

DOI:10.3969/j.issn.1671-0673. 2026. 01. 005

改进双延迟深度确定性策略梯度的库存控制方法

龚永奇, 郭基联, 张 亮, 唐希浪

(空军工程大学, 陕西 西安, 710038)

摘要: 针对不确定需求和供应延迟环境中库存控制难度大、成本偏高的问题, 提出一种改进双延迟深度确定性策略梯度(TD3)的库存控制方法。首先, 将库存控制抽象为满足率最高和成本最小双目标的马尔可夫决策过程, 作为 TD3 算法的训练环境; 其次, 采用优先经验回放机制, 提高 TD3 算法的采样效率, 将长短期记忆网络融入 TD3 算法的多层感知机, 优化网络结构; 最后, 通过 TD3 算法与环境交互, 实现库存控制中成本和满足率优化。实验结果表明, 所提方法在达到满足率阈值条件下, 其库存控制成本较原始 TD3 算法降低 22.2%。

关键词: 库存控制; 双延迟深度确定性策略梯度; 优先经验回放; 长短期记忆网络

中图分类号: TP18; E92

文献标识码: A

文章编号: 1671-0673(2026)01-0035-07

Inventory Control Method Based on Improved Twin Delayed Deep Deterministic Policy Gradient

GONG Yongqi, GUO Jilian, ZHANG Liang, TANG Xilang

(Air Force Engineering University, Xi'an 710038, China)

Abstract: To address the difficulty and high cost of inventory control in the environment of uncertain demand and supply delays, an inventory control method based on the improved twin delayed deep deterministic policy gradient (TD3) is proposed. Firstly, the inventory control is abstracted as a Markov decision process with dual objectives of service level maximization and cost minimization, which serves as the training environment for the TD3 algorithm. Secondly, the prioritized experience replay mechanism is adopted to improve the sampling efficiency of the TD3 algorithm, and the long short-term memory (LSTM) is integrated into the multi-layer perceptron of the TD3 algorithm to optimize the network structure. Finally, the TD3 algorithm is used to interact with the environment to optimize both cost and service level in inventory control. Experimental results demonstrate that the inventory control cost of the proposed method is 22.2% lower than that of the original TD3 algorithm when the service level threshold is reached.

Key words: inventory control; TD3; prioritized experience replay; LSTM

库存管理是供应链运营的关键, 库存过低会导致缺件, 无法满足使用需求, 而库存过高会导致呆滞备件, 造成资金浪费^[1]。但在实际管理中, 备件需求并非恒定不变, 存在随机性, 而且各级仓库之间需要供应时间, 存在供应延迟, 这加大了库存控制的难度^[2]。因此, 库存控制策略是供应链优化

的重难点问题。

面对动态变化的库存环境, 基于状态的库存控制策略应用广泛, 将库存控制抽象为马尔可夫决策过程(Markov Decision Process, MDP), 依据库存状态确定备件采购和供应动作, 其求解方法主要有动态规划(Dynamic Programming, DP)算法和深

收稿日期: 2025-09-17; 修回日期: 2025-10-28

基金项目: 国家自然科学基金(72201276)

作者简介: 龚永奇(1992—), 男, 助理工程师, 博士生, 主要研究方向为备件供应保障。E-mail: david_gyq123@163.com

度强化学习(Deep Reinforcement Learning, DRL)算法^[3]。DP算法包括价值迭代、策略迭代等^[4-5]。DRL算法包括优势演员—评论家(Advantage Actor-Critic, A2C)、近端策略优化(Proximal Policy Optimization, PPO)、软演员—评论家(Soft Actor-Critic, SAC)、双延迟深度确定性策略梯度(Twin Delayed Deep Deterministic Policy Gradient, TD3)等^[6-7]。DP算法只能求解小规模库存控制问题,当问题规模变大时,因维数灾难而无法求解;而DRL算法使用神经网络逼近函数值,能够解决大规模、多需求类型、多级库存控制问题^[8-9]。

