Master MVA: Reinforcement Learning **Lecture:**

Topics for Projects

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

1 Multi-arm Bandit Projects

1.1 Numerical Comparison of Bandit Algorithms

ASSIGNED: Julien Karadayi and William Pambrun (random UCB) / Coddy Levi (boot UCB).

Topic. Multi-arm bandit.

Category. Implementation (+Research).

Description. By now many different bandit algorithms are available. For most of them a detailed theoretical analysis with regret bounds is available. On the other hand, it is not always simple to have a convincing comparison of their actual empirical performance. While it could be easily assessed case by case on specific problems, it would be interesting to have a more detailed numerical comparison from which to draw general insights about which algorithm is better in which case.

The main objective of the project is to provide an extensive empirical analysis of existing bandit algorithms along different dimensions such as, expected regret, regret distribution, and computational complexity. Furthermore, two novel algorithms should be implemented and tested: random UCB and bootstrap UCB. The suggested steps for the project are (more details under request):

- Download and study the code of bandit algorithms from http://mloss.org/software/view/415/
- Implement random-UCB
- Implement boot-UCB
- Run experiments on different distributions studying the regret, the regret distribution, and the computational complexity.

References:

Contact: alessandro.lazaric@inria.fr

1.2 Extreme Values Bandits

ASSIGNED: Jessica Hoffman

Topic. Multi-arm bandit.

Category. Implementation (+Research).

Description. In many areas of finance, medicine, security, and advertising we aim to allocate limited resources to different channels in order to find some extreme value. In this project we will study efficient ways to allocate these resources sequentially. We consider we have several channels which correspond to different heavy-tailed distributions (such as Pareto) and our goal is to find the most extreme value. We will therefore design an algorithm based on the Hill's estimator of the tail index from which derive an allocation strategy. The goal of the project is to evaluate the efficiency of such strategy. both on the simulated data and on the real-data which were obtained by recording Internet traffic activity from the 270 laptop users. The description of the algorithm (and the implementation if needed) and the traffic data will be provided. The goal of the project is to devise and perform the experimental evaluation.

Contact. michal.valko@inria.fr

1.3 Learning the Max

ASSIGNED: Hugo Magaldi, Pauline Chavallard

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

1.4 Budgeted Bandits for the Coffee machines

ASSIGNED: Alexis Jacq, M. Chiapino

Topic. Multi-arm bandit.

Category. Research.

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 the service guy comes to restock. If we see that we are low on small cups, we can offer a big cup espresso for a discounted price (and in this way improve the regret until the next service).

This is a research project that would start with the formalization of the problem, looking over the relevant literature:

- http://arxiv.org/abs/1305.2545
- http://jmlr.org/proceedings/papers/v25/feraud12/feraud12.pdf
- http://arxiv.org/pdf/1204.1909v1.pdf
- http://research.microsoft.com/en-us/people/taoqin/budgetedmab.pdf,

and proposing a solution. In case of the good progress there is an opportunity to employ the algorithm on a real coffee machine.

Contact. michal.valko@inria.fr

1.5 Transfer, non-stationary bandit, change point detection, average best, contextual bandit: which one is the best?

ASSIGNED: Kevin Vu.

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 fact, it can be formalized as a contextual bandit problem if the identity of the user is available, it can be seen as a non-stationary bandit if users keep changing, it can be formalized as a standard bandit problem where users are just random, or it can be modeled as a transfer bandit problem whenever the switch between users is known. The objective of the project is to review all these approaches studying their assumptions and their range of validity in different applications and to provide a simple numerical comparison between some of them.

Contact. alessandro.lazaric@inria.fr

1.6 Online clustering with bandit information

Topic. Multi-arm bandit.

Category. Research.

Description. The problem of clustering is usually defined in a completely batch setting where data are available before hand. Nonetheless, it is often the case that data should be actively collected by the learning algorithm, with the objective of clustering together different sources of information under a time constraint. The objective of the project is to study the problem and develop a novel bandit algorithm for the solution of the online clustering problem.

Contact. daniil.ryabko@inria.fr, alessandro.lazaric@inria.fr

1.7 Online Submodular Minimization with the Bandit feedback

Topic. Multi-arm bandit.

Category. Research.

Description. The setting an the motivation are described here: http://www.satyenkale.com/papers/submodular.pdf;
The goal of this project is to try to improve the result using self-concordant barriers http://www-stat.wharton.upenn.edu/ r
More details and the 2 proposed algorithms can be given, the challenge is to provide their analyses.

Contact. michal.valko@inria.fr

1.8 Review of risk-aversion in multi-arm bandit

ASSIGNED: Despoina Ioannidou

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

1.9 Large Scale Online and Bandit Learning

ASSIGNED: Julien Bardonnet, Oana Camburu

Topic. Multi-arm bandit.

