Master MVA: Reinforcement Learning **Lecture:**

Course Projects and Internship Proposals

Lecturer: *Alessandro Lazaric* *http://researchers.lille.inria.fr/∼lazaric/Webpage/Teaching.html*

Notes:

- Projects can be done in groups of **two**, but no more than that.
- Additional projects may be included over time.
- Students are allowed to propose additional project and/or integrate this project with other courses' projects.
- We encourage students to finalize their assignment by the end of the month. Even if the project is not supervised by A. Lazaric, please notify your agreement with the person in charge of the project with A. Lazaric as well.
- A report (between 5 to max 10 pages) should be submitted to both the supervisor of the project and A. Lazaric by **January 5th**.
- A 10min presentation will be scheduled in the week 12-16 of January, which will finalize the course project.
- The final grades will be obtained by combining the grade of the project together with the homework and they will be submitted towards the end of January.

1 Course Projects

1.1 Exploration-exploitation in Reinforcement Learning

Topic. Reinforcement learning.

Category. Review.

Description. In the online RL setting, it is crucial to have an effective exploration strategy which allows to visit the whole state space to learn about the dynamics and the reward of the problem. Nonetheless, as in the bandit problem, the objective is to properly balance the exploration of the environment with exploiting the learned strategy so as to collect as much reward as possible over time. A number of exploration-exploitation strategies have been formulated over the last few years. In this project, the student should provide an overview of the literature in both the PAC-MDP and regret minimization settings.

Contact: alessandro.lazaric@inria.fr

1.2 The Arcade Learning Environment

Topic. Reinforcement learning.

Category. Implementation.

Description. The Arcade Learning Environment (http://www.arcadelearningenvironment.org/) is a platform that allows to easily integrate learning algorithms into an Atari emulator with more than 2600 games. The objective of this project is to become familiar with the platform and try to implement very simple RL algorithms in one of the game available and test the performance.

Contact: alessandro.lazaric@inria.fr

1.3 Hierarchical Reinforcement Learning

Topic. Reinforcement learning.

Category. Review.

Description. In the basic RL model, actions are atomic in time (they take one single step to terminate). Unfortunately, this model is often unsuitable to solve problems which very complex structure. In fact whenever the sequence of actions needed to achieve the goal is too long. the learning process becomes unfeasible. In hierarchical RL the idea is to provide a hierarchical structure of actions (and states) which provide more and more high-level/abstract representations of the problem so that low level tasks can be easily solved and composed together to solve higher and more difficult problems. In this project, the student should provide an overview of the hierarchical RL methods (options and MaxQ) and discuss about the approaches designed for an automatic generation of the hierarchical decomposition of a given problem.

Contact: alessandro.lazaric@inria.fr

1.4 Transfer and Multi-task Reinforcement Learning

Topic. Reinforcement learning.

Category. Review.

Description. In RL, one single task is solved at the time and, whenever the task changes, the learning process is restarted from scratch. The idea of transfer learning is to leverage over the experience collected over tasks to automatically build knowledge that can be profitable reused when solving new tasks. This intuitive concept can be applied in many different ways in RL, depending on the setting and the type of knowledge acquired and transferred across tasks. In this project, the student should review the basic literature in transfer RL with particular focus on the difference between the settings and the potential improvements coming from transfer.

Contact: alessandro.lazaric@inria.fr

1.5 Review of Applications of RL

Topic. Reinforcement learning.

Category. Review.

Description. This is an open project intended to review how RL algorithms have been applied to a specific domain of interest (e.g., energy management, robotics, finance). The student should contact me with a series of papers where it is clear the RL component in the solution of a problem in a specific application domain.

Contact: alessandro.lazaric@inria.fr

1.6 Policy Search Algorithms

Topic. Reinforcement learning.

Category. Review.

Description. RL algorithms derived from dynamic programming all share the idea of learning an optimal policy through the estimation of a value function, which is later used for improvement. Another category of approaches, try to directly work on the space of policies and perform a direct optimization in that space. This approach is referred to as "policy search" and it obtained significant results in complex problems, in particular in robotics. In this project, the student should provide an overview of the approach with different examples of algorithms.