面对不确定需求和供应时间延迟的库存环境,DRL算法只学习当前状态信息,生成相应动作,而忽视挖掘需求变化信息,往往优化效果不佳。对此,文献[10]在状态向量中添加历史需求信息,使DRL算法学习历史需求变化过程,将需求预测集成到策略中。文献[11]在文献[10]的基础上,采用PPO、SAC、TD3等主流DRL算法优化库存控制策略,验证实际应用价值。但面对多级库存的高维决策空间,DRL算法尤其是具有良好性能的TD3算法^[12],采用随机采样训练和多层感知机(Multi-Layer Perceptron, MLP),仍存在收敛速度慢、训练不稳定导致的库存成本偏高问题。此外,上述DRL算法主要以收入或成本为奖励函数,未能很好反映备件满足率要求。

针对上述问题,提出一种改进TD3的库存控制方法。首先,采用自适应权重方法,构建满足率和总成本双目标的库存控制模型,作为TD3算法的训练环境;其次,采用优先经验回放(Prioritized Experience Replay, PER)机制提高TD3算法的采样效率,引入长短期记忆网络(Long Short-Term Memory, LSTM)优化TD3算法网络结构;最后,以不确定需求和供应时间延迟的多级库存环境为例,验证算法性能,为库存控制提供技术参考。

1 基于MDP的库存控制模型

1.1 问题描述与模型假设

备件库存节点一般包括后方仓库和中继仓库。后方仓库具有采购和储备功能,向中继仓库供应备件,中继仓库满足用装单位备件需求。随着备件供应保障开展,每个阶段备件需求发生变化,后方仓库和中继仓库的采购供应动作和库存状态也随之改变,整个库存环境处于动态变化中。因此,在需求不确定和供应延迟时间的环境中,库

存控制策略的关键是确定后方仓库的动态采购数量和后方仓库向中继仓库的动态供应数量。

为便于建模,模型假设:1)不考虑同级仓库之间横向供应;2)后方仓库、中继仓库供应关系已知;3)各中继仓库的备件需求相互独立。

1.2 MDP建模

MDP模型的关键参数可定义为四元组 $M = (S, A, P, R)$ 。其中: S 表示状态空间; A 表示动作空间; P 表示状态转移规则; R 表示奖励函数值^[10]。相关数学符号及含义如表1所示。

表1 数学符号及含义

符号	含义
m	中继仓库总数,索引为 i
T	备件供应阶段总数,索引为 t
d_i^t	阶段 t 中继仓库 i 的需求量
s_0^t	阶段 t 后方仓库的库存数量
s_i^t	阶段 t 中继仓库 i 的库存数量
V_0	后方仓库的库存容量
V_i	中继仓库 i 的库存容量
U	后方仓库的最大采购量
x_0^t	阶段 t 后方仓库的采购数量
y_i^t	阶段 t 后方仓库向中继仓库 i 的供应数量
κ_i	后方仓库向中继仓库 i 的供应延迟时间
z_p	单位备件采购成本
z_c^i	中继仓库 i 的单位备件储备成本
z_s	单位运输车的运输成本
F	运输车的运输备件容量
z_c^0	后方仓库的单位备件储备成本
z_b	单位缺货停机损失成本

1.2.1 状态空间与动作空间

状态向量描述后方仓库和中继仓库的库存数量,状态向量定义如下:

$$S^t = [s_0^t, s_1^t, s_2^t, \dots, s_m^t] \quad (1)$$

由于库存控制策略受供应延迟时间影响,将每个中继仓库的历史需求序列作为额外状态信息,补充至状态向量,序列的长度依据后方仓库到中继仓库的供应延迟时间确定,即

$$S^t = [s_0^t, s_1^t, \dots, s_m^t, D_1^t, \dots, D_m^t] \quad (2)$$

式中, $D_i^t = [d_i^{t-\kappa_i}, \dots, d_i^t]$, $i \in [1, m]$, 表示中继仓库历史需求序列。

动作向量描述后方仓库的采购数量,以及后方仓库向中继仓库的供应数量,动作向量定义如下:

$$A^t = [x_0^t, y_1^t, y_2^t, \dots, y_m^t] \quad (3)$$

状态空间和动作空间分别是描述状态向量和

动作向量的约束或范围。状态空间和动作空间可定义如下:

$$0 \leq s_0^t \leq V_0 \quad (4)$$

$$s_i^t \leq V_i \quad (5)$$

$$0 \leq x_0^t \leq U \quad (6)$$

$$\sum_{i=1}^m y_i^t \leq x_0^t + s_0^{t-1} \quad (7)$$

其中:式(4)表示后方仓库的库存数量不超过其库存容量;式(5)表示中继仓库的库存数量不超过其库存容量;式(6)表示后方仓库采购数量不超过其最大采购能力;式(7)表示阶段 t 后方仓库向各中继仓库的供应数量不超过阶段 t 后方仓库采购数量和阶段 $(t-1)$ 库存数量之和。

