# P-MARL: Prediction-Based Multi-Agent Reinforcement Learning for Non-Stationary Environments

# (Extended Abstract)

Andrei Marinescu, Ivana Dusparic, Adam Taylor, Vinny Cahill, Siobhán Clarke
Distributed Systems Group, School of Computer Science and Statistics
Trinity College Dublin, Ireland
{marinesa, ivana.dusparic, tayloral, vinny.cahill, siobhan.clarke}@scss.tcd.ie

## ABSTRACT

Multi-Agent Reinforcement Learning (MARL) is a widely-used technique for optimization in decentralised control problems, addressing complex challenges when several agents change actions simultaneously and without collaboration. Such challenges are exacerbated when the environment in which the agents learn is inherently non-stationary, as agents' actions are then non-deterministic.

In this paper, we show that advance knowledge of environment behaviour through prediction significantly improves agents' performance in converging to near-optimal control solutions. We propose P-MARL, a MARL approach which employs a prediction mechanism to obtain such advance knowledge, which is then used to improve agents' learning. The underlying non-stationary behaviour of the environment is modelled as a time-series and prediction is based on historic data and key environment variables. This provides information regarding potential upcoming changes in the environment, which is a key influencer in agents' decision-making.

We evaluate P-MARL in a smart grid scenario and show that a 92% Pareto efficient solution can be achieved in an electric vehicle charging problem, where energy demand across a community of households is inherently non-stationary. Finally, we analyse the effects of environment prediction accuracy on the performance of our approach.

## Categories and Subject Descriptors

I.2.11 [**Distributed Artificial Intelligence**]: Multiagent systems

## Keywords

Multi-Agent Systems; Reinforcement Learning; Environment Prediction; Smart Grids

## 1. INTRODUCTION

Multi-Agent Reinforcement Learning (MARL) is being increasingly used in various domains such as computer networks, vehicular traffic, resource management, robotic teams and distributed control in general [1]. Many of these situations pose complex challenges to multi-agent systems due to the dynamicity of the environment, even more when the environment itself is characterised by non-stationary behaviour. Adding to the complexity in such circumstances is the situation where one might not only encounter dynamicity generated by stochastic interactions between agents; another level of stochasticity may occur independently of agent actions due to a continuously changing environment.

**Appears in:** *Proceedings of the 14th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2015), Bordini, Elkind, Weiss, Yolum (eds.), May, 4–8, 2015, Istanbul, Turkey.*


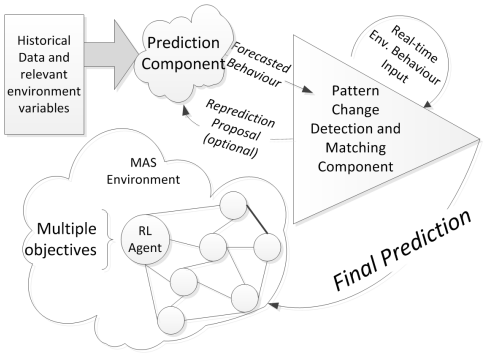


**Figure 1: P-MARL Algorithm Architecture**

The latter problem impacts on the convergence of a MARL, as agents encounter new situations all the time, for which they were not prepared in the exploration stages. There are previous proposals to deal with such non-stationary environments through partial models of the environment and context detection [2, 3], but these do not actually address the problem of handling environments where the number of environment states is infinite, thus not requiring a fixed number of models.

We propose to tackle the problems of such highly dynamic environments through an active prediction module. Our hypothesis is that prediction of future environment behaviour provides agents with a sufficiently good *a priori* training model in order to improve upon their performance in real-time. In this paper we propose Predictive-MARL (P-MARL), an online MARL augmented with environment prediction and pattern change detection capabilities for complex non-stationary environments.

## 2. P-MARL

P-MARL's architecture comprises three key components, illustrated in Fig. 1. Firstly, a **prediction model** component considers recent historic values and key environment variables that are correlated with the environment's historic behaviour in order to provide an estimate of future behaviour. Our particular model is a hybrid solution, and as such takes advantage of several techniques' strengths for time-series prediction [6]. Secondly, a **pattern change detection and matching** component detects when the prediction model fails in providing reasonable estimations of the

future state of the environment. The current behaviour of the environment is continuously evaluated. If the behaviour is classified as anomalous, adjustments need to made be since the prediction can become inaccurate. This triggers a new estimate which is particular to the anomalous class [5]. Finally, there is a **multi-agent system** (MAS) component which is based on reinforcement learning (RL). This employs the previous components as an input in order to improve its performance in non-stationary environments. The RL agents are implemented as a multi-objective W-Learning processes [4], where each objective is implemented separately as an independent Q-Learning process. Q-Values are obtained for each state-action pair. At every time-step an action is nominated based on these values. Through W-Learning, the winning action is selected based on the importance of all objectives.