Contact: alessandro.lazaric@inria.fr

1.7 Learning the Representation in RL

Topic. Reinforcement learning.

Category. Review.

Description. RL theory is based on the assumption that there exist a state under which the dynamics of the system is fully Markovian. Furthermore, the success of a RL algorithm often relies on the quality of the basis functions used for approximating either the value functions or the policies. Unfortunately, it is not always trivial to provide a suitable definition of state and/or basis functions, unless specific domain knowledge about the problem at hand is available. In order to solve this problem, a series of different approaches could be used to directly learn the representation and/or relax the Markovian assumption on the state definition. In this project, the student should review different approaches of representation learning and their application in RL.

Contact: alessandro.lazaric@inria.fr

1.8 Apprenticeship Learning

Topic. Reinforcement learning.

Category. Review.

Description. Reinforcement learning algorithms are designed to optimize the sum of rewards over time. Nonetheless, in a wide range of applications it is very difficult to provide a clear reinforcement signal which could lead to the desired policy. A typical example is the task of "learning how to drive". In this case, it is very easy to provide the learner with a good policy but it is very difficult to define a suitable reward function. Recently, it has been proposed to deal with these problems using an inverse reinforcement learning approach, where the learner has access to examples of good trajectories and the objective is to recover the reward which admits that behavior as optimal. This problem, usually referred to as *apprenticeship learning*,

has been tackled in many different ways in the past. In this project, the student should review the main approaches available in the literature.

Contact: alessandro.lazaric@inria.fr

1.9 Thompson Sampling: influence of the prior distribution

Topic. Multi-arm bandit.

Category. Research.

Description. Thompson Sampling was the first bandit algorithm, proposed by Thompson in 1933, and its good empirical performances were rediscovered around 2010, even in more complex (e.g. contextual) bandit models. However, theoretical guarantees are currently available only for simple bandit models, and often for a very specific choice of prior. This includes for example parametric bandit models that depends on a single parameter, like Bernoulli bandit models with uniform prior or exponential family bandit models with Jeffrey's prior (see. [1], [2]). Recently, the paper [3] showed that in a simple example of two-parameters bandit models, in which the arms are Gaussian distributions with both the mean and variance unknown, Thompson Sampling is not always optimal.

The goal of this project is to study in depth the influence of the prior on the efficiency of Thompson Sampling. Especially, you will discuss the following conjecture: Thompson Sampling is asymptotically optimal for every choice of prior in one-parameter models, but might be sub-optimal when the arms depend on more parameters.

References

1 *Thompson Sampling: an asymptotically optimal finite-time analysis*, Kaufmann, Korda and Munos, ALT 2012

2 *Thompson Sampling for one-dimensional exponential families*, Korda, Kaufmann and Munos, NIPS 2013

3 *Optimality of Thompson Sampling for Gaussian bandits depends on prior*, Honda and Takemura, AISTATS 2014

Contact: emilie.kaufmann@inria.fr

1.10 Optimistic approaches in Contextual Linear Bandit models

Topic. Multi-arm bandit.

Category. Review + Implementation

Description. In classic stochastic multi-armed bandit models, the arms are assumed to be independent, an assumption that is not realistic in many applications. One of the simplest model that allows for correlated arms is the linear bandit model, in which each arm is parametrized by a vector in $\mathbb{R}^d$ and the mean reward associated to each arm is the dot product with some unknown regression parameter.

In this project, you will present the counterpart of optimistic 'UCB-like' algorithms for this setting, e.g. Lin-UCB [1] or a variant [2] (in which the exploration parameter α is replaced by an adaptive exploration rate). In a 'contextual' setting (in which the set of arms you choose from might evolve over time), you

will implement these algorithms and study the influence of the exploration rate. You can compare with Thompson Sampling adapted for the contextual linear setting (see [3]), or try to understand the analysis of the OFUL algorithm (from [2] or [3]).

