Class project proposals - Reinforcement Learning - ENS MVA 2015/2016

Guidelines

Subscribe to a project by directly editing this document on https://goo.gl/2D3T0K

Group size: 1 or 2. Exceptionally 3 with justification why 3 is needed.
Project topics

- students are encouraged to come up with their own proposal
  - once students decide on a more specific topic, they should send (as a private piazza post to the instructor) a short paragraph of the project description in the same form as the project proposals written below
  - the natural requirement for the project is that they will be about learning and graphs

Project format

- 4-6 weeks of work
- final report will have 5-10 pages
- formatting of submission: NIPS format
- projects will be then presented in class

Work flow for taking the projects

- some projects announced and described in class: **17. 11. 2015** (proposals can arrive later)
- DL of assignment for students **1. 12. 2015**
- supervisor writes the proposal here
- students contact the supervisor listed as "Contact" (not necessarily the instructor)
- supervisor writes the names and **emails** of the students in this document

Submission

- DL of submission **12. 1. 2016**
- on the DL for the submission, students send their **project report** to the supervisor listed as "Contact" **and** cc the course instructor (michal.valko@inria.fr)

Project presentation

- project presentations in class: **(TBD)** about 15+5 minutes per project
- time your presentation for **15 minutes** (otherwise your grade will be affected)
- for supervisors that proposed a project and want to participate in presentation we can organize a skype/hangouts call. Similarly for students that are not available to present in person or if some projects do not fit into the schedule.

Class website: http://researchers.lille.inria.fr/~lazaric

Internships: internships are available in the Spring. Please contact me directly to investigate possible topics.

Proposals:

Name: A "Dual"-Based Reward Function for the Game of Go
Topic: RL and games
Category: implementation
Contact: Odalric-Ambrym Maillard, **odalric.maillard@inria.fr**

Assigned to: TBD

Description: The game of Go (https://en.wikipedia.org/wiki/Go_%28game%29) is a typical example of application of Reinforcement Learning for Games. It is still an open challenge to beat world champions at this game, and insights from Reinforcement Learning via the basic Monte Carlo Tree Search significantly improved the level of automatic players a few years ago. The goal of this project is to study a specific construction of reward function (detailed below) for the game of Go, and compare the performance of the resulting strategies obtained by applying standard Reinforcement Learning algorithms to the induced Markov Decision Process built with this reward function, to the performance obtained when considering a standard reward function that only gives 1 or -1 at the very end of the game depending on who is winning. Further info here.

Name: Exploration-exploitation in Reinforcement Learning

Topic: RL

Category: review

Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**

Assigned to: TBD

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.

Name: The Arcade Learning Environment

Topic: RL

Category: implementation

Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**

Assigned to: TBD

Description: The Arcade Learning Environment 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.

Name: Hierarchical Reinforcement Learning

Topic: RL

Category: review, implementation

Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**

Assigned to: TBD

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.

Name: Transfer and Multi-task Reinforcement Learning
Topic: RL
Category: review
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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.

Name: Review of Applications of RL
Topic: RL
Category: review
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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.

Name: Policy Search Algorithms
Topic: RL
Category: review
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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.

Name: Learning the Representation in RL
Topic: RL
Category: review
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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.

Name: Apprenticeship Learning
Topic: RL
Category: review
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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 \textit{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.

Name: Thompson Sampling: influence of the prior distribution
Topic: MAB
Category: research
Contact: Emilie Kaufmann, **emilie.kaufmann@inria.fr**
Assigned to: TBD
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.

[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

Name: Optimistic approaches in Contextual Linear Bandit models
Topic: MAB
Category: review, implementation
Contact: Emilie Kaufmann, **emilie.kaufmann@inria.fr**

Assigned to: TBD
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]).

[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

Name: Thompson Sampling in Contextual Bandits
Topic: MAB
Category: implementation, research
Contact: Emilie Kaufmann, **emilie.kaufmann@inria.fr**
Assigned to: TBD
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.

[1] An emirical evaluation of Thompson Sampling, Chapelle, Li, NIPS 2011
[2] Parametric bandits: the generalized linear case, Filippi, Cappe, 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

Name: Best arm identification versus regret minimization
Topic: MAB
Category: review, implementation
Contact: Emilie Kaufmann, **emilie.kaufmann@inria.fr**
Assigned to: TBD
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.

[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

Name: Stochastic versus adversarial bandits
Topic: MAB
Category: review, implementation
Contact: Emilie Kaufmann, **emilie.kaufmann@inria.fr**
Assigned to: TBD
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.

[1] http://jmlr.org/proceedings/papers/v23/bubeck12b/bubeck12b.pdf
[2] http://jmlr.org/proceedings/papers/v32/seldinb14-supp.pdf

---

Name: Multi-action bandits
Topic: MAB
Category: research
Contact: Emilie Kaufmann, **emilie.kaufmann@inria.fr**
Assigned to: TBD
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.

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

Topic: MAB
Category: research
Contact: Michal Valko, **michal.valko@inria.fr**
Assigned to: TBD
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.

Name: Coffee Bandits
Topic: MAB
Category: algorithmic
Contact: Michal Valko, **michal.valko@inria.fr**
Assigned to: TBD
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.

Name: Global optimization of very difficult function
Topic: MAB
Category: algorithmic
Contact: Michal Valko, **michal.valko@inria.fr**
Assigned to: TBD
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.

Name: Sample Complexity of Lineary Solvable MDPs
Topic: MAB
Category: research
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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.

http://homes.cs.washington.edu/~todorov/papers/TodorovNIPS06.pdf

---

Name: Numerical Comparison of Bandit Algorithms for Best-arm Identification
Topic: MAB
Category: implementation
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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.

Name: Transfer in the Multi-arm Bandit Framework
Topic: MAB
Category: implementation
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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 http://researchers.lille.inria.fr/~lazaric/Webpage/Publications_files/transfer-bandit.pdf and compare it with variants that will be developed during the project.

Name: Non-Stationary Multi-arm Bandit
Topic: MAB
Category: review, implementation
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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.

Name: Distributed Bandit
Topic: MAB
Category: review
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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.

Name: Review of risk-aversion in multi-arm bandit
Topic: MAB
Category: review
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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:

http://arxiv.org/abs/1301.1936
http://faculty.cse.tamu.edu/nikolova/papers/yu-nikolova-ijcai13-full.pdf
http://hal.inria.fr/hal-00821670
http://jmlr.org/proceedings/papers/v29/Galichet13.pdf

Name: Review of risk-aversion in MDPs
Topic: MAB
Category: review
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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)

http://www.optimization-online.org/DB_FILE/2009/12/2497.pdf
http://arxiv.org/abs/1202.3755
https://cs.uwaterloo.ca/~ppoupart/nips08-workshop/accepted-papers/nips08paper03-final.pdf
http://www.auai.org/uai2012/papers/88.pdf

Name: Linear Programming for MDPs
Topic: MAB
Category: review
Contact: Alessandro Lazaric, **alessandro.lazaric@inria.fr**
Assigned to: TBD
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.

Name: xxx
Topic: xxx
Category: xxx
Contact: xxxx, **XXXX@XXXX**
Assigned to: TBD
Description: XXXX.