MARLIM: Multi-Agent Reinforcement Learning for Inventory Management

Rémi Leluc[1], Elie Kadoche[2], Antoine Bertoncello[2], Sébastien Gourvénec[2]
[1] LTCI, Télécom Paris, Institut Polytechnique de Paris, [2]TotalEnergies OneTech

Inventory Management

• **GOAL:**
Find the right balance between the **supply** and **demand** of products by **optimizing replenishment decisions** and **minimizing costs**.

• **BENEFITS:**
↪ Better inventory accuracy
↪ Insights to cost savings
↪ Avoidance of stock-outs

• **FRAMEWORK**:
A controller observes the past demands and local information of the inventory and has to decide about the next ordering values.

• **MAIN ISSUE:** Environment uncertainty

↪ demands and lead-times are stochastic with potentially high volatility.

↪ controller may exceedingly order, leading to unnecessary *ordering* and *holding* costs.

↪ controller may insufficiently order, leading to *shortage* costs and may jeopardize the company's performance.

Contributions

(1) We develop a novel reinforcement learning framework, called **MARLIM**, to address the inventory management problem for a single-echelon multi-products supply chain on a production line with stochastic demands and lead-times.

(2) We provide the methodology to train agents in different scenarios for fixed or shared capacity constraints with specific handling of storage overflows.

(3) We perform various numerical experiments on real-world data to demonstrate the benefits of our method over classical baselines.

Inventory Costs

For any item $i \in \mathcal{N}$, denote by $C_o^{(i)}, C_h^{(i)}$ and $C_s^{(i)}$ the unit ordering, holding and shortage costs respectively.

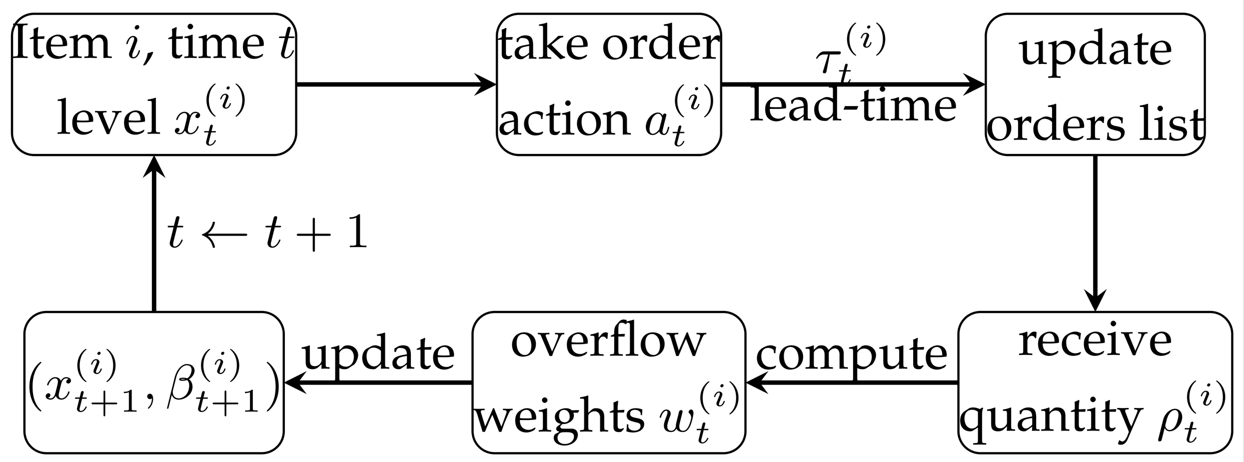
(a) **Ordering costs**: functioning costs, reception costs, salaries personnel, labor costs, rent of a factory, energy consumption allocated for production.

(b) **Holding costs**: financial and functional costs, rent and maintenance of required space, insurance costs, transportation and obsolescence costs.

(c) **Shortage costs**: demand exceeds available inventory ↪ backlogging costs and penalty shortage cost.

Inventory Dynamics

At time t, for each product $i = 1, \dots, n$, inventory controller decides about the order $a_t^{(i)}$ to take based on the current inventory level $x_t^{(i)}$ and the stochastic demand $\delta_t^{(i)}$. Order arrives after stochastic lead-time $\tau_t^{(i)}$.



