 2018, vol. 80, pp. 1578-1586, https://hal.inria.fr/hal-01941206

[30] P. GAJANE, T. URVOY, E. KAUFMANN. *Corrupt Bandits for Preserving Local Privacy*, in "ALT 2018 - Algorithmic Learning Theory", Lanzarote, Spain, Proceedings of Machine Learning Research, April 2018, https://hal.archives-ouvertes.fr/hal-01757297

[31] J.-B. GRILL, M. VALKO, R. MUNOS. *Optimistic optimization of a Brownian*, in "NeurIPS 2018 - Thirty-second Conference on Neural Information Processing Systems", Montréal, Canada, December 2018, https://hal.inria.fr/hal-01906601

[32] J.-H. JACOBSEN, A. SMEULDERS, E. OYALLON. *i-RevNet: Deep Invertible Networks*, in "ICLR 2018 - International Conference on Learning Representations", Vancouver, Canada, April 2018, https://arxiv.org/abs/1802.07088 , https://hal.archives-ouvertes.fr/hal-01712808

[33] E. OYALLON, E. BELILOVSKY, S. ZAGORUYKO, M. VALKO. *Compressing the Input for CNNs with the First-Order Scattering Transform*, in "European Conference on Computer Vision", Munich, Germany, 2018, https://hal.inria.fr/hal-01850921

[34] M. PAPINI, D. BINAGHI, G. CANONACO, M. PIROTTA, M. RESTELLI. *Stochastic Variance-Reduced Policy Gradient*, in "ICML 2018 - 35th International Conference on Machine Learning", Stockholm, Sweden, Proceedings of Machine Learning Research, July 2018, vol. 80, pp. 4026-4035, https://hal.inria.fr/hal-01940394

[35] E. PEREZ, F. STRUB, H. DE VRIES, V. DUMOULIN, A. COURVILLE. *FiLM: Visual Reasoning with a General Conditioning Layer*, in "AAAI Conference on Artificial Intelligence", New Orleans, United States, February 2018, https://arxiv.org/abs/1707.03017 , https://hal.inria.fr/hal-01648685

[36] J. PÉROLAT, B. PIOT, O. PIETQUIN. *Actor-Critic Fictitious Play in Simultaneous Move Multistage Games*, in "AISTATS 2018 - 21st International Conference on Artificial Intelligence and Statistics", Playa Blanca, Lanzarote, Canary Islands, Spain, April 2018, https://hal.inria.fr/hal-01724227

[37] J. SEZNEC, A. LOCATELLI, A. CARPENTIER, A. LAZARIC, M. VALKO. *Rotting bandits are no harder than stochastic ones*, in "International Conference on Artificial Intelligence and Statistics", Okinawa, Japan, 2019, https://hal.inria.fr/hal-01936894

[38] A. TIRINZONI, A. SESSA, M. PIROTTA, M. RESTELLI. *Importance Weighted Transfer of Samples in Reinforcement Learning*, in "ICML 2018 - The 35th International Conference on Machine Learning", Stockholm, Sweden, Proceedings of Machine Learning Research, July 2018, vol. 80, pp. 4936-4945, https://hal.inria.fr/hal-01941213

Conferences without Proceedings

[39] E. KAUFMANN, W. KOOLEN, A. GARIVIER. *Sequential Test for the Lowest Mean: From Thompson to Murphy Sampling*, in "Advances in Neural Information Processing Systems (NIPS)", Montréal, Canada, December 2018, https://arxiv.org/abs/1806.00973 , https://hal.archives-ouvertes.fr/hal-01804581

[40] E. LEURENT, Y. BLANCO, D. EFIMOV, O.-A. MAILLARD. *Approximate Robust Control of Uncertain Dynamical Systems*, in "32nd Conference on Neural Information Processing Systems (NeurIPS 2018) Workshop", Montréal, Canada, December 2018, https://hal.archives-ouvertes.fr/hal-01931744

[41] X. SHANG, E. KAUFMANN, M. VALKO. *Adaptive black-box optimization got easier: HCT only needs local smoothness*, in "European Workshop on Reinforcement Learning", Lille, France, October 2018, https://hal.inria.fr/hal-01874637