References

1 A Contextual-Bandit Approach to Personalized News Article Recommendation, Li, Chu, Langford, Schapire, WWW 2010

2 Improved algorithms for Linear Stochastic Bandits, Abbasi-Yadkori, Pal, Szepesvari, NIPS 2011

3 Chapter 4 of my PhD thesis, available on my website

Contact: emilie.kaufmann@inria.fr

1.11 Thompson Sampling in Contextual Bandits

Topic. Multi-arm bandit.

Category. Implementation + Research.

Description. Thompson Sampling is an old algorithm (it dates back to 1933) but was rediscovered (see e.g. [1]) because of its efficiency in so-called contextual linear bandit models, that can be used for the display of advertising. The goal of this project is to implement Thompson Sampling in the model described in Section 4 of [1], on synthetic (and maybe real) data, and maybe compare with an optimistic approach for this model (see [2]), or with other heuristics of your choice. In Section 4.5 of [3], you will find a more precise description of the logistic model considered by [1]. The papers [4] explains that Thompson Sampling is used by Criteo for displaying advertisement.

References

1 *An emirical evaluation of Thompson Sampling*, Chapelle, Li, NIPS 2011

2 *Parametric bandits: the generalized linear case*, Filippi, Cappé, Garivier, Szepesvari, NIPS 2010

3 Chapter 4 of my PhD thesis (available on my website)

4 Simple and scalable response prediction for display advertising, Chapelle, Manavoglu, Rosales, 2014

Contact. emilie.kaufmann@inria.fr

1.12 Best arm identification versus regret minimization

Topic. Multi-arm bandit.

Category. Review + Implementation.

Description. A stochastic multi-armed bandit model is simply a collection of arms - that are probability distributions - from which an agent has to sequentially choose from. Usually, the samples from the arms collected by the agent are interpreted as rewards (as in the more general reinforcement learning framework), and his goal is to maximize his rewards. An other possible possible objective is to identify the best arm, without suffering a loss when drawing arms with small means. In this setup, called best-arm identification,

the agent has to find a strategy for drawing the arms and has also to decide when to stop so that he can find the best arm with probability larger than $1-\delta$, where δ is some risk parameters. Some algorithms have been proposed in the literature, like Successive Elimination ([1]) or LUCB ([2]).

In this project, you will review algorithms for best-arm identification, explaining for example how they are different from or similar to algorithms for regret minimization. You may want to propose others heuristics for best-arm identification, based on the UCB algorithm, and compare them numerically with algorithms from the literature.

References

1 Action Elimination and Stopping Conditions for the Multi-Armed Bandit and Reinforcement Learning, Even-Dar, Mannor, Mansour, JMLR 2006

2 PAC Subset Selection in Stochastic Multi-Armed Bandits, Kalyanakrishnan, Tewari, Auer, Stone, 2010

Contact. emilie.kaufmann@inria.fr

1.13 Stochastic versus adversarial bandits

Topic. Multi-arm bandit.

Category. Review + Implementation.

Description. In stochastic bandit models, if we make specific assumptions on the distributions of the arms, one can propose very efficient algorithms (like KL-UCB, Thompson Sampling). However, often it is unrealistic to know in advance the distribution of the arms, or even the fact that the rewards are drawn in an i.i.d. fashion. In case of bounded rewards, the EXP-3 algorithm can still be implemented without precise stochastic assumptions. The first part of the project is to compare stochastic and adversarial bandit algorithms studied in class on bandit problems that follow (or don't follow) the stochastic assumption. Then, you will add in the pictures recent algorithms proposed in the literature ([1],[2]), that should be suited for both the stochastic and adversarial frameworks.

References

1 http://jmlr.org/proceedings/papers/v23/bubeck12b/bubeck12b.pdf

2 http://jmlr.org/proceedings/papers/v32/seldinb14-supp.pdf

Contact. emilie.kaufmann@inria.fr

1.14 Multi-action bandits

Topic. Multi-armed bandit.

