第 46 卷第 4 期
Vol. 46 No. 4 2025

青岛理工大学学报
Journal of Qingdao University of Technology

基于交叉熵策略优化的强化学习油藏注采优化方法

陈泽宇[1]，刘浩旻[2]，刘丕养[1,*]，张　凯[1]，胡丹丹[3]

（1. 青岛理工大学 土木工程学院，青岛 266525；2. 中国石油化工股份有限公司石油勘探开发研究院，北京 100089；
3. 中国石油勘探开发研究院，北京 100083）

摘　要：在水驱油藏开发过程中，注采优化是调整油藏渗流场分布、提高原油采收率的关键技术。针对现有方法中策略探索能力不足和历史经验样本利用效率低等问题，提出了一种基于交叉熵策略优化的强化学习油藏注采优化方法。该方法将注采优化问题建模为马尔可夫决策过程，以交叉熵策略优化的软演员-评论家算法为框架，利用最大熵策略探索机制增强策略探索能力，以避免陷入局部最优。同时结合交叉熵方法对策略采样过程进行优化，减少无效探索，快速适应油藏流场动态变化。利用与环境交互获得历史调控经验数据，不断迭代优化直至学会最优的注采调控优化策略。将所提出的算法应用于二维油藏模型实例上进行测试，实验结果表明，所提出的基于交叉熵策略优化的强化学习油藏注采优化方法在优化性能和增油控水方面优于传统的进化算法和强化学习算法。

关键词：注采优化；强化学习；马尔可夫决策过程；最大熵策略探索机制；交叉熵策略优化

中图分类号：TE341　　**文献标志码**：A　　**文章编号**：1673-4602(2025)04-0001-10

The oil reservoir injection and production optimization method based on cross-entropy policy optimization in reinforcement learning

CHEN Zeyu[1], LIU Haomin[2], LIU Piyang[1,*], ZHANG Kai[1], HU Dandan[3]

(1. School of Civil Engineering, Qingdao University of Technology, Qingdao 266525, China;
2. Sinopec Petroleum Exploration and Production Research Institute, Beijing 100089, China;
3. Research Institute of Petroleum Exploration and Development, Beijing 100083, China)

Abstract: In the process of waterflood oil reservoir development, injection and production optimization is a key technology for adjusting the reservoir flow field distribution and improving oil recovery factor. To address the issues of insufficient policy exploration capability and low efficiency in utilizing historical experience samples in existing methods, a reinforcement learning-based oil reservoir injection and production optimization method was proposed, based on cross-entropy policy optimization. This method modeled the injection-production optimization problem as a Markov Decision Process (MDP) and utilized the Soft Actor-Critic with Cross-Entropy Policy Optimization (SAC-CEPO) algorithm framework. It employed the maximum entropy policy exploration mechanism to enhance policy exploration capability to avoid local optima. At the same time, the cross-entropy method was used to optimize the

收稿日期：2025-05-08
基金项目：国家自然科学基金(52274057)
作者简介：陈泽宇(2000—　)，男，河南信阳人。硕士，研究方向为渗流力学。E-mail：czy20220222@163.com。
*通信作者：刘丕养(1988—　)，男，山东菏泽人。博士，教授，主要从事油气田开发工程方面的研究。E-mail：Piyang.Liu@qut.edu.cn。

policy sampling process, reducing ineffective exploration and enabling rapid adaptation to the dynamic changes in the reservoir flow field. Historical control experience data was obtained through interaction with the environment and constantly optimized until the optimal injection-production control policy was learned. The proposed algorithm was tested on a two-dimensional reservoir model. The results showed that the oil reservoir injection and production optimization method based on cross-entropy policy optimization in reinforcement learning proposed in this paper demonstrated superior performance in optimization efficiency, enhanced oil recovery, and water control compared to traditional evolutionary algorithms and existing reinforcement learning approaches.

Key words: injection and production optimization; reinforcement learning; Markov Decision Process; the maximum entropy policy exploration mechanism; cross-entropy policy optimization

我国超过80%的原油产量源于水驱油藏[1]，普遍具有强非均衡性特征，主要受沉积岩成岩持续作用，地下岩石各向受力异性强、砂泥岩分布不均，导致储层渗透率差异较大。在长期注水开发过程中会出现“指进-突进”效应：注入水受储层渗透率各向受力异性驱动，优先沿着高渗通道形成优势渗流路径，导致低渗区域原油难以有效驱替，驱替效率普遍低于30%[2-3]。传统油藏流场调控方法虽通过等效均质化处理建立流动方程，并借助物质平衡原理进行开发参数优化，但其经验性回归模型难以精确表征储层非均质场对渗流矢量的空间调制作用，致使开发方案动态预测偏差率通常较低。