1.2.2 状态转移规则

后方仓库的库存状态 s_0^t 随采购数量 x_0^t 和供应数量 $\sum_{i=1}^m y_i^t$ 动态变化,其状态转移方程为

$$s_0^t = s_0^{t-1} + x_0^t - \sum_{i=1}^m y_i^t \quad (8)$$

由于后方仓库与中继仓库存在供应延迟时间,任一中继仓库的储备数量 s_i^t 随延迟供应数量 $y_i^{t-\kappa_i}$ 和需求 d_i^t 动态变化,其状态转移方程为

$$s_i^t = s_i^{t-1} + y_i^{t-\kappa_i} - d_i^t, i \in [1, m] \quad (9)$$

1.2.3 奖励函数

奖励函数是MDP模型的核心,库存控制策略优化目标为满足率和成本,以自适应权重调节方式^[13],构建奖励函数。

优化目标1:满足率最高。阶段 t 满足率为

$$v_{d_rate}^t = \frac{1}{m} \sum_{i=1}^m v_i^t \quad (10)$$

式中, v_i^t 表示阶段 t 中继仓库 i 的满足率。

任意阶段 t 备件需求未满足作缺件保持,故 s_i^t 需同时表征库存数量和需求满足情况。若 $s_i^t \geq 0$,表示阶段 t 中继仓库备件需求满足后的库存数量;若 $s_i^t < 0$,表明阶段 t 中继仓库缺件数。中继仓库的满足率由其储备状态确定,即:

$$v_i^t = \begin{cases} 1, & s_i^t \geq 0; \\ 1 + s_i^t/d_i^t, & s_i^t < 0. \end{cases} \quad (11)$$

优化目标2:成本最小。成本主要包括采购成本 w_p^t 、运输成本 w_s^t 、储备成本 w_c^t 及缺件停机损失成本 w_b^t 。阶段 t 成本为

$$w^t = w_p^t + w_s^t + w_c^t + w_b^t = z_p x_0^t + \sum_{i=1}^m z_s \left\lceil \frac{y_i^t}{F} \right\rceil + z_c^0 s_0^t + \sum_{i=1}^m z_c^i \max(s_i^t, 0) - \sum_{i=1}^m z_b \min(s_i^t, 0) \quad (12)$$

式中: $z_p x_0^t$ 表示后方仓库的采购成本; $\sum_{i=1}^m z_s \left\lceil \frac{y_i^t}{F} \right\rceil$ 表示

后方仓库向中继仓库的运输成本,其中 $\left\lceil \frac{y_i^t}{F} \right\rceil$ 表示运输车数量, $\lceil \cdot \rceil$ 为向上取整,即备件运输不满运输车容量,向上取1辆车; $z_c^0 s_0^t$ 表示后方仓库的储备成本, $\sum_{i=1}^m z_c^i \max(s_i^t, 0)$ 表示中继仓库的储备成本,其中

$\max(s_i^t, 0)$ 为正,表示储备数; $-\sum_{i=1}^m z_b \min(s_i^t, 0)$ 表示缺件停机损失成本, $\min(s_i^t, 0)$ 为负,表示缺件数。

依据满足率最高和成本最小两个优化目标,为消除成本数据纲量影响,引入纲量系数 δ ,同时设置权重系数 θ ,阶段 t 的奖励值为

$$R^t = \theta^t v_{d_rate}^t - (1 - \theta^t) \frac{w^t}{\delta} \quad (13)$$

为避免过度追求满足率,导致备件闲置而产生浪费。引入自适应调整因子 μ ,引导满足率 $v_{d_rate}^t$ 向设定区间 $[v_{d_rate}^{\min}, v_{d_rate}^{\max}]$ 内运动,即

$$\theta^t = \begin{cases} \theta^{t-1} - \mu, & v_{d_rate}^{t-1} > v_{d_rate}^{\max}; \\ \theta^{t-1} + \mu, & v_{d_rate}^{t-1} \leq v_{d_rate}^{\min}. \end{cases} \quad (14)$$