Category. Research.

Description. In the context of movie recommendation, one can imagine that instead of recommending just one movie to a user, we recommend a bunch of movies, and qualify our recommendation as good if the user clicks on one of the films. This situation can be modeled (simply) by a multi-armed bandit models in which at each round, you have to choose a number, say m, of arms to draw.

In this project, you will try to adapt the algorithms you know for the classical stochastic MAB to this context (UCB, Thompson Sampling). This problem is a particular case of what is sometimes called in the literature a combinatorial bandit problem.

Contact. emilie.kaufmann@inria.fr

1.15 Submodular bandits with $\sqrt{T}$ regret (open problem - difficult)

Topic. Multi-arm bandit.

Category. Theory.

Description. The setting an the motivation are described in the introduction of the following paper:
http://www.satyenkale.com/papers/submodular.pdf
The goal is to propose a better algorithm. One idea is to try to improve the result using self-concordant barriers
http://www-stat.wharton.upenn.edu/~rakhlin/papers/AbeHazRak08.pdf
Another is to attempt to solve it using a modified version of Follow-The-Perturbed-Leader. More details and few proposed ways can be given, the challenge is to provide their analyses.

Contact: michal.valko@inria.fr

1.16 Coffee Bandits

Topic. Multi-arm bandit.

Category. Algorithmic.

Description. Consider a real-life setting in the (contextual) bandit setting where we have a budget constraints of how many times we can pull one arm. For example, that are only 100 small cups in the coffee vending machine until a service guy comes to restock. To be efficient (maximize the profit), we need to learn the distribution of the customers, their taste and our policy should be time-dependent (saving some items for high-gain clients). Parting from http://arxiv.org/abs/1305.2545 (we can provide an idea of of a new algorithm) and the goal is to come up with an efficient solution (approximate knapsack) with regret guarantees.

Contact: michal.valko@inria.fr

1.17 Random Graph Bandits with side information

Topic. Multi-arm bandit.

Category. Algorithmic.

Description. We consider the setting called bandits on graphs with similarity information, which formalizes learning also from the feedback from friends in social networks. The goal is first to investigate and implement the existing solution for Erdos-Renyi graphs and second, come up with a new algorithm for Barabsi-Albert model, which is a more realistic random graph model for social networks.

Contact: michal.valko@inria.fr

1.18 Bandits for Feature Selection

Topic. Multi-arm bandit.

Category. Algorithmic.

Description. Feature selection is a classical problem in ML. We consider a bandit approach to a setting when we have a large (potentially infinite) set of features and we aim to find the best one. This is often the case in biology, when one wants to find a biomarker (a good predictor) of a disease or an adverse condition. Formally, we model this problem as an infinitely many arm bandit where we are interested in an algorithm minimizing simple regret (finding the best or a nearly-best feature).

Contact: michal.valko@inria.fr

1.19 Global optimization of very difficult function

Topic. Multi-arm bandit.

Category. Algorithmic.

Description. We consider efficient and provably optimal search algorithms based on Upper Confidence Trees (https://hal.inria.fr/hal-00747575), which is often used for parameter search in computer games. In this project, we focus on difficult functions with unknown smoothness properties. The goal is to investigate(both theoretically and empirically) several search strategies.

Contact: michal.valko@inria.fr

1.20 Sample Complexity of Lineary Solvable MDPs

Topic. Reinforcement learning.

Category. Theory.

Description. The class of linearly solvable MDPs relies on additional assumptions on the structure of the MDP which corresponds to a significant simplification in the computation of the optimal policy. Although this improvement has been empirically studied, there is no careful sample complexity analysis showing how the complexity of linearly solvable MDPs actually compares to the traditional MDPs and where the advantage actually comes from. The objective of the project is to develop a preliminary sample complexity analysis of batch algorithms for linearly solvable MDPs.