After receiving replenishment quantity $\rho_t^{(i)}$ the inventory levels are temporarily updated through $\lambda_t^{(i)} = x_t^{(i)} + \lfloor w_t^{(i)} \rho_t^{(i)} \rfloor$. The levels and backlogs β_t are updated to evaluate the inventory costs $C_t^{(i)}$.

$$x_{t+1}^{(i)} = \left(\lambda_t^{(i)} - \delta_t^{(i)}\right)_+ \quad \beta_{t+1}^{(i)} = \beta_t^{(i)} + \left(\lambda_t^{(i)} - \delta_t^{(i)}\right)_-$$

$$C_t^{(i)} = \alpha_o \underbrace{a_t^{(i)} C_o^{(i)}}_{\text{ordering}} + \alpha_h \underbrace{x_t^{(i)} C_h^{(i)}}_{\text{holding}} + \alpha_s \underbrace{\beta_{t+1}^{(i)} C_s^{(i)}}_{\text{shortage}}$$

where $\alpha_o, \alpha_h, \alpha_s \in [0,1]$ with $\alpha_o + \alpha_h + \alpha_s = 1$ are weighting coefficients that translate some expert's knowledge about the desired strategy.

Inventory Features

• The inventory costs of each agent are associated to the single reward defined as: $r^i(s_t, a_t) = -C_t^i$. Inside a product subspace $\mathcal{N}_k$, the agents are working in a cooperative setting in order to optimize the average reward $r_k(s_t, a_t) = \sum_{i \in \mathcal{N}_k} r^i(s_t, a_t)/|\mathcal{N}_k|$.

• warehouse with n independent products (one agent per item). capacity of each product is either: (1) finite and non-variable for each product (single agents) or (2) shared with finite capacity in cluster (multi-agents RL).
• stochastic demands and lead-times (e.g. Poisson, Geometric) with stationary distributions that may be infered from historical data.

Numerical Details

• Lead-time geometric $\tau^{(i)} \sim \mathcal{G}(p_i)$.
• Demand $\delta^{(i)} \sim X_i Y_i$ with $X_i \sim \mathcal{B}(b_i), Y_i \sim \mathcal{P}(\mu_i)$.
MinMax agents: a standard min-max strategy (s, S) from operation research.
Oracle agents: order at each time according to a normal law $\mathcal{N}(\hat{\mu}_\delta, \hat{\sigma}_\delta^2)$ which is clamped to fit the bounds of the action space.
MARL agents: PPO algorithm, when working with capacity constraints per item, both discrete and continuous policies are considered, denoted by PPO-D and PPO-C respectively. When the items compete for storage space, we implement IPPO.

Numerical Results

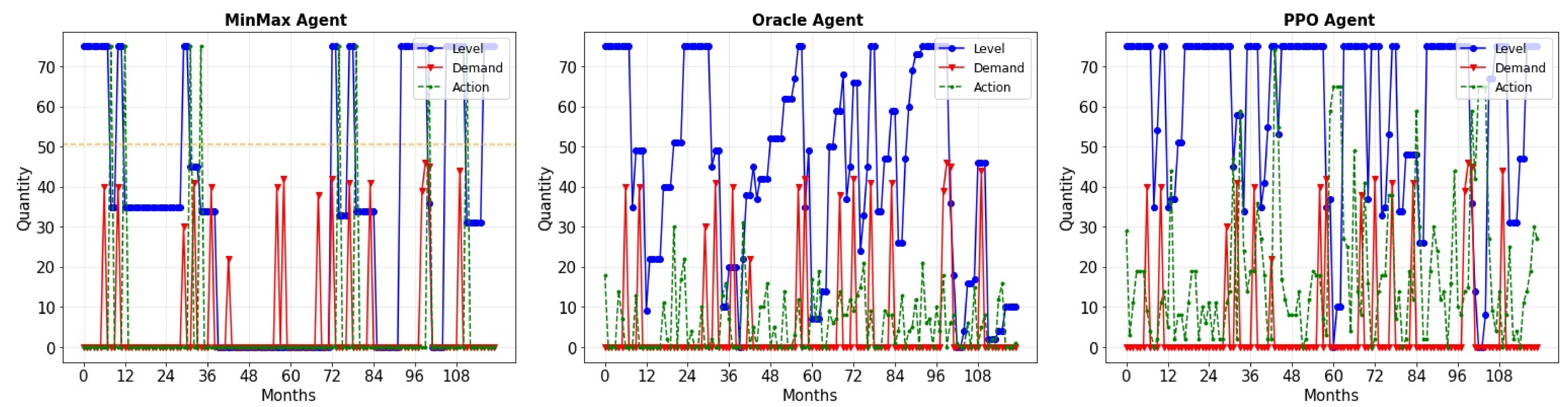


Figure 1: Inventory level (*blue*) over $T = 120$ months: MinMax (left), Oracle (center), PPO (right). The demand is plotted in *red* and the order actions are plotted in *green*. The safety stock MinMax agent is displayed in *orange*.

