The Title of your Report

Anonymous Authors

***Abstract*—A short summary of your project. You should change also the title, but do *not* enter any author names or anything that unnecessarily identifies any of the authors. It is suggested you use a similar structure (sections, etc.) as demonstrated in this document, but you can make the section headings more descriptive if you wish. Of course *you should delete all the text in this template and write your own*! – this text simply provides detailed instructions/hints on how to proceed.**

I. Introduction

Describe what you did. Provide access to your anonymized code[1].

Note that results should be reproducible using the technologies from the labs (i.e., Python, and selecting among Scikit-Learn, OpenAI Gym, TensorFlow, PyGame, ...).

Do not change the formatting (columns, margins, etc). Hint: shared tools like `http://sharelatex.com/` and `http://overleaf.com/` are great tools for collaborating on a multi-author report in latex. If you wish to use Word, base it on the IEEE template[2] and convert to `pdf` for submission.

II. Background and Related Work

Elaborate (in your own words) the background material required to understand your work. It should cover a subset of the topics touched upon in the course. You are encouraged to cite topics in lectures, e.g., structured output prediction in [2], book chapters, e.g., Chapter 9 from [1], or articles from the literature, e.g., [3], [4]. Basically, you should prepare the reader to understand what you are about to present in the following sections. Eq. (1) shows a random equation.

$$\hat{\mathbf{y}} = \underset{\mathbf{y}\in\{0,1\}}{\operatorname{argmax}}\, p(\mathbf{y}|\mathbf{x}) \tag{1}$$

III. The Environment

In this report we work with a supply-chain optimization environment over an infinite amount of periods. The environment consists of one factory and up to $k \in \mathbb{N}$ warehouses where in each period it needs to be decided how many units of a product (e.g. butter) should be produced and how many should be shipped to the individual warehouses. A representation of a network with 5 warehouses is depicted in figure III. The state space consists of $k+1$ elements describing the current stock level of the factory / the warehouses where each stock level is limited by some $c_j \in \mathbb{N}$. Furthermore we include an environment process D describing the (stochastic) demand d_j at each warehouse $j = 1, ..., k$. In each period the agent can now set the production level for the next

[1]Our code is available here: http://anonymouslinktoyourcode.zip

[2]https://www.ieee.org/publications_standards/publications/conferences/2014_04_msw_a4_format.doc

Fig. 1. Example of a supply chain network with $k = 5$ warehouses $(s_1, .., s_5)$ the factory (s_0).

period $a_0 \in \{0, .., \rho_{max}\}$ where $\rho_{max} \in \mathbb{N}$ is the maximum production aswell as the amount of products that is shipped to each location $a_j \in \mathbb{N}^k$ that is naturally limited by the current storage level in the factory ($\sum_{j=1}^{k} a_j \leq s_0$). Based on this information and the demand d we can now describe the transition for s_0, s_j, $j = 1, ..., k$ by

$$\begin{aligned} s_0' &= \min\{s_0 + a_0 - \sum_{j=1}^{k} a_j\}, \\ s_j' &= \min\{s_j + a_j - d_j\},\ j = 1, ..., k. \end{aligned} \tag{2}$$

The reward within each period consists of the revenue from sold products at a fix price p less production costs $\kappa_{pr} a_0$, storage costs $\sum_{j=0}^{k} \kappa_{st,j} \max\{s_j, 0\}$, penalty costs $\kappa_{pe} \sum_{j=1}^{k} \min\{s_j, 0\}$ and transportation costs $\sum_{j=1}^{k} \kappa_{tr,j} \lceil a_j/\zeta_j \rceil$. We let $p, \kappa_{pr}, \kappa_{pe}, \kappa_{st,j}, \kappa_{tr,j} \in \mathbb{R}$ and $\zeta_j \in \mathbb{N}$ for $i = 1, ..., k$. We can now define the one-step reward function by

$$\begin{aligned} r(s, d, a) :=& p \sum_{j=1}^{k} d_i - \kappa_{pr} a_0 - \sum_{j=0}^{k} \kappa_{st,j} \max\{s_j, 0\} \\ & - \kappa_{pe} \sum_{j=1}^{k} \min\{s_j, 0\} - \sum_{j=1}^{k} \kappa_{tr,j} \left\lceil \frac{a_j}{\zeta_j} \right\rceil \end{aligned} \tag{3}$$

with $\lceil x \rceil$ the ceiling of x. Based on the model description we can now formulate the task as an infinite horizon markov decision process that is subject to a varying environment. Thereby we receive the tuple $(S, D, A, \mathcal{X}, T, Q, r, \gamma)$ with

Definition III.1.