[42] F. STRUB, M. SEURIN, E. PEREZ, H. DE VRIES, J. MARY, P. PREUX, A. COURVILLE, O. PIETQUIN. *Visual Reasoning with Multi-hop Feature Modulation*, in "ECCV 2018 - 15th European Conference on Computer Vision", Munich, Germany, V. FERRARI, M. HEBERT, C. SMINCHISESCU, Y. WEISS (editors), Part of the Lecture Notes in Computer Science book series - LNCS, September 2018, vol. 11205-11220, n^{o} 11209, pp. 808-831, https://arxiv.org/abs/1808.04446 , https://hal.archives-ouvertes.fr/hal-01927811

[43] R. WARLOP, A. LAZARIC, J. MARY. *Fighting Boredom in Recommender Systems with Linear Reinforcement Learning*, in "Neural Information Processing Systems", Montreal, Canada, December 2018, https://hal.inria.fr/hal-01915468

Other Publications

[44] R. ALAMI, O.-A. MAILLARD, R. FÉRAUD. *Memory Bandits: Towards the Switching Bandit Problem Best Resolution*, August 2018, MLSS 2018 - Machine Learning Summer School, Poster, https://hal.archives-ouvertes.fr/hal-01879251

[45] L. BESSON. *A Note on the Ei Function and a Useful Sum-Inequality*, July 2018, https://hal.inria.fr/hal-01847480

[46] L. BESSON. *SMPyBandits: an Experimental Framework for Single and Multi-Players Multi-Arms Bandits Algorithms in Python*, July 2018, working paper or preprint, https://hal.inria.fr/hal-01840022

[47] L. BESSON, E. KAUFMANN. *What Doubling Tricks Can and Can't Do for Multi-Armed Bandits*, February 2018, https://arxiv.org/abs/1803.06971 - working paper or preprint, https://hal.inria.fr/hal-01736357

[48] N. CARRARA, R. LAROCHE, J.-L. BOURAOUI, T. URVOY, O. PIETQUIN. *A Fitted-Q Algorithm for Budgeted MDPs*, August 2018, Workshop on Safety, Risk and Uncertainty in Reinforcement Learning. https://sites.google.com/view/rl-uai2018/, https://hal.archives-ouvertes.fr/hal-01867353

[49] G. GAUTIER, R. BARDENET, M. VALKO. *DPPy: Sampling Determinantal Point Processes with Python*, September 2018, working paper or preprint, https://hal.inria.fr/hal-01879424

[50] E. KAUFMANN, W. M. KOOLEN. *Mixture Martingales Revisited with Applications to Sequential Tests and Confidence Intervals*, October 2018, https://arxiv.org/abs/1811.11419 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01886612

[51] E. LEURENT. *A Survey of State-Action Representations for Autonomous Driving*, October 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01908175

[52] Y. LIU, G. RADANOVIC, C. DIMITRAKAKIS, D. MANDAL, D. C. PARKES. *Calibrated Fairness in Bandits*, December 2018, working paper or preprint [*DOI :* 10.1145/NNNNNNN.NNNNNNN], https://hal.inria.fr/hal-01953314

[53] O.-A. MAILLARD, M. ASADI. *Upper Confidence Reinforcement Learning exploiting state-action equivalence*, December 2018, working paper or preprint, https://hal.archives-ouvertes.fr/hal-01945034

[54] K. VILLATEL, E. SMIRNOVA, J. MARY, P. PREUX. *Recurrent Neural Networks for Long and Short-Term Sequential Recommendation*, July 2018, https://arxiv.org/abs/1807.09142 - working paper or preprint, https://hal.inria.fr/hal-01847127

[55] H. VAN HASSELT, Y. DORON, F. STRUB, M. HESSEL, N. SONNERAT, J. MODAYIL. *Deep Reinforcement Learning and the Deadly Triad*, December 2018, https://arxiv.org/abs/1812.02648 - working paper or preprint, https://hal.archives-ouvertes.fr/hal-01949304

References in notes

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

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

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

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

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

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

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

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