考虑到油藏流场的非均衡性，注采优化是提升生产效率和降低生产成本的核心技术手段。这项技术将优化算法、渗流力学原理和数值模拟方法相融合，目的是对地下油藏中的流体运动状态进行有效调控。不过在具体实施过程中，常规的注采优化方法[4-9]存在着明显不足：仅能针对某一特定的油藏流场状态进行调控优化，而每次优化后的结果通常只能得到单一的最优解。当油藏流场状态随着开采进程发生变化时，就需要重新从头开始进行优化训练，这种操作模式不仅花费大量时间，还会影响油田开发的整体效率[10-12]，特别是在处理地质条件复杂的油藏时，这种局限性就显得更加突出。

伴随着人工智能技术的飞速发展，机器学习的方法受到广泛关注，已在多个专业领域取得显著进展，如自然语言处理、医学图像识别、自动驾驶和油藏流场调控生产优化等[13-16]。强化学习(Reinforcement Learning，RL)是一种前沿的机器学习方法，通过让智能体与复杂不确定性的环境进行持续的交互和试错，根据环境的奖惩反馈来进行策略自主学习和优化，从而实现最优决策[17]。2022年，ZHANG等提出了一种基于软演员-评论家算法[18](Soft Actor-Critic，SAC)的油藏全生命周期生产优化方法，实验结果表明该方法能在有限的模拟评估次数下最大化经济净现值并显著增强驱油性能[19]。尽管SAC方法能够利用历史经验数据和最大熵策略探索机制，但在复杂或高维环境中探索效率不高且容易陷入局部最优。

因此，本文聚焦于复杂的水驱油藏地下渗流流场的非均衡性注采优化问题展开研究，通过融合交叉熵方法[20](Cross-Entropy Method)和先进的强化学习算法解决现有方法中策略探索能力不足和历史经验样本利用效率不高等问题，提出一套适应于我国地质特征的复杂水驱油藏注采优化解决方案。

1 油藏注采优化数学模型

油藏注采优化是指通过合理配置油水井不同时间步的注采方案(生产井的产液速率和注水井的注入速率)来改变油藏流场的分布状态，提高油藏波及系数，从而最大化目标函数(如经济净现值或累积产油量)。在实际油田开发过程中，大多采用经济净现值作为油藏注采优化问题的性能指标，反映为在整个油藏生产周期内项目预期经济效益与增产措施所产生的投资成本之间的差额。其定义如下：

$$J(u,v)=\sum_{n=1}^{N_{\text{step}}}\left\{\left[\sum_{j=1}^{N_{\text{pro}}}(r_{\text{o}}\cdot q_{\text{o},j}^{n}-r_{\text{w}}\cdot q_{\text{w},j}^{n})-\sum_{i=1}^{N_{\text{inj}}}(r_{\text{inj}}\cdot q_{\text{inj},i}^{n})\right]\frac{\Delta t^{n}}{(1+b)^{t^{n}/365}}\right\} \tag{1}$$

式中：$J(u,v)$为经济净现值，元；u为井控制变量；v为油藏状态变量；N_{step}为控制步数，步；N_{pro}和N_{inj}

分别为生产井数和注水井数，口；r_o、r_w 和 r_{inj} 分别为油价、产出水处理成本和注入水成本，元/m^3；$q^n_{o,j}$ 和 $q^n_{w,j}$ 分别为第 n 时间步下第 j 口生产井的产油速率和产水速率，m^3/d；$q^n_{inj,i}$ 为第 n 时间步下第 i 口注水井的注水速率，m^3/d；b 为年折现率，%；Δt 和 t^n 分别为第 n 时间步的总时长和已经过的时长，d。

考虑到实际油田生产操作中，必须满足所要求的工程约束条件，以确保注采方案的实施可行性。因此，油藏注采优化数学模型定义如下：

$$\max J(u,v),\quad u\in\Omega \tag{2}$$

受限于：

$$u^{low}\leqslant u\leqslant u^{up} \tag{3}$$

$$g_j(u,v)\leqslant 0,\quad j=1,\cdots,m \tag{4}$$

式中：Ω 为油藏边界内所有活动网格边界；u^{low} 和 u^{up} 分别为井控制变量的下界和上界；$g_j(u,v)$ 为第 j 个非线性状态约束；m 为非线性状态约束总数。