Category. Review.

Description. The goal of this review is to summarizes techniques, approaches, and application of the large scale bandits. A good starting point can be the following thesis [link].

Contact. michal.valko@inria.fr

1.10 Thompson Sampling for Permutation Bandit

ASSIGNED: Matthieu Labeau

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

1.11 Learning the Gradient of Zero-sum Games

ASSIGNED: Robin Bénesse, Pierre-Alexandre Mattei

Topic. Multi-arm bandit.

Category. Implementation+Theory(+Research).

Description. Let us consider a 2 players zero-sum game, where the reward matrix depend on some parameter α. For any fixed value of α, we know that if both players play a regret-minimization algorithm, then the empirical average of the rewards obtained by player A converges to the value of the game, i.e., the average reward at the Nash equilibrium.

In this project we want to study how the value of the game changes w.r.t. α. In particular, we want to adapt the previous regret-minimisation algorithm so that the empirical average of a quantity computed along a run converges to the gradient of the value. This problem is not trivial since a modification of α implies a change in the corresponding equilibrium probabilities, thus it is crucial to compute the sensitivity of the strategy w.r.t. variations of α. This problem can be of interest when the objective is to "reverse-engineer" a game, i.e., to find the (parameterized) reward function which is actually maximised by greedy agents.

The project will need to:

- write the gradient estimate and prove that it is consistent,
- run simulations in a simple game, with a finite number of actions. Or even a simple poker game where the actions are: either fold or raise by a certain amount which has to be optimized (defined by the parameter α).

Contact. remi.munos@inria.fr

1.12 Resteless Bandit

Topic. Multi-arm bandit.

Category. Theory(+Research).

Description. In the standard multi-arm bandit, the rewards observed from each arm are considered as i.i.d. realizations from stationary distributions. In the restless bandit model, arms are more complicated stochastic processes which evolve over time independently from the learner's actions. In this setting, the objective is not to compete against the best arm but against the best policy, which in general switches between arms pulling the best sequence of arms. The project is about extending current near-optimal algorithms for the Markov case to the more general case of k-order Markov and mixing processes (with known or unknown mixing rate). The objective of the project is to devise an algorithm that has a sublinear regret w.r.t. the best policy of choosing arms. The reference for Markov (of order 1) is http://daniil.ryabko.net/maba.pdf but note that the algorithm and the analysis for higher order Markov processes and mixing processes would be much simpler, since the objective is just to obtain sublinear regret.

Contact. daniil.ryabko@inria.fr

1.13 ???

ASSIGNED: Romain Warlop

Topic. Multi-arm bandit.

Category. ???.

Description. ???

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

1.14 Bandits with the side information

ASSIGNED: Arthur Roullier, Edwin Grappin

Topic. Multi-arm bandit.

Category. Research (+Theory + Implementation).

Description. We consider the setting called bandits on graphs with similarity information that formalizes feedback from ”friends” in the social networks. We will consider a more realistic setting: for example, the graph is weighted and you can get the feedback from the neighbors only with some probability (related to these weights).

Contact. michal.valko@inria.fr

1.15 Recommender Systems

ASSIGNED: Thomas Belhalfaoui, Claire Vernade

Topic. Multi-arm bandit.

Category. Review.

Description. ???

Contact. jeremie.mary@inria.fr

1.16 Numerical Comparison of Bandit Algorithms for Best-arm Identification

ASSIGNED: Nicolas Keriven.

Topic. Multi-arm bandit.

Category. Implementation.

Description.

References:

Contact: alessandro.lazaric@inria.fr

2 Reinforcement Learning Projects

2.1 Exploration-exploitation in RL

ASSIGNED: Ismael Belghiti

Topic. Reinforcement learning.

Category. Implementation+Theory.

Description. The UCRL algorithm solves the problem of exploration-exploitation in RL taking inspiration from the UCB technique developed in multi-arm bandit. While in the multi-arm bandit framework the upper confidence bounds have been refined using more sophisticated techniques than the simple Chernoff-Hoeffding inequality. The objective of the project is two-fold: *(i)* implement UCRL and test its performance on simple grid-world problems with different sizes and diameters, *(ii)* develop more refined concentration inequalities and plug them into the algorithm and check whether any improvement can be achieved.

Contact. alessandro.lazaric@inria.fr

2.2 Semi-Supervised Maxent Inverse RL

ASSIGNED: Sébastien Forestier

Topic. Reinforcement learning.

Category. Implementation.