From the first two components, an estimate of the environment's future expected behaviour is provided. The agents evaluate the future behaviour and attempt to optimally reach their goals with respect to the imposed estimate. A process of exploration-exploitation is performed by the agents based on the provided environment estimate. This helps them learn the best behaviour for the expected states of the environment. Once agents reach a near-optimal solution, they are ready to switch to online mode. Even though the actual environment they will face will differ, the previously obtained knowledge will help them perform well as conditions in the environment are similar to the the ones in the estimate.

## 3. EXPERIMENTAL STUDY

We apply P-MARL to a non-stationary environment, a real-world scenario occurring in the Smart Grid. The state of the environment is characterised by energy consumption, which introduces a certain amount of randomness due to the behaviour of human users. The environment can be represented as a time-series, which exhibits non-stationary characteristics. We consider a neighbourhood of 230 residential users which contains a set of 90 EVs, each controlled by an intelligent agent; the task of each EV agent is to achieve a desired battery charge for the next day's trip. Additionally, this charging process might be constrained by periods of high demand which occur in particular during the evening, when charging is to be avoided. Three charging algorithms are evaluated in this scenario: a benchmark *Centralised* solution, which computes an optimal charging scheme for each EV given an initial environment estimate; a *Night Tariff-Aware Greedy* solution - which charges the EVs as soon as possible starting from 23:00, by adjusting to a night-saver time tariff; and the *P-MARL* solution based on decentralised control through intelligent agents.

The experiments are run over 3 different sub-cases involving an anomalous day: assuming simple prediction of demand (less accurate), reprediction of demand (more accurate), and finally assuming perfect prediction of the day's demand as input to MARL.

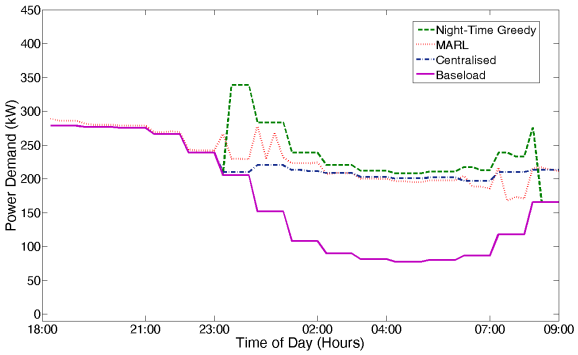


**Figure 2: Algorithm Performance in Reprediction Case**

**Table 1: Comparison of Performance**

| Method | P-MARL | N. Greedy | Centralised |
|---|---|---|---|
| **Perfect Pred.** | 92.4% | 83.1% | 100% |
| **Repredicted** | 92.2% | 83.5% | 97.6% |
| **Simple Pred.** | 89.6% | 83.8% | 97.9% |

The reprediction sub-case is presented in Fig. 2. The centralised solution attempts to evenly schedule the EV's demand over the available hours (18:00-9:00) based on a forecast of the baseload. The resulting load shows some irregularities due to errors in forecasting. For the greedy solution all vehicles start charging at 23:00 resulting in a peak in demand higher than the evening peak, an undesired situation. As for P-MARL, once the base demand is low enough several EV agents start charging, but since this leads to high demand a few of them immediately back off. Once the base demand lowers, more EVs start to charge.

The performance of the algorithms in terms of Pareto optimality is summarised in Table 1. The best performance is for the two methods relying on prediction (P-MARL and Centralised), and is achieved assuming perfect prediction of the future power demand. The increase in forecasting errors comes with a price, as P-MARL performance is brought down to 89.6% Pareto optimality in the case of simple prediction. It is worth noting that having perfect prediction improves the MARL performance only by 0.4% compared to the accurate reprediction.

## 4. CONCLUSIONS

This paper presents P-MARL, a prediction-based solution which improves MARL performance in non-stationary environments. P-MARL's performance is closely related to the ability to accurately forecast future states of the environment. The loss in accuracy results in diminished performance, therefore good environment prediction mechanisms are essential in achieving efficient solutions.

## Acknowledgment

This work was ssupported, in part, by Science Foundation Ireland grant 10/CE/I1855.

## 5. REFERENCES

[1] L. Busoniu, R. Babuska, and B. De Schutter. A comprehensive survey of multiagent reinforcement learning. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 38(2):156–172, 2008.

[2] B. C. Da Silva, E. W. Basso, A. L. Bazzan, and P. M. Engel. Dealing with non-stationary environments using context detection. In *ICML*, pages 217–224. ACM, 2006.

[3] K. Doya, K. Samejima, K.-i. Katagiri, and M. Kawato. Multiple model-based reinforcement learning. *Neural computation*, 14(6):1347–1369, 2002.

[4] M. Humphrys. W-learning: Competition among selfish q-learners. *Computer Laboratory Technical Report*, 362, 1995.

[5] A. Marinescu, I. Dusparic, C. Harris, V. Cahill, and S. Clarke. A dynamic forecasting method for small scale residential electrical demand. In *IJCNN*, pages 3767–3774, July 2014.

[6] A. Marinescu, C. Harris, I. Dusparic, V. Cahill, and S. Clarke. A hybrid approach to very small scale electrical demand forecasting. In *ISGT, 2014 IEEE PES*, pages 1–5, Feb 2014.