2　基于强化学习的注采优化求解方法

强化学习是一种不同于监督学习和无监督学习方法的机器学习方法，该方法更侧重于研究智能体采取何种动作与环境之间的交互。与传统的监督学习方法不同，强化学习没有明确的输入、输出对应关系，它是通过智能体与环境的交互试错，不断探索并调整自己的决策策略。

2.1　强化学习模型

强化学习的核心思想是智能体与环境的不断交互，依赖环境提供的奖惩反馈机制不断优化策略，使得智能体逐步学习到最优策略，从而最大化累积奖励。强化学习的智能体-环境的反馈框架如图 1 所示。在第 t 时刻，智能体根据当前状态 s_t 选择了一个动作 a_t，并将该动作传递给环境。环境根据该动作改变当前状态，并反馈给智能体新的状态 s_{t+1} 和即时奖励 r_t，智能体根据这个奖励调整其行为策略。在一个回合中，假设智能体会与环境进行 T 次交互，每次交互的信息使用一个四元组 $<s_t,a_t,r_t,s_{t+1}>$ 表示。由于当前时刻的奖励 r_t 与下一个时刻的奖励 r_{t+1} 的重要性不同，为了使智能体能够平衡当前奖励和未来奖励，引入折扣因子 $\gamma\in(0,1)$，得到累积折扣奖励 $R=\sum_{k=1}^{T-t}\gamma^{k-1}r_{t+k}$。智能体的目标是最大化累积折扣奖励，通过与环境交互不断优化行为策略，最后学习得到最优策略 π^*。

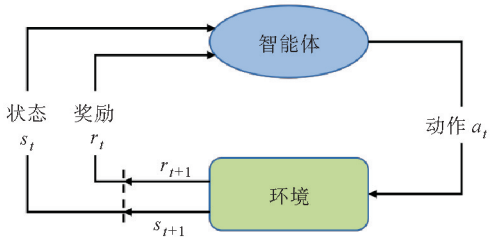


图 1　强化学习的智能体-环境的反馈框架

2.2　马尔可夫决策过程

马尔可夫决策过程[21]（Markov Decision Process，MDP）是强化学习中一个至关重要的理论框架，它通过数学建模的方式，精准地描述了强化学习的学习过程，为理解和解决强化学习问题提供了坚实的理论基础。MDP 通常用一个五元组 $<S,A,P,R,\gamma>$ 表示，其中状态空间 S 表示系统中所有可能的状态集合；动作空间 A 表示系统中所有可能的动作集合；P 是状态转移概率，$P(s_t,a_t,s_{t+1})$ 用来描述在当前状态 s_t 执行动作 a_t 转移到下一个状态 s_{t+1} 的概率；R 是奖励函数 $R:S\times A\to\mathbb{R}$，$R(s_t,a_t)$ 用来描述在当前状态 s_t 下执行动作 a_t 可以获得的即时奖励期望；γ 是奖励的折扣因子。

在 MDP 中，价值函数是评估一个策略 π 好坏的标准，它反映了从当前状态出发，智能体按照某个策略能够获得的期望总回报。价值函数分为两类：状态价值函数（状态值函数）和状态-动作价值函数（Q 值函数）。在某一时刻 t，状态值函数表示从状态 s_t 开始，遵循策略 π 执行所有动作后所获得的期望累积奖励；Q 值函数表示在状态 s_t 下采取动作 a_t 后，继续按照策略 π 执行所有动作后所获得的期望累积回报。其形式如下：

$$V_\pi(s_t)=E_\pi\left[\sum_{k=1}^{\infty}\gamma^{k-1}r_{t+k}\mid s_t\in S\right] \tag{5}$$

$$Q_\pi(s_t,a_t)=E_\pi\left[\sum_{k=1}^{\infty}\gamma^{k-1}r_{t+k}\mid s_t\in S,a_t\in A\right] \tag{6}$$

式中：$V_\pi(s_t)$ 为在状态 s_t 下的状态值函数；$Q_\pi(s_t)$ 为在状态 s_t 下执行动作 a_t 的 Q 值函数；$E_\pi[\cdot]$ 为在策略 π 上的数学期望。

根据贝尔曼方程，二者之间的转换关系如下：

$$V_\pi(s_t)=\sum_{a_t}\pi(a_t\mid s_t)Q_\pi(s_t,a_t) \tag{7}$$

$$Q_\pi(s_t,a_t)=r(s_{t+1},a_{t+1})+\gamma E_{s_{t+1}\sim P}[V_\pi(s_{t+1})] \tag{8}$$