2 改进TD3算法

2.1 TD3算法

TD3算法将MDP视为环境交互过程,采用Actor-Critic框架,并通过环境反馈奖励值指导策略更新^[13]。

该算法包括Actor网络 φ 和目标Actor网络 φ' ,使用目标Actor网络计算 S^{t+1} 状态下的动作策略 A^{t+1} ,可表示为

$$A^{t+1} = \varphi'(S^{t+1}) + \varepsilon' \quad (15)$$

式中, ε' 为随机噪声,服从裁剪正态分布(均值为0,方差为 σ^2),即 $\varepsilon' \sim \text{clip}(N(0, \sigma^2), -\varepsilon, \varepsilon)$ 。

Critic网络(又称 Q 网络)由 Q_1 和 Q_2 两个网络组成,每个 Q 网络对应1个目标 Q 网络,分别为 Q'_1 和 Q'_2 。将 (S^{t+1}, A^{t+1}) 输入两个目标 Q 网络,取最小值确定目标 Q 值,即:

$$r^t = R^t + \gamma \min_{k=1,2} (Q'_k(S^{t+1}, A^{t+1})) \quad (16)$$

式中: γ 表示折扣因子; R^t 表示阶段 t 奖励值; k 表示 Q 网络序号。

Q 网络参数 ϕ_k 的损失函数为

$$L^{\phi_k} = E[(Q_k(S^t, A^t) - r^t)^2], k = 1, 2 \quad (17)$$

Actor网络参数 ψ 的损失函数为

$$L^{\psi} = E[\nabla Q_1(S^t, A^t) | A^t = \varphi(S^t)] \quad (18)$$

目标 Q 网络参数 ϕ_k^{target} 和目标 Actor 网络参数 ψ^{target} 均通过软更新方法,实现参数训练,即:

$$\begin{cases} \phi_k^{\text{target}} = \tau \phi_k + (1 - \tau) \phi_k^{\text{target}_0}, k = 1, 2; \\ \psi^{\text{target}} = \tau \psi + (1 - \tau) \psi^{\text{target}_0}. \end{cases} \quad (19)$$

式中: τ 为软更新系数; $\phi_k^{\text{target}_0}$ 为目标 Q 网络更新前参数; ψ^{target_0} 为目标 Actor 网络更新前参数。

2.2 PER 机制

TD3 算法因采用同等概率采样,导致不太重要的经验样本易被反复采样,面对高维动作空间往往收敛速度慢。而 PER 机制通过赋予不同样本的优先级权重,使优先级重要的样本能够被更大概率抽取。因此,将 PER 机制和 TD3 算法相结合,抽取高价值样本训练,提升模型的鲁棒性。

PER 机制采用时序差分误差表示样本的重要程度,而时序差分误差是当前状态价值的估计值与目标值之间的误差,误差值越大,样本越重要,优先级权重越大。设经验回放池样本总量为 L , 样本 l 的时序差分误差为 η_l , 则样本 l 被采样概率为

$$P(l) = \left(\eta_l \cdot \left(\sum_{l=1}^L \eta_l \right)^{-1} \right)^\alpha \quad (20)$$

式中, α 为优先级系数, $\alpha \in [0, 1]$ 。

针对非均匀采样导致梯度估计偏差,校正重要性采样权重为

$$\omega_l = \left(\frac{1}{L \cdot P(l)} \right)^\beta \quad (21)$$

式中, β 为重要性采样指数, $\beta \in [0, 1]$ 。

2.3 网络设计

由于 TD3 算法中, Actor 网络和 Critic 网络均采用 MLP 网络,在学习状态向量中的历史需求信息

时,其探索性和稳定性还有欠缺。LSTM 网络通过引入门控机制,能够捕捉过去的关键信息,适用于处理和预测时间序列的间隔和延迟事件。因此,将 LSTM 网络结构引入 TD3 算法,使其充分挖掘时间序列的历史需求信息,迭代训练时感知和预测需求的变化过程,进而生成供应时间延迟条件下的备件补充动作和动作状态价值。

