Reinforcement learning for supply chain optimization

L. Kemmer, H. von Kleist, D. de Rocheboüet, N. Tziortziotis, J. Read
LIX, École Polytechnique, France

Problem Setting

Motivation

- Optimizing supply chains consisting of a factory and multiple warehouses is a problem faced by many companies
- **Main decision**: production quantities at the factory and quantities shipped to the warehouses
- Small firms can manage supply chains manually but big companies need more elaborate tools. This motivates our adaptation of reinforcement learning (RL) agents that we test in 2 scenarios against a heuristic baseline algorithm

The Model

- The problem is modeled by a Markov decision process [1,2] where t denotes the current period and j a specific warehouse
- It is defined by a **state-space $\mathbf{s_t}$** that represents stock levels at the factory and warehouses, the **demand $\mathbf{d_t}$** at each warehouse, an **action space $\mathbf{a_t}$** that sets the production level and transportation quantities to the warehouses, a set of **feasible actions** (it is not possible to ship more that what is in stock), a **transition function $\boldsymbol{T(s_t, d_t, a)}$**, a **one-step reward function** r_t and a **discount factor** γ
- The demand $d_{j,t} = \left\lfloor \frac{d_{max}}{2} \sin\left(\frac{2\pi(t+2j)}{12}\right) + \frac{d_{max}}{2} + \epsilon_{j,t} \right\rfloor$ (with $P(\epsilon_{j,t} = 0) = P(\epsilon_{j,t} = 1)$=0,5) incorporates a seasonal trend
- The one step-reward function r_t consists of **revenue from sold products**, **production costs**, **storage costs**, **penalty costs** for unsatisfied demands and **transportation costs** from the factory to the warehouses

Fig. 1: Supply chain network with 5 warehouses and 1 factory

Approximate SARSA

- Use a **linear Q-function approximation** $Q_w(s,a) = w^T\phi\,(s,a)$ with a parameter-vector w and a feature map $\phi\,(s,a)$ to deal with exponentially growing state and action spaces
- Reasonable knowledge about the structure of the MDP is assumed and used to construct **over 15 different features** to design $\phi\,(s,a)$
- In order to achieve reasonable knowledge about the dynamics of the demand process past demands are used to create a **simple forecast of future demands**

REINFORCE

- **Discretize action space:** 3 actions per location, total of $3^{(K+1)}$ actions
- **Maximize policy function:** $p(a = a^{(i)}|s) = \frac{e^{\phi(s)^T w_{a^{(i)}}} \cdot f(a^{(i)}|s)}{\sum_{j=1}^{n_a} e^{\phi(s)^T w_{a^{(j)}}} \cdot f(a^{(j)}|s)}$

  where $f(a^{(j)}|s) = \begin{cases} 1, & \text{if } a^{(j)} \text{ is allowed in state } s, \\ 0, & \text{otherwise.} \end{cases}$
- **Different options for feature map $\phi\,(s)$:** Linear, Quadratic and RBF

Results

- Approximate SARSA [3] and REINFORCE [4] agents are compared to the baseline (ς, Q) agent [5] that replenishes each warehouse j by an amount Q_j if the current stock is below c_j and there is still stock left in the factory
- **Scenario 1**: 1 warehouse and 1 factory with costs for production, storage (only warehouse), transportation and penalty costs
- **Scenario 2**: 3 warehouses and 1 factory with costs for production, storage (only for warehouse no. 1), transportation and penalty costs
- The **(s,Q)-policy** is **outperformed by REINFORCE and approximate SARSA** in scenario 1, and by REINFORCE in scenario 2
- RL agents learn to invest in stock and transportation, despite short-term negative rewards

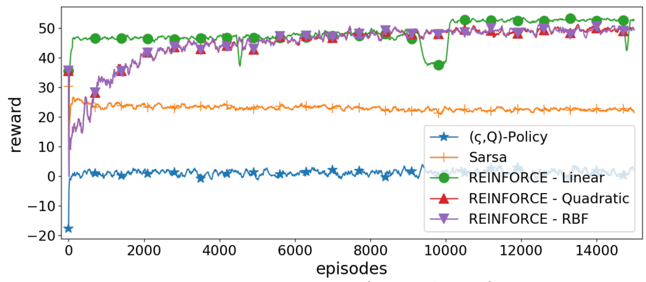

Fig. 2: Scenario 1 (1 warehouse).

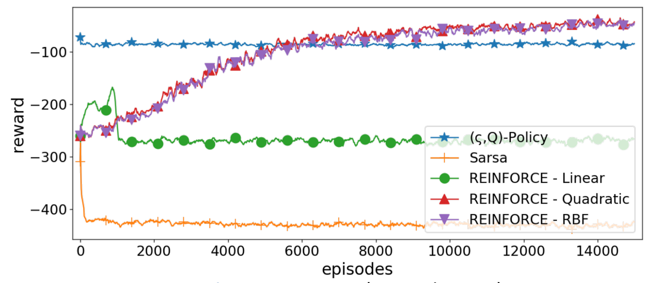

Fig. 3: Scenario 2 (3 warehouses).

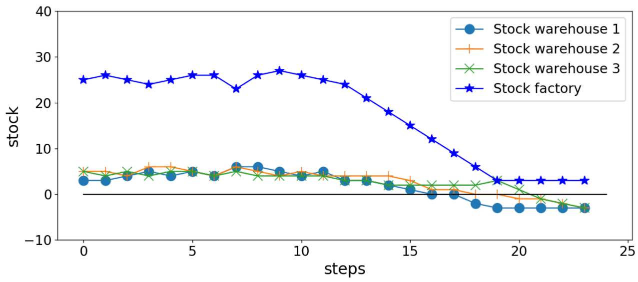

Fig. 4: Stocks for the REINFORCE agent using a quadratic feature map ϕ.

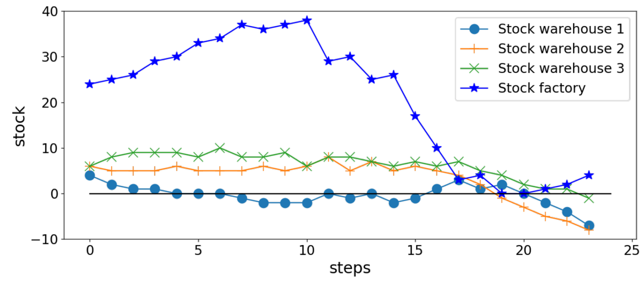

Fig. 5: Stocks for the (ς, Q)-Policy based agent.

References

[1] Warren B. Powell. Approximate dynamic programming : solving the curses of dimensionality. Wiley series in probability and statistics. Wiley-Interscience, Hoboken, NJ,2007. ISBN 978-0-470-17155-4.
[2] Lars Norman Moritz. Target Value Criterion in Markov Decision Processes. PhD thesis, 2014.
[3] Gavin A Rummery and Mahesan Niranjan. On-line q-learning using connectionist systems. Technical report, Cambridge University, 1994.
[4] Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning, 8(3):229{256, 1992.
[5] Horst Tempelmeier. Inventory management in supply networks : problems, models, solutions. Books on Demand, Norderstedt, 2. ed. edition, 2011. ISBN 978-3-8423-4677- 2.