式中：$\pi(a_t\mid s_t)$ 为策略 π 在状态 s_t 下选择动作 a_t 的概率；$r(s_{t+1},a_{t+1})$ 为在状态 s_{t+1} 下采取动作 a_{t+1} 获得的即时奖励。

在 MDP 中智能体的目标是找到一个最优策略 π^*，使得智能体在一个回合中从任意状态 s_t 开始都能获得最大累积奖励。最优策略 π^* 可以通过状态值函数或 Q 值函数进行表示：

$$\pi^*=\underset{\pi}{\operatorname{argmax}}\,V_\pi(s_t)=\underset{\pi}{\operatorname{argmax}}\,Q_\pi(s_t,a_t) \tag{9}$$

2.3 注采优化问题建模

为了在油藏注采优化问题上更有效地运用强化学习算法，本文采用了 MDP 框架对这一问题进行建模，将其转化为一个序列决策过程。在该过程中，智能体被认为是一个控制器，能够根据当前油藏流场状态实时调整注采调控方案，而环境和状态转移过程通过油藏数值模拟器来描述，提供实时的油藏流场状态更新和相应的奖励反馈信息。油藏注采优化问题中的马尔可夫决策过程如图 2 所示。

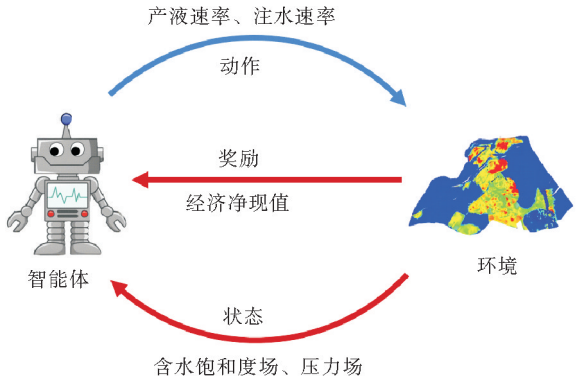


图 2 油藏注采优化问题中的马尔可夫决策过程

2.3.1 状态空间

在状态空间设计时，应选择能够充分反映系统动态变化的变量。在油藏流场调控优化问题中，通常选择含水饱和度场和压力场作为主要的观测状态变量。这些变量不仅能够反映油藏流场在不同调控时间步的动态变化情况，还能够为优化决策提供丰富的特征信息。t 时刻的状态 s_t 可定义为油藏流场模型中各个网格点的含水饱和度值和压力值：

$$s_t=[w_{t,1},w_{t,2},\cdots,w_{t,n};p_{t,1},p_{t,2},\cdots,p_{t,n}] \tag{10}$$

式中：$w_{t,i}$，$p_{t,i}$ 分别为第 i 个网格在 t 时刻的含水饱和度值和压力值。

此外，考虑到含水饱和度值一般在[0,1]，压力值通常在数千数量级之间，直接使用原始数据可能会对神经网络的训练效果产生影响，因此使用 max-min 归一化技术将压力场数据调整至统一区间[0,1]：

$$f_{\text{max-min}}(p_{t,i})=\frac{p_{t,i}-p_{\min}}{p_{\max}-p_{\min}} \tag{11}$$

式中：$f_{\text{max-min}}(p_{t,i})$ 为第 t 时刻第 i 个网格归一化后的压力值；$p_{\min}$、$p_{\max}$ 分别为油藏的最小压力值和最大压力值。

2.3.2 动作空间

动作空间作为 MDP 框架中的一个重要组成部分，定义了智能体在每个状态下可以采取的所有可能动作集合。动作空间的设计需要紧密结合具体问题的特征，才能保证智能体能够在状态空间中进行有效地探索。在油藏流场调控优化问题中，通常选择生产井的产液速率和注水井的注水速率作为动作调控变量。t 时刻的动作 a_t 可定义为

$$a_t=[l_{t,1}^{\text{pro}},l_{t,2}^{\text{pro}},\cdots,l_{t,j}^{\text{pro}},\cdots,l_{t,N_{\text{pro}}}^{\text{pro}};w_{t,1}^{\text{inj}},w_{t,2}^{\text{inj}},\cdots,w_{t,i}^{\text{inj}},\cdots,w_{t,N_{\text{inj}}}^{\text{inj}}] \tag{12}$$

式中：$l_{t,j}^{\text{pro}}$、$w_{t,i}^{\text{inj}}$ 分别为在 t 时刻的第 j 口生产井的产液速率和第 i 口注水井的注水速率。