References. Linearly-solvable Markov decision problems

Contact. alessandro.lazaric@inria.fr

1.21 Numerical Comparison of Bandit Algorithms for Best-arm Identification

Topic. Multi-arm bandit.

Category. Implementation.

Description. The objective of the project is to provide a thorough comparison among different best-arm identification algorithms in the settings of fixed budget and fixed confidence. Beside reviewing the current

literature, the student is expected to produce a Matlab code which allows to easily implement and compare additional algorithms. An example of a code which could serve as a basis for this project is available at http://mloss.org/software/view/415/.

Contact: alessandro.lazaric@inria.fr

1.22 Transfer in the Multi-arm Bandit Framework

Topic. Multi-arm bandit.

Category. Implementation.

Description. In many applications, the bandit algorithm is applied on a stream of users which interact for some finite amount of time with the system. This very general scenario can be tackled in many different ways depending on the information and resources available. In this project we consider the case where the setting is modeled as a transfer bandit problem in which the switch between users is known but the identify of the user remains unknown. The objective of the project is to run intensive tests on the algorithm proposed in "Sequential Transfer in Multi-armed Bandit with Finite Set of Models" and compare it with variants that will be developed during the project.

Contact. alessandro.lazaric@inria.fr

1.23 Non-Stationary Multi-arm Bandit

Topic. Multi-arm bandit.

Category. Implementation.

Description. In the bandit literature, two main settings are considered: stochastic or adversarial. While the assumption of perfectly stationary environments of the former is often unrealistic, the worst-case analysis of the latter is too conservative. In this project, we want to study the performance of bandit algorithms in non-stationary environments. The student will be asked to review the papers available in the literature on the topic, implement them, and propose variants to effectively deal with non-stationary problems.

Contact. alessandro.lazaric@inria.fr

1.24 Distributed Bandit

Topic. Multi-arm bandit.

Category. Review.

Description. With the increasing application of MAB strategies in many different domains, new problems and settings arise. One of them is the problem of distributed bandit, where multiple MAB algorithms are (more or less partially) connected and they can exchange data in order to improve their performance. This distributed scenario poses a number of challenges where MAB theory overlaps with routing, multi-agent, and distributed control. In this project, the student should provide an overview of the problem and review a few algorithmic approaches which have been formulated in the literature.

Contact. alessandro.lazaric@inria.fr

1.25 Learning the Max

Topic. Multi-arm bandit.

Category. Implementation+Research.

Description. While standard multi-arm bandit has the objective of pulling as much as possible the optimal arm, in many applications it is critical not only to identify the optimal arm but also to have an accurate estimate of its value, i.e., the maximum expected reward of the problem. This suggests that a quite different exploration strategy should be implemented. The objective of the project is to study the problem and implement the proposed algorithm.

Contact. alessandro.lazaric@inria.fr

1.26 Review of risk-aversion in multi-arm bandit

Topic. Multi-arm bandit.

Category. Review.

Description. In multi-arm bandit the focus is often to pulled as much as possible the arm with the largest expected reward and the performance is measured w.r.t. to the expected regret. Other measures of optimality can be defined. The objective of the project is to review recent advances in the direction of including risk aversion in online learning and multi-arm bandit. The review should mostly cover the settings and the results from the following papers:

- Risk-Aversion in Multi-armed Bandits
- Sample Complexity of Risk-averse Bandit-arm Selection
- Robust Risk-averse Stochastic Multi-Armed Bandits
- Exploration vs Exploitation vs Safety: Risk-Aware Multi-Armed Bandits

Contact. alessandro.lazaric@inria.fr

1.27 Thompson Sampling for Permutation Bandit

Topic. Multi-arm bandit.

Category. Review+Implementation.

Description. In many applications of multi-arm bandit, more than one arm has to selected at the same time (e.g., multi-user channel allocation). This corresponds to the general case of the combinatorial bandit problem. The first objective of the paper is to review the current research available on the topic with particular attention to the permutation bandit case. The second objective is to implement the three algorithms available for the permutation bandit case and compare their performance to a variation of the Thompson sampling algorithm, which is believed to obtain competitive results with a better computational complexity.