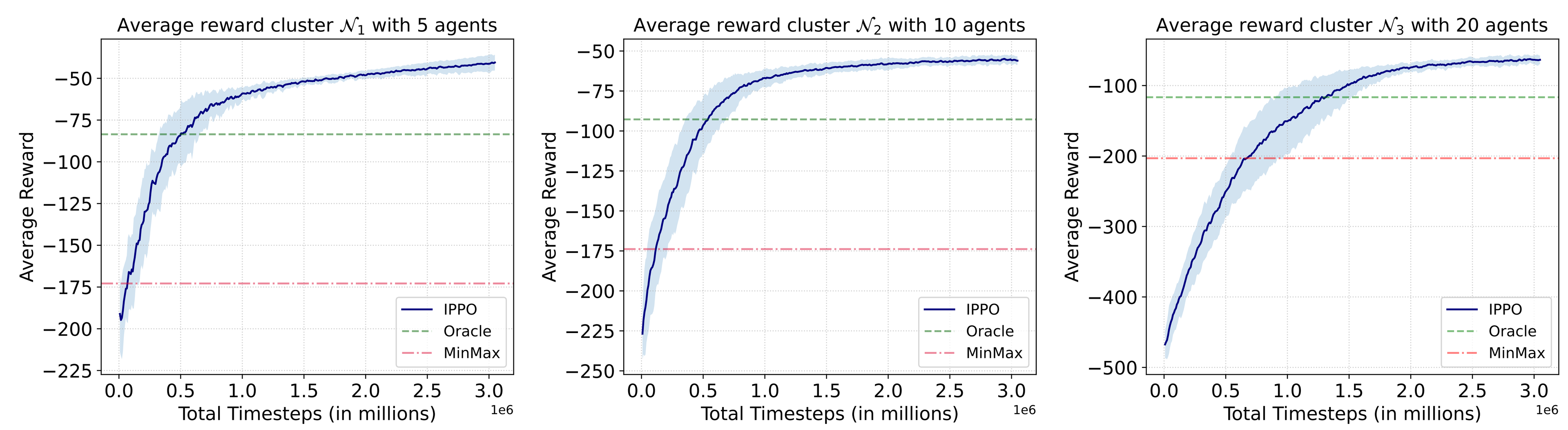


Figure 2: Learning curves of clusters $\mathcal{N}_1, \mathcal{N}_2$ and $\mathcal{N}_3$ where the mean and standard deviation of IPPO are plotted in *blue* and the horizontal lines are average reward for baselines Oracle (*green*) and MinMax (*red*) computed over 100 replications.

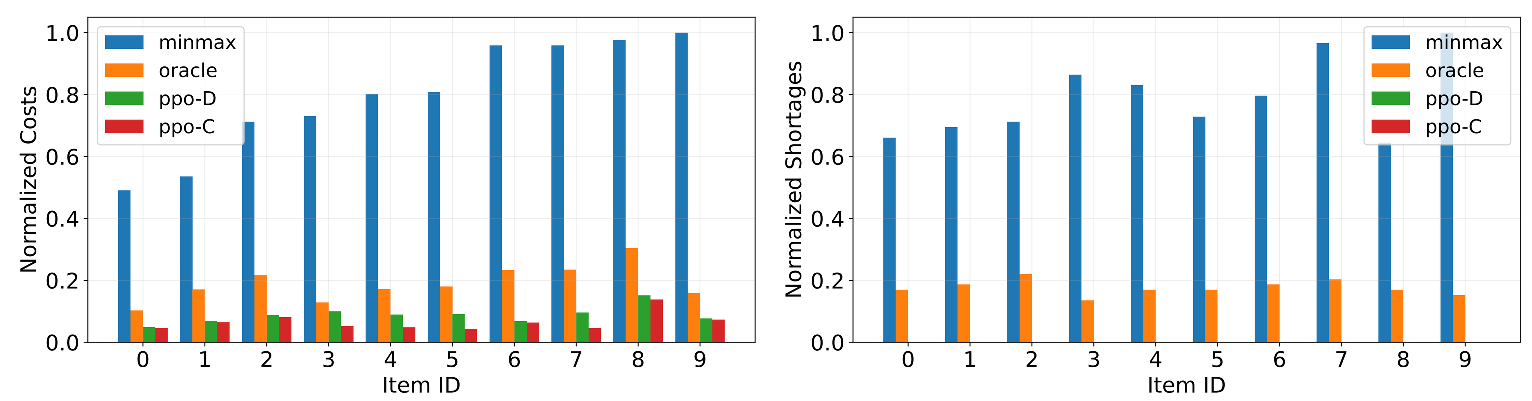


Figure 3: Average Cumulative Costs and Item Shortages obtained over 100 replications, horizon $T = 240$ months for Items 0-9.