设采样批量 B , 初始状态空间维度为 (B, M) , 初始动作空间维度为 (B, K) , LSTM 网络隐藏层神经元个数为 128, MLP 网络采用两个线性层 (Linear) 和 1 个激活函数 (ReLU), 改进 TD3 算法的网络层及参数如表 2 所示。

表 2 改进 TD3 算法的网络层及参数

网络结构	网络层	参数	输出尺寸
Actor 网络	Reshape		$(B, 1, M)$
	LSTM	128	$(B, 1, 128)$
	Reshape		$(B, 128)$
	Linear1	256	$(B, 256)$
	Linear2	512	$(B, 512)$
	ReLU		$(B, 512)$
Critic 网络	状态动作拼接		$(B, M+K)$
	Reshape		$(B, 1, M+K)$
	LSTM	128	$(B, 1, 128)$
	Reshape		$(B, 128)$
	Linear1	256	$(B, 256)$
	Linear2	512	$(B, 512)$
	ReLU		$(B, 512)$

2.4 算法流程

改进 TD3 算法的网络结构如图 1 所示,具体描述如下。

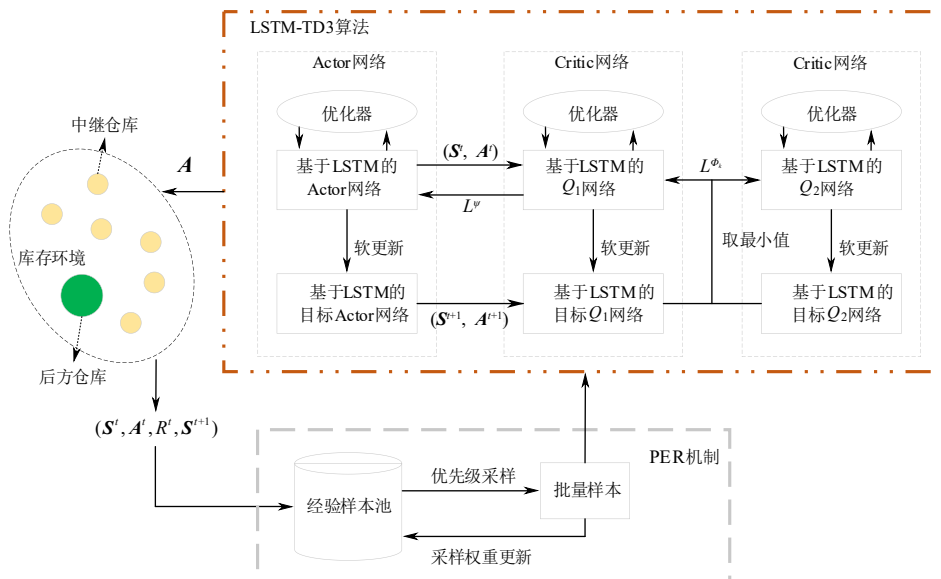


图 1 改进 TD3 算法的网络结构

1)依据库存环境,生成库存动作状态,并将历史需求数据 D' 引入状态向量,以 (S', A', R', S'^{+1}) 作为基础元组样本,构造经验样本池。

2)采用PER机制,依据式(20)以经验优先级权重确定采样概率,并按概率随机抽取训练样本;同时,在每次迭代训练中,依据时序差分误差,更新样本优先级权重,进而更新采样概率。

3)将LSTM按照表2中的网络结构设置,融入TD3算法中两个Actor网络和4个Q网络。每次迭代训练中,首先,状态 S'^{+1} 输入到基于LSTM的目标Actor网络,输出动作 A'^{+1} ;其次, (S'^{+1}, A'^{+1}) 输入到基于LSTM的两个目标Q网络,同时将 (S', A') 输入到基于LSTM的两个Q网络,依据式(16)和(17),用于Q网络参数更新,依据式(18),在Q值增加的方向更新Actor网络参数;最后,依据式(19),对所有目标网络进行软更新。