References.

- A New UCB-Like Algorithm for Permutation Bandit Problem (pdf under request)

- On the Combinatorial Multi-Armed Bandit Problem with Markovian Rewards
- Combinatorial Bandits
- Combinatorial Multi-Armed Bandit: General Framework, Results and Applications

Contact. alessandro.lazaric@inria.fr

1.28 Review of risk-aversion in MDPs

Topic. Reinforcement learning.

Category. Review.

Description. The standard definition of optimal policy involves the maximization of the expected sum of rewards. Whenever the problem requires some form of risk aversion, maximizing the expected return is no longer desirable. In order to formalize risk-aversion, a large number of notions of risk have been introduced over years. The project should focus on reviewing the notions of risk which are related to a multi-stage problem, such as in MPDs. In particular, the review should focus on the following papers (and references therein if needed)

- Risk-Averse Dynamic Programming for Markov Decision Processes
- Iterated risk measures for risk-sensitive Markov decision processes with discounted cost
- Risk-Aware Decision Making and Dynamic Programming
- An Approximate Solution Method for Large Risk-Averse Markov Decision Processes

Contact. alessandro.lazaric@inria.fr

1.29 Linear Programming for MDPs

Topic. Reinforcement learning.

Category. Review.

Description. Unlike the standard dynamic programming algorithms, the linear programming approach to the solution of MDP is particularly appealing since it targets the computation of the optimal value function in a direct, non-iterative way. The objective of the project is to review the literature about this approach with a particular focus on the empirical and theoretical performance of the LP algorithms.

Contact. alessandro.lazaric@inria.fr

1.30 Avoiding Chattering in Policy Iteration

Topic. Reinforcement learning.

Category. Review.

Description. Approximate policy iteration is known to converge only in a region. This creates practical problems in applications where continuous chattering between different policies can pose serious issues. The objective of the project is to review the algorithms developed so far which try to avoid these effects.

Contact. alessandro.lazaric@inria.fr

2 Internship

Note: Most of the topics covered during the course and available in the project list are active research topics at *SequeL* (INRIA Lille) and open the possibility for internships and Ph.D. programs. If you are interested in these opportunities, please contact A. Lazaric directly. All internships are payed by INRIA Lille and require spending part of the internship period in the group at INRIA Lille.

2.1 Deep recurrent neural networks for man-machine dialog state tracking

Topic. Reinforcement Learning.

Description. Speech is a natural way of interaction between humans and it has been studied for decades as a mean to interact with machines. Speech is not only an acoustic signal, it embeds meanings and allows establishing long-term interactions between speakers. To decide what to say at a given time in a conversation, each participant relies on the current context of the interaction which is based on the history of the dialogue, i.e. the pieces of information exchanged so far in the dialogue.

Automatically inferring the current context of the dialogue for a machine is quite a hard task because the don't recognize speech so well and it's hard to extract meanings from badly recognized speech. Yet, automatic speech recognition (ASR) and natural language understanding (NLU) systems are able to provide uncertainty measures about the result of their process.

During this internship, it is proposed to study the opportunity to use deep and recurrent neural networks to automatically build the context of a dialogue using this uncertainty information as input. Reccurrence will help to deal with the time dependency of the context while the deep architecture will help in dealing with the large number of possible inputs [1,2,3]. This context, also called dialog state, will then be used by a reinforcement learning algorithm to optimize the sequence of decisions the machine will have to perform (e.g. ask a question, provide information, ask for a confirmation) so as to generate an efficient dialog [4,5].

The internship will be funded by the French National Agency for Research (ANR) in collaboration with the Computer Science lab of Avignon (LIA) and the (LAAS) lab (robotics) in Toulouse. Results could be used to participate to the Dialogue State Tracking Challenge [1,2] (DTSC) in 2015 previously organised by the Cambridge University and Microsoft Research. More details at http://www.lifl.fr/~pietquin/internship.html.