Description. Start from this article of Ziebart et al; 2008, [link], which describes a maximum entropy approach to IRL. We will consider a setting where we have beside the expert trajectories also an access to a set of other trajectories. These trajectories may have a good performance or not. In order to incorporate them into the learning, we modify the objective function with penalty R, which penalizes reward function that would assign unsmooth rewards to all trajectories we have available: $\bar{\theta}^* = \arg\max_{\bar{\theta}} \Big(P(\bar{\theta}|\tilde{\zeta}) + \gamma R(\bar{\theta})\Big)$, where R can be

$$R_0(\bar{\theta}) = \frac{2}{m(m-1)} \sum_{i>j} s\left(\zeta_i, \zeta_j\right) \left|\bar{\theta}^\top (f_i - f_j)\right|,$$

where s is a user defined objective function. This objective function can be optimized by a modification to the gradient descent algorithm described in the article by Ziebart et al. The goal of this project would be to execute this approach on a interesting task, such as highway car driving[1]. The algorithm can be implemented within the IRL toolkit[2] that already contains the implementation of the domain. Suggested steps 1) studying the approach 2) designing the similarity function for the driving trajectories 3) implementation of the MaxEnt SSIRL algorithm and the experiments 4) possibly improving the similarity or the objective function.

Contact. michal.valko@inria.fr

[1] http://graphics.stanford.edu/projects/gpirl/highway.avi

[2] http://graphics.stanford.edu/projects/gpirl/index.htm

2.3 Sample Complexity of Lineary Solvable MDPs

Topic. Reinforcement learning.

Category. Theory+Research.

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

2.4 Exploration-exploitation in RL with options

ASSIGNED: Oana JEAN-MARIE

Topic. Reinforcement learning.

Category. Review+Implementation.

Description. The framework of options (i.e., macro-actions) is known to provide a (potential) significant improvement in the performance of online RL algorithms. Nonetheless, at the moment there is no completely principled algorithm able to profitable use options and, in particular, it is not clear what is the best way to integrate options in efficient exploration-exploitation algorithms. The objective of the project is to review the option framework and the limited theoretical analysis available for it. Then it should propose a way to integrate options with the UCRL algorithm in a principled way and prove (empirically) the advantage (if any) of carefully designed options.

Contact. alessandro.lazaric@inria.fr

2.5 Review of risk-aversion in MDPs

ASSIGNED: Mahmoud Ezzaki.

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

2.6 Linear Programming for MDPs

ASSIGNED: Emilien Lomet, Lucas Plaetevoet

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

2.7 Inverse Reinforcement Learning

ASSIGNED: Rafael Parpinel Cavina

Topic. Reinforcement learning.

Category. Review.

Description. n this review, the aim is to write down a coherent summary and comparison of different approaches to IRL. The good subset of relevant papers can be found at: http://researchers.lille.inria.fr/ valko/hp/project-irl

Contact. michal.valko@inria.fr

2.8 Avoiding Chattering in Policy Iteration

ASSIGNED [Review]: Monneret Gilles

ASSIGNED [Implementation]: Jean-Baptiste Alayrac and Thomas Moreau

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

2.9 ???

ASSIGNED: Bertrand Rondepierre, Vincent Bodin

Topic. Reinforcement learning.

Category. ???.

Description. ???

Contact. alessandro.lazaric@inria.fr

2.10 Reinforcement Learning for Cross-Channel Marketing

ASSIGNED: Jonathan Tinkeu-Ngatchou

Topic. Reinforcement learning.

Category. Review.

Description. ???

Contact. alessandro.lazaric@inria.fr

2.11 Reinforcement Learning for Sailing

ASSIGNED: Marc Abeille

Topic. Reinforcement learning.

Category. Implementation.

Description. ???

Contact. alessandro.lazaric@inria.fr

2.12 Policy Search for Robotic Juggling

ASSIGNED: Antoine Biard, Vincent Roulet

Topic. Reinforcement learning.

Category. Implementation.

Description. ???

Contact. alessandro.lazaric@inria.fr

2.13 Deep learning of representations for Reinforcement Learning

ASSIGNED: Bertrand Rondepierre, Vincent Bodin

Topic. Reinforcement learning.

Category. Review.

Description. In the recent years, deep learning achieved impressive results in solving high-dimensional problems by first learning a suitable representation for the problem using multiple layers of neural networks. Very few attempts have been made in bringing this approach to the reinforcement learning problem. The objective of this project is to review the basics of deep learning and the few applications to reinforcement learning.

Contact. alessandro.lazaric@inria.fr

2.14 Safe exploration in MDPs

ASSIGNED: Felipe Yanez Lang, Ketan Bacchuwar.

Topic. Reinforcement learning.

Category. Implementation.

Description. Implementation of the paper "Risk Aversion in Markov Decision Processes via Near-Optimal Chernoff Bounds"

Contact. alessandro.lazaric@inria.fr

2.15 Application of ADP to energy-related problems

ASSIGNED: Matthew Trager.

Topic. Reinforcement learning.

Category. Review.

Description.

Contact. alessandro.lazaric@inria.fr