3 仿真实验

3.1 场景构建

以1个后方仓库和两个中继仓库为库存场景,备件供应保障阶段时长为30 d,每天最大采购量为150个,后方仓库和中继仓库的库存容量分别为300、150个,后方仓库到两个中继仓库的供应延迟时间分别为1、2 d,单个运输车的最大运输容量为30个备件,每辆运输车的费用分别为0.1、0.2万元,单位备件的采购费用为1万元,后方仓库和两个中继仓库的单位备件储备费用分别为0.01、0.2、0.3万元,单位缺件停机损失费用为1万元。

备件需求场景包括不连续型和连续型需求,其中不连续型需求为有限备件需求场景,而连续型需求以概率分布形式表示无限备件需求场景,如泊松分布、正态分布、指数分布及正弦函数(季节性分布)等。在备件库存控制问题中,一方面受任务强度、阶段转换等因素影响,备件需求呈现上下波动,正弦函数更贴近实际保障中的备件需求不确定变化,而其他分布难以表征此类不确定性;另一方面正弦函数的需求变化幅度大,可扩展与其他分布相结合,增加备件需求的不确定性,也被广泛应用于检验不确定需求的库存控制效果^[14-15]。因此,本文以正弦函数为基础、最大需求服从均匀分布为例,构建备件不确定的需求场景。

设置两个中继仓库的需求服从不同正弦函数,最大需求数量 $d_{\max}$ 服从 $[50, 80]$ 区间的均匀分布。

中继仓库1的需求公式为

$$\begin{cases} d_1' = \text{round}\left(\frac{1}{2}d_{\max} \times \sin\left(\frac{\pi}{6}t\right) + \frac{1}{2}d_{\max}\right); \\ d_{\max} = \text{uniform}(50, 80). \end{cases} \quad (22)$$

中继仓库2的需求公式为

$$\begin{cases} d_2' = \text{round}\left(-\frac{1}{2}d_{\max} \times \sin\left(\frac{\pi}{4}t\right) + \frac{1}{2}d_{\max}\right); \\ d_{\max} = \text{uniform}(50, 80). \end{cases} \quad (23)$$

式中: $\text{round}(\cdot)$ 表示四舍五入取整; $\text{uniform}(\cdot)$ 表示均匀分布采样。

3.2 库存控制优化实验

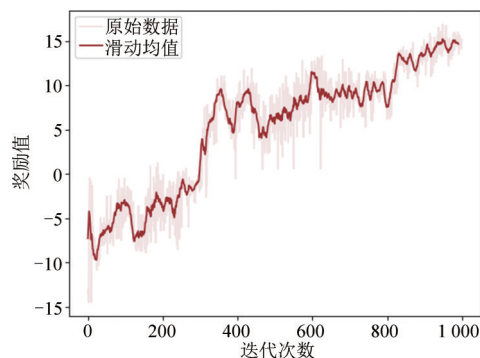
实验关键参数设置如下。

1)迭代次数为1 000,折扣因子为0.95,经验回放区样本范围为2 000,优先级系数为0.6,重要性采样系数为0.4;

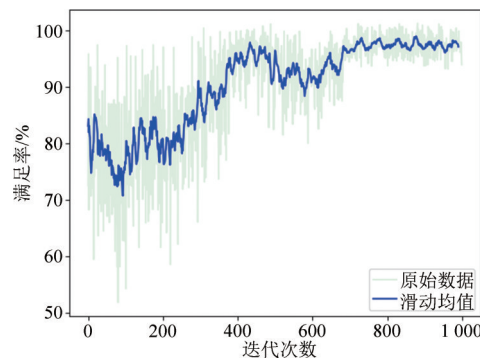
2)满足率和成本优化目标初始权重为0.5,权重调整因子为0.01,纲量系数为10 000,满足率阈值范围为 $[0.95, 0.98]$;

