particularly relevant in item discovery scenarios, consists in assuming that the observations start from the ideas considered in [7, 9] for reward maximization. An alternative observation model, same order of magnitude than the overall length of the experiment. To address this issue, one can occurs when the observations are only available after very long delays, which are typically of the models that are of practical interest in the use-cases mentioned above. The first situation of interest negative rewards - i.e., costs associated with the action - instead).

A simple and illustrative example of the long term goals of the project would be to certify the fairness of a black box algorithm where, for instance, the actions consists of categories of individuals with attached attributes and the goal of the learner is to certify, taking into account experimental costs, that the black box algorithm does not discriminate between the different categories of individuals. Another natural application of this approach would be to tackle model drift by checking whether a previously learned optimal policy can still be considered as appropriate after some time of use. From a more practical view point, we will focus in a first phase on relevant scenarios of interest found in two current important use-cases of multi-armed bandit models: online marketing applications, in collaboration with an industrial partner, as well as intensional web data management [5], relying on the expertise of the VALDA team.

Main activities

A natural starting point of this PhD will be to consider recent works on best-arm identification [4,6].

In this PhD we will consider more specifically contexts in which the actions that can be performed, are structured - consisting, for instance, of pairs of attributes and values - and the focus is put on reaching, as fast as possible, a statistically certified complex decision. Where by "complex", we mean a decision that does not reduce to the usual choice of a single best action. We are also interested in mixed scenarios where one needs to reach decisions with some statistical guarantees while controlling the regret with respect to the optimal sequence of cases, given some reward function.

In this PhD we will consider more specifically contexts in which the actions that can be performed, are structured - consisting, for instance, of pairs of attributes and values - and the focus is put on reaching, as fast as possible, a statistically certified complex decision. Where by "complex", we mean a decision that does not reduce to the usual choice of a single best action. We are also interested in mixed scenarios where one needs to reach decisions with some statistical guarantees while controlling the regret with respect to the optimal sequence of cases, given some reward function.

A simple and illustrative example of the long term goals of the project would be to certify the fairness of a black box algorithm where, for instance, the actions consists of categories of individuals with attached attributes and the goal of the learner is to certify, taking into account experimental costs, that the black box algorithm does not discriminate between the different categories of individuals. Another natural application of this approach would be to tackle model drift by checking whether a previously learned optimal policy can still be considered as appropriate after some time of use. From a more practical view point, we will focus in a first phase on relevant scenarios of interest found in two current important use-cases of multi-armed bandit models: online marketing applications, in collaboration with an industrial partner, as well as intensional web data management [5], relying on the expertise of the VALDA team.

Specific additional challenges that need to be addressed are listed below.

Structured actions This is a topic that has been well-studied recently in the context of reward maximization [7,10] but that needs to be revisited in the decision making context considered here.

Taking into account action reward/cost Traditional best-arm identification ignores the reward of actions (focusing only on the time needed to reach a decision) while it is well-known that, conversely, reward-maximization policies are not well suited for decision-making. Here, we are interested in developing novel strategies for mixed objectives in which one aims to reach statistically certified decisions while also taking into account the accumulated reward (note that to keep up with common terminology, we refer to rewards but in some scenarios it may be relevant to consider negative rewards - i.e., costs associated with the action - instead).

Delayed and diminishing observations During the PhD, we will also consider alternative feedback models that are of practical interest in the use-cases mentioned above. The first situation of interest occurs when the observations are only available after very long delays, which are typically of the same order of magnitude than the overall length of the experiment. To address this issue, one can start from the ideas considered in [7, 9] for reward maximization. An alternative observation model, particularly relevant in item discovery scenarios, consists in assuming that the observations
correspond to the findings of unique (rare) items, yielding a particular form of non-stationary problem, with diminishing returns, for which one can consider adapting the approach followed in [3, 8].

References


Benefits package
- Subsidised catering service
- Partially-reimbursed public transport

Remuneration
Gross Salary per month: 1 982 € the first 2 years and 2 085 € the last year