Extended Abstract: Reinforcement Learning for the Beer Distribution Game

Maria Alcala-Durand[*†]
Instituto Tecnologico Autonomo de Mexico

Abstract

This work presents a Reinforcement Learning approach for the Beer Distribution Game. Furthermore, assumptions are removed to follow trends that are real world-like, as an example of succesfully applying RL to supply chain problems.

1 Introduction

Reinforcement Learning algorithms can help solve problems that are prone to variance due to human risk aversion, such as the Beer Distribution Game, which is a famous academic supply chain problem. Furthermore, while it is both impossible and impractical to model reality to the slightest detail, it is important to keep realistic assumptions in order to ensure model applicability.

The Beer Distribution Game (BDG) is based on a $4-$level echelon chain, with the ultimate goal of fulfilling a consumer's beer demand. Each period, each echelon demands beer units to the next echelon. However, humans are notoriously inefficient in managing uncertainty and rarely find the optimal solution. The causing underlying effect is called "bullwhip effect", since the variance on the information received by each echelon is higher the farther away it is from the origin (the consumer).

Reinforcement Learning (RL) is a subset of Machine Learning models based in psychology: an *agent* is looking to obtain *reward*. The agent will look for a collection of actions, a *policy*, that maximizes its total reward. On each experiment iteration, it will try to learn the *optimal policy*. In-depth definitions can be found in Sutton [20].

2 Methodology

In order to establish the BDG as a RL problem, the actors were defined as agents and split into two types: learning agents (those who have to make decisions that will affect their overall income, which are factory, regional warehouse, wholesale and retail) and non-learning agents (those who have set trends and don't make decisions, which are consumer and fields). Each learning agent's utility function, for them to attempt to maximize individually, was defined as their daily income, which is comprised by their sales, purchases, inventory costs and backlog penalty.

RL alorithm: Policy Iteration As defined in Sutton [20], policy iteration is a type of on-policy RL algorithm. Each iteration strives to find a better policy π, for each state $s \in S$:

$$\pi(s) \leftarrow \max_a \sum_{s' \in S} T(s, a, s')[R(s, a, s') + \gamma V(s')]$$

Until π converges. During each step, each of the agents explore policies in a *greedy* way. For this problem, each agents' policy is a vector of length 365, and it contains its daily beer demand towards the next echelon.

[*] under the aegis of Dr. Adolfo de Unanue Tiscareno
[†] mfadurand@gmail.com

Challenged assumptions The original statement of the BDG has two assumptions deemed unrealistic for real world applications. First, the customer demand is set at a constant level throughout time. A realistic demand was modeled after a study made on USA beer consumers, where beer consumption is higher towards the weekend and during the holidays. An example month with four weeks can be seen in Figure 1. Second, the fields are able to supply as much material as the factory needs, at any point in time. These were modeled after the yearly harvest of barley in the USA. This means that availability of raw materials for the factory is limited to the summer months, with the main peak happening in August, as can bee seen in Figure 2.

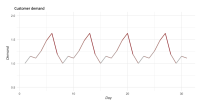

Figure 1: Consumer demand

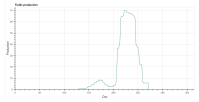

Figure 2: Fields production

3 Results

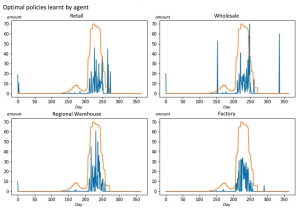

Figure 3: Policies learnt by agent

The policy iteration method found optimal policies in $10,000$ iterations, which took around 5 minutes to train. The sufficiency of these policies was established contrasting stability with training scenarios between $100,000$ and $1,000,000$ iterations. Figure 3 shows, in blue, the learnt policy for each of the 4 agents. An orange line representing the fields supply was added to show that the agents not only learn to follow material availability, but also to overstock during harvest season in order to cover demand during the rest of the year.

Figure 3 shows, for each one of the four learning agents, three scenarios that were built to understand the overall and individual impact of different strategies. 250 simulations were run for each scenario, initializing the world parameters randomly and adding realizations of a gaussian random variable to both the consumer and the fields trends.