3)采样批量为100,策略函数网络学习率为0.000 5,值函数网络学习率为0.000 5,软更新系数为0.005。

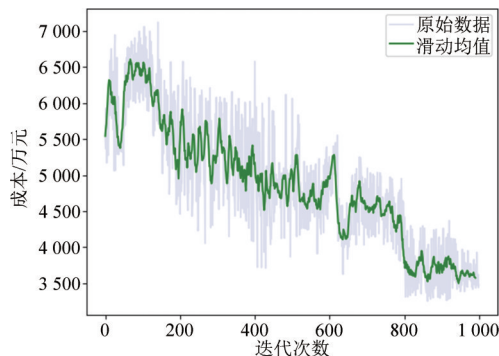
采用原始数据和滑动均值描述仿真迭代变化过程,其各自的奖励函数、满足率和成本的变化情况如图2所示。



(a) 奖励函数变化情况



(b) 满足率变化情况



(c)成本变化情况

图2 迭代训练过程

从图2可知,奖励值随迭代次数趋向于收敛,满足率随迭代次数呈上升趋势,滑动均值浮动于 $[0.95, 0.98]$ 区间范围内,成本随迭代次数呈下降趋势,表明本文所提方法在达到备件满足率阈值要求下,降低了库存成本,验证了所构建奖励函数的合理性,实现了满足率和成本的双目标库存控制优化。

3.3 消融实验

为进一步分析PER机制、LSTM网络对TD3算法的改进效果,分别采用TD3算法、PER-TD3算法、LSTM-TD3算法,以及本文改进TD3算法(PER-LSTM-TD3算法)开展消融实验,其迭代对比如图3所示。

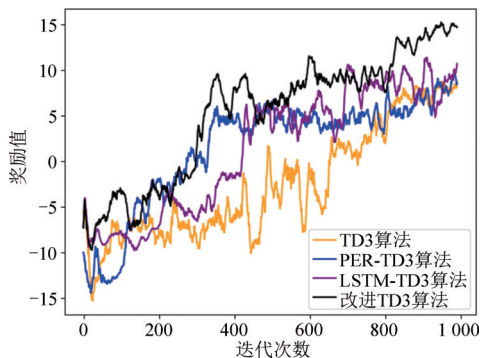


图3 消融实验的迭代图

从图3可得,PER机制和LSTM网络均对TD3算法性能有提升,其中PER机制提升算法收敛速度,LSTM网络提高算法最大奖励值。分别运行10次,不同算法指标取平均,对比结果如表3所示。

表3 消融实验的指标对比

算法	最大奖励值	满足率/%	成本/万元
TD3	7.82	95.3	4 656
PER-TD3	9.13	96.1	4 431
LSTM-TD3	11.40	97.1	3 937
改进TD3	14.60	97.3	3 621

由表3可知,TD3算法、PER-TD3算法在满足率达到阈值约束后,未能有效降低总成本,最大奖励值分别为改进TD3算法最大奖励值的53.6%和62.5%。LSTM-TD3算法满足率与改进TD3算法相近,但收敛速度慢,成本下降缓慢,其最大奖励值为改进TD3算法最大奖励值的78.1%。因此,在达到满足率阈值条件下,改进TD3算法较原始TD3算法库存控制成本降低了22.2%,表明改进TD3算法性能更优,训练效果更明显。

3.4 对比实验

为进一步验证改进TD3算法的性能,分别采用A2C、PPO、SAC等主流DRL算法进行对比分析,其迭代对比如图4所示。

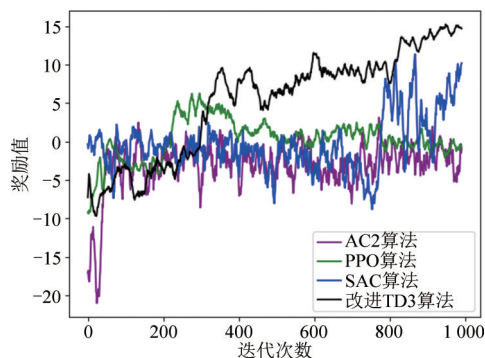


图4 不同算法的迭代图

由图4可知,改进TD3算法随迭代次数的增加,其奖励值呈现增长趋势,且波动幅度缩小,表明该算法收敛性和稳定性更好。分别运行10次,不同算法指标取平均,对比结果如表4所示。