1. $S \times D := \prod_{j=0}^{k} \{s_j \in \mathbb{Z} \mid s_j \leq c_j\} \times D, c_j \in \mathbb{N}$ the state space consisting of elements $(s, d) = ((s_0, s_1, ..., s_k), d)$;
2. $A := \{0, ..., \rho_{max}\} \times \mathbb{N}_0^k$ the action space consisting of elements $a = (a_0, ..., a_k)$;

3. $\mathcal{X}(s) \coloneqq \{0, ..., \rho_{max}\} \times \{a \in \mathbb{N}_0^k \mid \sum_{j=1}^{k} a_j \leq s_0\}$, the set of all feasible actions in a state s and $\mathcal{X} \coloneqq \{(s, d, a) \in S \times D \times A \mid a \in \mathcal{X}(s)\}$;
4. $T : S \times D \times A \to S$ the transition function defined by

$$\begin{aligned} T(s,d,a) \coloneqq (&\min\{s_0 + a_0 - \sum_{j=1}^{k} a_j, c_0\}, \\ &\min\{s_1 + a_1 - d_1, c_1\}, \\ &... \\ &\min\{s_k + a_k - d_k, c_k\}); \end{aligned}$$

5. $Q : \mathcal{X} \times S \times D \to [0, 1]$ the transition probabilities defined by $Q(s', d' \mid s, d, a) \coloneqq q_d(d')$ for $s' = T(s, d, a)$ and 0 otherwise;
6. $r : S \times D \times A \to \mathbb{R}$ the one-step reward function as defined in Eq. (3);
7. $\gamma \in (0, 1)$ the discounting factor.

For the calculation of the expected discounted revenue for each tuple (s, d) we obtain the value function V with

$$\begin{aligned} V(s,d) = \max_{a \in \mathcal{X}(s)} \{&r(s,d,a) \\ &+ \gamma \sum_{d' \in D} q_d(d') V(T(s,d,a), d')\}, \\ &s \in S, d \in D. \end{aligned} \tag{4}$$

IV. The Agent

The agent you designed for your environment. Justify your choice and design and explain briefly how you implemented/configured it. Naturally, if you took a ready-made environment, you should invert relatively much more effort into this section than the previous one.

V. Results and Discussion

This is one of the most important sections. You put your agent to the test in the environment, you show – and most importantly – you interpret the results.

A. Performance of your Agent in your Environment

Show plots, graphs, tables (e.g., Table I), etc. You may wish to encourage readers to reproduce results for themselves, e.g., run `runDemo.py` in our source code. Show how your agent performs well, or, if it doesn't perform well, it is better to explain why (this is a result in itself!). In any case, you *must* highlight the weaknesses of your agent as well as its strengths.

TABLE I
THIS TABLE IS JUST AN EXAMPLE.

Environment config.	Standard SARSA	Our Improved Agent
Simulation 1	10	15
Simulation 2	12	11

B. Performance of your Agent in the ALife Environment

You deploy your agent in the ALife[3] environment (a random screenshot shown in Figure 2). Does it work well? Why? Why not? Justify the adaptation you think is best.

[3]https://github.com/jmread/alife

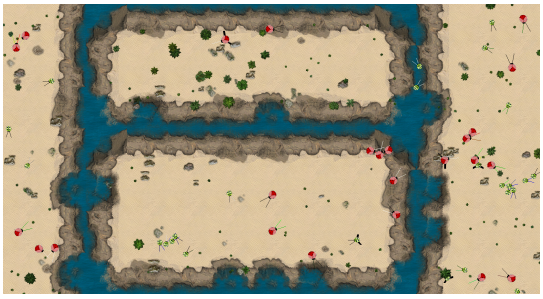

Fig. 2. An example figure

VI. Conclusion and Future Work

This section summarizes the paper: Your environment and agent, its strength and its weaknesses. Also remark about what would be the next steps you would take if you or someone else were to continue/extend this project. Note that for the initial submission you are limited strictly to 4 pages (double column), *not including references*. An extra page will be allowed for final submission (after the initial reviews).

References

[1] D. Barber. Bayesian Reasoning and Machine Learning, *Cambridge University Press*, 2012.

[2] J. Read. Lecture III - Structured Output Prediction and Search. *INF581 Advanced Topics in Artificial Intelligence*, 2018.

[3] D. Mena et al. A family of admissible heuristics for A* to perform inference in probabilistic classifier chains. *Machine Learning*, vol. 106, no. 1, pp 143-169, 2017.

[4] O. Vinyals et al. StarCraft II: A New Challenge for Reinforcement Learning. https://arxiv.org/abs/1708.04782, 2017.