The blue line shows the baseline with no randomness, which is a simple $5-$day rolling average demand every day. The red distribution shows the results, normalized by the baseline, of all the agents following the "smart" policy (the one learnt with the RL model). In the green scenario, the relevant agent follows the smart policy and all the others follow the baseline. Finally, in the orange scenario, another agent follows the smart policy, and all the others (including the relevant one) follow the baseline policy. The main results are:

- If all the agents follow their smart policies in parallel, their results are consistently above the baseline
- For each agent, following its smart policy overperforms following the baseline policy, when the other agents follow the baseline
- The agents towards the middle of the chain have different distributions than the ones on the edges, most likely due to them being affected by the double bullwhip effect rooted in them being removed from both information sources simultaneously, while the edge agents have direct contact with at least one

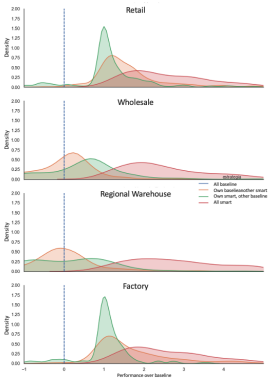

Figure 4: Intelligence scenarios

4 Conclusion

The policy iteration method found actionable, good polices quickly enough for it to be flexible when faced with daily demand changes. Furthermore, the optimal policy proved better than the baseline policy for each agent, regardless of which actions the rest of the agents take. Finally, the double bullwhip effect caused by two sources of information showed that the middle echelons need better forecasting models and have higher stakes for choosing smart strategies.

References

[1] Alexander, J.A. & Mozer, M.C. (1995) Template-based algorithms for connectionist rule extraction. In G. Tesauro, D.S. Touretzky and T.K. Leen (eds.), *Advances in Neural Information Processing Systems 7*, pp. 609–616. Cambridge, MA: MIT Press.

[2] Bower, J.M. & Beeman, D. (1995) *The Book of GENESIS: Exploring Realistic Neural Models with the GEneral NEural SImulation System.* New York: TELOS/Springer–Verlag.

[3] Hasselmo, M.E., Schnell, E. & Barkai, E. (1995) Dynamics of learning and recall at excitatory recurrent synapses and cholinergic modulation in rat hippocampal region CA3. *Journal of Neuroscience* **15**(7):5249-5262.

[1] Alloway, T. (2021) What pandemic puppies can tell us about supply shortages *Bloomberg news*

[2] Bloembergen, D. & Hennes, D. (2013) Fundamentals of multi-agent reinforcement learning *AAMAS Conference Proceedings: MARL (multi-agent reinforcement learning)*

[3] Busoniu, L. & Babuska, R. & De Schutter, B. (2010) Multi-agent reinforcement learning: an overview. Studies in Computational Intelligence **310**:183–221

[4] Chaharsooghi, S. K. & Heydari, J. & Zegordi, S. H. (2008) A reinforcement learning model for supply chain ordering management: An application to the beer game *Decision Support Systems* **45**:949–959

[5] Conneely, B. & Duggan, J. & Lyons, G. (2019) A distributed system dynamics-based frame- work for modelling virtual organisations *National University of Ireland internal publication*

[6] Dizikes, P. (2012) The secrets of the system *MIT News*

[7] Frazzoli, E. (Fall 2010) Principles of Autonomy and Decision Making *Massachusetts Institute of Technology: MIT OpenCouseWare* 16. 410
/ 16. 413, Publisher

[8] Ganapathy, V. (2015) Jay forrester, discoverer of the bullwhip effect *Bookbon*

[9] Glatzel, C. & Helmcke, S. & Wine, J. (2009) Building a flexible supply chain for uncertain times *McKinsey Business Functions*

[10] Grasl, O. (2015) Understanding the beer game using system thinking to improve game result *Trasentis consulting Play the Beer Game series*

[11] Hu, J. (2016) Reinforcement learning explained *O'Reilly*

[12] Jacobs, F. & Chase, R. (2011) Administracion de operaciones. Produccion y cadena de suministros *McGraw Hill-Education*

[13] Kimbrough, S. & Wu, D. & Zhong , F. (2002) Computers play the beer game: can artificial agents manage supply chains? *Decision Support Systems* **33**:323–333

[14] United States Department of Agriculture (2010) Field crops usual planting and harvesting dates *Agricultural Handbook* **628**:6–7

[15] Saad, L. & Author, B. (2014) Beer is americans adult beverage of choice this year *Gallup News*

[16] Shiflet, A. & Shiflet, G. (2006) Introduction to Computational Science *Princeton University Press*

[17] Sterman, J. & Author, B. (2000) Business Dynamics *Irwin/McGraw-Hill* pp. 100, Publisher

[18] Sterman, J. & Author, B. (1989) Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment *Management Science* **35**:321–339

[19] Strozzi, F. & Bosch, J. & Zaldivar, J. (2007) Beer game order policy optimization under changing customer demand *Decision Support Systems* **42**:2153–2163

[20] Sutton, R. & Barto, A. (1998) Reinforcement learning: an introduction *The MIT Press*

[21] Zarandi, M. & Anssari, M. & Avazbeigi, M. & Mohaghar, A. (2011) A multi- agent model for supply chain ordering management: An application to the beer game *Supply Chain Management* 433–44