Contact: olivier.pietquin@inria.fr

2.2 Online Imitation and Apprenticeship Learning

Topic. Reinforcement Learning.

Description. Reinforcement learning algorithms are designed to optimize the sum of rewards over time. Nonetheless, in a wide range of applications it is very difficult to provide a clear reinforcement signal which could lead to the desired policy. A typical example is the task of "learning how to drive". In this case, it is very easy to provide the learner with a good policy but it is very difficult to define a suitable reward function. Recently, it has been proposed to deal with these problems using an inverse reinforcement learning approach, where the learner has access to examples of good trajectories and the objective is to recover the reward which admits that behavior as optimal. This problem, usually referred to as *apprenticeship learning*, has only been considered in the batch setting, where expert trajectories are provided in advance. In this

research project, we would like to investigate the "online" setting where the learner can ask for "examples" from an expert at any time, while learning from a direct interaction with the environment. The objective is to minimize the number of "calls" to the expert and, at the same time, to maximize the performance (measured in terms of accuracy of the behavior or the reward accumulated over time).

During this internship, the student should review the basic literature on apprenticeship learning and on selective sampling and develop a novel algorithm for selective sampling in MDPs. The internship would then proceed with either the implementation or the theoretical analysis of the proposed method(s).

Contact. alessandro.lazaric@inria.fr

2.3 Transfer in the Multi-arm Bandit Framework

Topic. Multi-arm bandit.

Description. In many applications, the bandit algorithm is applied on a stream of users which interact for some finite amount of time with the system. This very general scenario can be tackled in many different ways depending on the information and resources available. In this internship we consider the case where the setting is modeled as a transfer bandit problem in which the switch between users is known but the identify of the user remains unknown. The objective of the internship is to focus on the contextual bandit problem where the space of functions used to approximate the score of different items is adapted over time using the experience accumulated while solving similar tasks. The internship will mostly focus on the theoretical aspects of the problem, trying to show to which extent a reduction in the regret can be actually showed. Furthermore, a preliminary numerical simulation supporting the theoretical findings will be developed.

Contact. alessandro.lazaric@inria.fr

2.4 Segmentation spectrale et thorie des matrices alatoires

Topic. Other.

Description. La segmentation spectrale est une (famille de) mthode de classification non supervise de donnes (voir [1]). tant donn un ensemble de donnes segmenter, ces mthodes s'appuient sur la dcomposition spectrale de la matrice de similarit de chaque paire de donnes. Cette dcomposition est ralise de manire standard, en utilisant des algorithmes classiques en algbre linaire.

Nanmoins, si l'on se place dans un contexte statistique, chaque donne peut-tre vue comme une ralisation d'un processus stochastique et la matrice de similarit construite partir d'un ensemble de donnes est donc la ralisation d'une matrice alatoire. Dans les approches traditionnelles de calcul du spectre de cette matrice, le caractre alatoire de la matrice de similarit est pour l'instant peu pris en compte.

Les travaux sur la thorie des matrices alatoires, fortement dveloppe depuis une dizaine, nous permettent d'aborder ce problme de manire discipline. Ces travaux proposent des mthodes pour estimer les proprits spectrales du processus alatoire sous-jacent [2].

Il semble donc pertinent d'essayer de runir ces deux points de vue.

Par consquent, nous proposons dans ce stage d'essayer d'appliquer la thorie des matrices alatoires aux algorithmes de segmentation spectrale. Il s'agira de mettre au clair les lments utiles de la thorie des matrices alatoires aux algorithmes existants et d'valuer l'impact pratique de cette thorie sur l'amlioration des performances des algorithmes. Les aspects algorithmiques, en termes de complexit en particulier, ne devront pas tre ngligs.

References:

1. U. von Luxburg, A Tutorial on spectral clustering.
2. G. Anderson, A. Guionnet, O. Zeitouni, An introduction to random matrices, Cambridge University Press, 2010.

