# PhD Thesis proposition

## Non-stationary and robust Reinforcement Learning methodologies for surveillance applications

**Supervisors**:

- *Stefano Fortunati* EC IPSA Paris/L2S (70 % of supervision),
- *Alexandre Renaux* MCF (HDR) Université Paris-Saclay/L2S (Directeur de thèse, 30 % of supervision).

**Ecole Doctorale de rattachement**: Sciences et technologies de l'information et de la communication (STIC) at Université Paris-Saclay.

**Abstract**

*EG:* One of the underlying assumptions of Reinforcement Learning (RL) methodologies is the stationarity of the environment embedding the agent. Specifically, the three main elements characterizing a Markov Decision Process (MDP), i.e. the set of states, the set of actions and the reward function are assumed to be constant/invariant over time. In surveillance applications however, such assumption is generally unrealistic since the environment (i.e. the area that the agent has to monitor) is constantly changing. The amount and types of objects or different disturbance statistics are just two examples of non-stationarity. The main aim of this project is then the development of original RL schemes able to cope with time-dependent MDP. This challenging goal may bring significant benefits in many theoretical and applicative AI-subfield: from the statistical learning theory, non-stationary random processes and sets to Signal Processing applications. The theoretical findings will be validated in an emerging crucial issue: the detections of drones using massive antenna arrays.

*FR:* Une des hypothèses de l'apprentissage par renforcement est la supposée stationnarité de l'environnement dans lequel évolue l'agent. En effet, les trois éléments qui caractérisent le processus de décision markovien (PDM), c'est-à-dire l'ensemble des états, l'ensemble des actions et la récompense sont supposés constants au cours du temps. Cependant, pour des applications de surveillance, une telle hypothèse est peu réaliste puisque l'environnement (c'est-à-dire la zone surveillée par l'agent) est en perpétuelle évolution. Le nombre et le type d'objets à surveiller ainsi que les perturbations statistiques de l'environnement sont deux exemples de non-stationarité. L'objectif principal de ce projet consistera en le développement de méthodes

originales d'apprentissage par renforcement avec la prise en compte de PDM évoluant au cours du temps. De telles méthodes auront des retombées dans des domaines théoriques et applicatifs de l'intelligence artificielle : apprentissage statistique, processus et ensembles aléatoires non stationnaires et traitement statistique du signal. Notre application principale concernera un problème qui est désormais d'une grande importance : la détection de drones à l'aide d'un grand réseau multi-antenne.

**Keywords**: Markov decision process, Reinforcement Learning (RL), Multi-agent RL, Non-Stationary Stochastic Learning, Robust Statistics.

# 1 Detailed description of the PhD project

## 1.1 Scientific context

Reinforcement Learning (RL) methodologies are currently adopted in different context requiring sequential decision-making tasks under uncertainty [1]. The RL paradigm is based on the perception-action cycle, characterized by the presence of an agent that senses and explores the unknown environment, tracks the evolution of the system state and intelligently adapts its behavior in order to fulfill a specific mission. This is accomplished through a sequence of actions aiming at optimizing a pre-assigned performance metric (reward). There are countless applications that can benefit from this perception-action cycle (traffic signal control, robots interactions the physical objects, just to cite a few), each of which is characterized by a peculiar definition of "uncertainty" or "unknown environment". A more precise definition of this uncertainty strongly depends on the particular domain considered. However, there is at least one crucial assumption underlying the majority of classical RL algorithms: the stationarity of the environment, i.e. the statistical and physical characterization of the scenario, is assumed to be time-invariant. This is clearly a quite restrictive limitation in many real-world RL applications, where the agent is usually embedded in a changing scenario whose both statistical and physical characterization may evolve over time. Due to the crucial importance of including the non-stationarity in the RL framework, both theoretical and application-oriented non-stationary approaches have been proposed recently in the RL literature (e.g. [2,3]). Among the numerous potential applications, in this project we will focus on the problem of Cognitive Radar (CR) detection in unknown and non-stationary environment. Specifically, building upon the previous works [4–6], we will aim at proposing an RL based algorithm for cognitive multi-target detection in the presence of unknown, non-stationary disturbance statistics. The radar acts as an agent that continuously senses the unknown environment (i.e., targets and disturbance) and consequently optimizes transmitted waveforms in order to maximize the probability of detection (PD) by focusing the energy in specific range-angle cells.

## 1.2 Scientific goals

The scientific goal of the proposed PhD thesis is twofold. Firstly, the PhD candidate will get familiar and develop original RL-based algorithms for non-stationary environments. These theoretical outcomes will be then applied to a specific scenario of great interest nowadays: the radar detection of drones. More specifically, the PhD thesis will be structured in two steps:

1. *Theoretical foundation of non-stationary RL algorithms*
   The aim of this first step is to develop an original theoretical foundation of non-stationary

Markov Decision Processes (MDP) [2]. In particular, the candidate will investigate the possibility to generalize classical RL methodologies to MDP characterized by a time-varying sets of states, actions and reward functions. This non-stationary generalization is of crucial importance for a wide variety of applications and it is an almost unexplored research field.

2. *Surveillance applications and drone detection*
The theoretical results obtained in the first part of the PhD thesis will then be used to derive and implement new algorithms for drones detection and tracking using radar systems [4–6]. Specifically, we will consider a co-located Multiple-Input-Multiple-Output (MIMO) radar with a large ("massive") number of transmitters and receivers. It has been shown, in fact, that this massive MIMO configuration allows one to dispense with unrealistic assumptions about the a-priori knowledge of the statistical model of the disturbance [4].