表4 不同算法的指标对比

算法	最大奖励值	满足率/%	成本/万元
A2C	3.42	96.9	5 112
PPO	6.36	96.8	4 891
SAC	10.70	97.2	4 202
改进TD3	14.60	97.3	3 621

由表4可知,改进TD3算法最大奖励值最高,在达到满足率阈值条件下,库存成本最低。因此,与其他主流算法相比,改进TD3算法在库存控制优化问题上更具优势,性能更优。

4 结束语

针对备件需求的不确定性和供应延迟时间影响,构建基于MDP的库存控制模型,并结合PRE机制和LSTM网络对DRL算法中的TD3算法进行改进,有效提升算法性能,并实现满足率和成本的优

化。在实际供应环境中,多个后方仓库、多个中继仓库相互交织,后方仓库之间的采购分配及中继仓库之间的多式调拨更加复杂。下一步,将构建多目标优化算法与深度强化学习算法联合优化的跨区域、大规模备件库存控制模型。

参考文献:

- [1] PUTTEMANS I, CARIS A, BRAEKERS K. The integration of location and inventory decisions in supply chain networks: a literature review and future prospects [J]. *Transportation Research Part E: Logistics and Transportation Review*, 2025, 195: No.103970.
- [2] ZHANG S, HUANG K, YUAN Y F. Spare parts inventory management: a literature review [J]. *Sustainability*, 2021, 13(5): No.2460.
- [3] WANG J, ZHU X Y. Joint optimization of condition-based maintenance and inventory control for a k -out-of- n : F system of multi-state degrading components [J]. *European Journal of Operational Research*, 2021, 290(2): 514-529.
- [4] 康子旭,周栋,李会欣,等.基于飞机状态的备件动态规划技术[J]. *北京航空航天大学学报*, 2024, 50(1): 276-285.
- [5] KHEDR A M, RANI S S. Enhancing supply chain management with deep learning and machine learning techniques: a review[J]. *Journal of Open Innovation: Technology, Market, and Complexity*, 2024, 10(4): No.100379.
- [6] CUARTAS C, AGUILAR J. Hybrid algorithm based on reinforcement learning for smart inventory management [J]. *Journal of Intelligent Manufacturing*, 2023, 34(1): 123-149.
- [7] TIAN R, LU M, WANG H P, et al. IACPPPO: a deep reinforcement learning-based model for warehouse inventory replenishment [J]. *Computers & Industrial Engineering*, 2024, 187: No.109829.
- [8] ZHOU Y F, GUO K, YU C, et al. Optimization of multi-echelon spare parts inventory systems using multi-agent deep reinforcement learning [J]. *Applied Mathematical Modelling*, 2024, 125: 827-844.
- [9] 余诚.基于深度强化学习的多级备件库存策略研究[D]. 南京:东南大学, 2022: 1-71.
- [10] KEMMER L, VON KLEIST H, ROCHEBOUËT D D, et al. Reinforcement learning for supply chain optimization [C]//*Proceedings of the 14th European Workshop on Reinforcement Learning*. [S.l.]: [s.n.], 2018: 1-9.
- [11] 李雨情,潘超凡.基于深度强化学习的配件库存决策研究[J]. *现代计算机*, 2023, 29(11): 16-21.
- [12] ZHANG F J, LI J, LI Z. A TD3-based multi-agent deep reinforcement learning method in mixed cooperation-competition environment[J]. *Neurocomputing*, 2020, 411: 206-215.
- [13] WU J L, WU Q M J, CHEN S Y, et al. A-TD3: an adaptive asynchronous twin delayed deep deterministic for continuous action spaces[J]. *IEEE Access*, 2022, 10: 128077-128089.
- [14] ALVES J C, SILVA D M D, MATEUS G R. Applying and comparing policy gradient methods to multi-echelon supply chains with uncertain demands and lead times [C]//*Proceedings of the 20th International Conference on Artificial Intelligence and Soft Computing*. Cham, Switzerland: Springer, 2021: 229-239.
- [15] STRANIERI F, FADDA E, STELLA F. Combining deep reinforcement learning and multi-stage stochastic programming to address the supply chain inventory management problem [J]. *International Journal of Production Economics*, 2024, 268: No.109099.

(编辑:李志豪)