Contact. philippe.preux@inria.fr

2.5 Recommender systems as combinatorial bandits

Topic. Multi-arm Bandit.

Description.

Internet services and social networks are more and more linked and attempt to use these connections to deliver more structured recommendations. For example, we no longer want just the best movie to watch but a set of diverse movies that we can watch today. Recommender systems also strive to used the connections (e.g., from a social or a professional network) in order to maximize the chances of being successful. All these setting naturally possess complex actions sets: social networks have graph structure, selecting offers to customers is a bipartite matching, diverse sets of movies recommendation forms a polymatroid, ...

This internship will investigate both theoretical and practical aspects of complex modern recommender system. Departing from fixed preferences of user, we will also study the cases where the preference of users change. We will model a sensational effect in recommendation and news: wherever an important event happens, we should consider recommending it to a large population despite their preferences. During the internship we formally define one or several practically relevant recommendation settings and study it both theoretically and empirically.

References:

1. http://arxiv.org/abs/1306.0811
2. http://jmlr.org/proceedings/papers/v32/valko14.pdf
3. http://arxiv.org/abs/1405.7752
4. http://papers.nips.cc/paper/5542-efficient-learning-by-implicit-exploration-in-bandit-problems-with-side-observations.pdf

Contact. michal.valko@inria.fr

2.6 Adaptation de la dimension du modle pour la recommandation squentielle de produits.

Topic. Multi-arm Bandit.

Description.

Depuis les annes 2000 et suite notamment la comptition organise par Netflix, le problme de la recommandation est trs tudi dans la communaut de l'apprentissage statistique. L'enjeu est de concevoir un algorithme qui recommande chaque utilisateur un ou des produits susceptibles de les intresss.

Aujourd'hui, les approches les plus efficaces formalisent cet objectif comme un problme de compltion de matrice : on s'intresse la matrice M telle que $M_{i,j}$ contient l'apptence de l'utilisateur i pour le produit j ; seules quelques valeurs de cette matrice sont connues ; l'algorithme infre les valeurs inconnues et recommande sur la base de ces valeurs infres.

L'valuation des algorithmes se fait en mode "batch" : partir d'un certains nombre de valeurs connues, on mesure la qualit des valeurs infres. Ce mode d'valuation correspond la situation o un seul produit est recommand et o il n'y a plus dinteractions par la suite avec l'utilisateur. Pourtant dans la pratique ces algorithmes sont utiliss en ligne, on itre successivement les tapes : (i) l'algorithme recommande un produit un utilisateur, (ii) l'utilisateur indique son apptence pour le produit recommand. Dans ce contexte, la thorie dite des "bandits manchots" indique que pour tre optimal, un algorithme doit trouver un compromis entre (1) recommander pour maximiser l'apptence retourne, et (2) recommander pour explorer les gots de l'utilisateur.

Forte de son exprience en bandits manchots, notre quipe travaille depuis deux ans l'extension des algorithmes de bandits manchots au contexte de la recommandation. Ce travaille nous a permis de poser les bases d'un algorithmes de recommandation squentiel. Lors du stage nous explorerons l'adaptation en ligne de la dimension du modle appris.

Ce sujet fait appel des connaissances de base en algbre linaire et en statistiques. Il peut tre abord sous deux angles complmentaires : une approche thorique fixant la dimension du modle de sorte contrler le regret ; une comparaison empirique des divers approches proposes sur des donnes relle volumineuse.

References:

1. www.research.rutgers.edu/ lihong/pub/Li10Contextual.pdf
2. www.hpl.hp.com/personal/Robert.../netflix_aaim08(submitted).pdf
3. http://homes.di.unimi.it/ cesabian/Pubblicazioni/ml-02.pdf
4. http://david.palenica.com/papers/linear-bandit/linear-bandits-NIPS2011-camera-ready.pdf
5. https://hal.inria.fr/hal-01022628/PDF/RR-8563.pdf

Contact. romaric.gaudel@inria.fr, jeremie.mary@inria.fr