### 1.3 Expected impact

The generalization to non-stationary environment of the actual RL algorithm are of central importance for many real-world applications. Intelligent traffic control and pilotage of robots and drones for rescue missions are two classical examples of non-stationary environments that may cause a violation of the basic assumptions underlying the actual RL methodologies. The first desired expected outcome of this project is to provide a rigorous mathematical framework general enough to be able to guarantee the possibility to develop, by independent researchers, new and original non-stationary RL algorithms that may be used in different fields. Moreover, the specific application proposed in the project, i.e. non-stationary RL algorithm for target detection, can potentially represent a breakthrough in the radar signal processing community. In fact, such an algorithm will be able to provide optimal performance (in term of probability of false alarm and probability of detection) without the need of any a-priori information on the disturbance statistical model, unlike the classical likelihood ratio-based detection schemes commonly used since 40 years in radar applications.

### 1.4 Organization of the thesis work

The three-years research program will be organized as follows:

1. *First step (4 months).* The first part of the research project will be dedicated to a review of the existing literature on stationary RL methodologies in order to formalize the mathematical framework underlying classical stationary learning methodologies. Specifically, the attention will be focused on the formal definition of Markov Decision Process and on the related concepts of set of actions, set of states and reward function [2,7] from a statistical point of view. This will allow us to better clarify how the non-stationarity may be introduced in the main plot.

2. *Second step (6 months).* After an in-depth analysis of the state-of-the-art on this topics, we will formally define what we will consider as non-stationary environments. Having in mind the application to target detection that will be developed later, we will characterize the non-stationarity with both physical abrupt changes in the scenario (e.g. objects appearing and disappearing from the field of view) and statistical evolution of the environmental data (e.g. changes in the data time correlation from a learning cycle to the next one).

3. *Third step (10 months).* After having posed the mathematical foundation of the main framework, we will move to the core of the proposed project: the development of learning

methodologies for non-stationary environments. Original RL strategy will be proposed and their effectiveness and robustness to non-stationarity assessed through extensive simulation results.

4. *Fourth step (10 months).* The theoretical outcomes obtained in the previous steps will be exploited to derive and implement original RL-based detection algorithm for surveillance applications. Specifically, following [5,6], the abstract notions of agent, actions, states and reward will be specialized in the radar detection framework. Moreover, the non-stationarity of the environment assumes here a precise characterization: the number of targets/sources to be detected changes over time along with the statistical characterization of the disturbance.

5. *Fifth step (6 months).* Finally, the obtained results will be applied to a specific scenario of great interest nowadays: the detection of drones. Due to its highly non-stationarity and physical variability, this is a challenging problem that cannot be addressed with classical, time-invariant, RL algorithms.

### 1.5 Expected results and perspectives in research and applications

The aim of this PhD project is twofold. From the theoretical research side, we aim at developing advanced RL methodologies able to handle the non-stationarity of the environment to be explored by the agent. Since the vast majority of real-world scenario is affected by some sort of non-stationarity, this original line of research may pave the way to countless practical exploitations (intelligent traffic control, pilotage of robots in harsh conditions, and so on) as well as further theoretical investigation (statistical optimality condition under high non-stationarity, data-driven selection of the hyper-parameters used in the RL algorithm). From the application side, the derivation of a cognitive detection algorithm for radars may be of interest to integrate with new functionalities existing surveillance systems. As research products, during the duration of the PhD scholarship, we plan to publish at least one journal paper and two conference papers in Machine Learning (ML) and related field where our theoretical findings will be discussed and presented to the ML community. Moreover, the original application of the theoretical outcomes to the detection problem will be disseminated in the Signal Processing and Radar community through another journal paper and three additional conferences.

### 1.6 Hosting laboratory

The PhD thesis will be developed at the L2S Laboratory (Laboratoire des signaux et systèmes, UMR8506) where the candidate will join the Modeling and Estimation Group (GME) in the Signals and Statistics group. An office and a PC will be made available to the candidate. It is important to underline that this PhD thesis are fully in line with an active international collaboration already established among the L2S (France), the University of Pisa (Italy) and the Ruhr-University Bochum (Germany). Depending on the availability of funds, the candidate will have the opportunity to visit these two institutions for short research periods (from 2 up to 6 months).

## References


[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. The MIT Press, second ed., 2018.

[2] E. Lecarpentier and E. Rachelson, "Non-stationary markov decision processes, a worst-case approach using model-based reinforcement learning," *Advances in neural information processing systems*, vol. 32, 2019.

[3] S. Padakandla, K. J. Prabuchandran, and S. Bhatnagar, "Reinforcement learning algorithm for non-stationary environments," *Applied Intelligence*, vol. 50, p. 3590?3606, 2020.

[4] S. Fortunati, L. Sanguinetti, F. Gini, M. S. Greco, and B. Himed, "Massive MIMO radar for target detection," *IEEE Transactions on Signal Processing*, vol. 68, pp. 859–871, 2020.

[5] A. M. Ahmed, A. A. Ahmad, S. Fortunati, A. Sezgin, M. S. Greco, and F. Gini, "A reinforcement learning based approach for multitarget detection in massive MIMO radar," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 57, no. 5, pp. 2622–2636, 2021.

[6] F. Lisi, S. Fortunati, , M. S. Greco, and F. Gini, "Enhancement of a state-of-the-art RL-based detection algorithm for Massive MIMO radars," *IEEE Transactions on Aerospace and Electronic Systems (accepted)*, 2022.

[7] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.