2.3.3 奖励函数

奖励函数是智能体在学习策略过程中的核心部分，它通过反馈智能体在每个时间步采取的动作所带

来的奖励或惩罚，帮助智能体在多次决策中逐步学习到最优策略。在油藏流场调控优化问题中，目标是在整个生命周期内最大化经济净现值。因此在 t 时刻的奖励 r_t 根据目标函数确定，可定义为

$$r_t=\left[\sum_{j=1}^{N_{\mathrm{pro}}}(r_{\mathrm{o}}\cdot q_{\mathrm{o},j}^{t}-r_{\mathrm{w}}\cdot q_{\mathrm{w},j}^{t})-\sum_{i=1}^{N_{\mathrm{inj}}}(r_{\mathrm{inj}}\cdot q_{\mathrm{inj},i})\right]\cdot\Delta t^{n} \tag{13}$$

3　交叉熵策略优化算法构建

3.1　算法原理

基于交叉熵策略优化的软演员-评论家算法[22]（Soft Actor-Critic with Cross-Entropy Policy Optimization，SAC-CEPO）是在 SAC 算法基础上进行改进的无模型深度强化学习算法，与 SAC 算法相似，SAC-CEPO 算法的评估模块仍然保留双 Q 值网络结构设计，即 2 个独立且网络结构相似的 Q 值网络 $Q_{\theta_1}(s_t,a_t)$ 和 $Q_{\theta_2}(s_t,a_t)$。但决策模块略有不同，将 SAC 原始高斯策略网络解耦为 2 个独立的网络，即均值网络 $\pi_{\phi_1}^{\mu}(s_t)$ 和标准差网络 $\pi_{\phi_2}^{\sigma}(s_t)$，$\phi_1$ 和 ϕ_2 为各自网络参数。通过对策略网络解耦，能够消除策略参数更新时的梯度耦合效应，使得均值网络专注于策略收敛，标准差网络动态适配环境的不确定性，从而保证策略稳定性的同时提升策略探索效率。其中均值网络输出高斯分布中的均值 μ，通过交叉熵方法优化；而标准差网络输出标准差 σ，通过梯度下降方法更新。两者协同作用重构原始策略：

$$\pi_{\phi}(a_t\mid s_t)=N(\mu,\sigma) \tag{14}$$

式中：$\pi_{\phi}(a_t\mid s_t)$ 为策略，表示在状态 s_t 下选择动作 a_t 的概率分布；$N(\mu,\sigma)$ 为高斯分布，其中，μ 为均值网络 $\pi_{\phi_1}^{\mu}(s_t)$ 输出值，σ 为标准差网络 $\pi_{\phi_2}^{\sigma}(s_t)$ 输出值。

在训练更新过程中，从经验回放缓冲区 D 中随机小批量采集样本数据对策略和 Q 值网络进行迭代更新训练。通过最小化柔性贝尔曼残差来更新 Q 值网络的参数，公式如下：

$$Q_{\theta}(s_t,a_t)=E_{\substack{(s_t,a_t,r_t,s_{t+1})\sim D,\\ a_{t+1}\sim\pi_{\phi}(\cdot\mid s_{t+1})}}\left[r_t+\gamma\left(\min_{j=1,2}Q_{\theta_j}(s_{t+1},a_{t+1})-\alpha\log\pi_{\phi}(a_{t+1}\mid s_{t+1})\right)\right] \tag{15}$$

$$L_Q(\theta_j)=E_{(s_t,a_t)\sim D}\left[\frac{1}{2}\left(Q_{\theta_j}(s_t,a_t)-Q_{\theta}(s_t,a_t)\right)^2\right],\quad j=1,2 \tag{16}$$

式中：$Q_{\theta}(s_t,a_t)$ 为在状态 s_t 下执行动作 a_t 的目标估计 Q 值；$Q_{\theta_j}(s_{t+1},a_{t+1})$ 为第 j 个目标 Q 值网络中，在状态 s_{t+1} 下执行动作 a_{t+1} 的目标估计 Q 值；$\min\limits_{j=1,2}Q_{\theta_j}(s_{t+1},a_{t+1})$ 为选取 2 个目标 Q 值网络中最小的 Q 值来计算 $Q_{\theta}(s_t,a_t)$，有效避免过大偏差估计；α 为温度参数；$L_Q(\theta_j)$ 为第 j 个 Q 值网络的损失函数，通过最小化该损失函数来训练 Q 值网络参数 θ_j。

为了保证训练的稳定性，目标 Q 值网络以软更新的方式进行。目标 Q 值网络的更新公式如下：

$$\bar{\theta}_j\leftarrow\tau\theta_j+(1-\tau)\bar{\theta}_j \tag{17}$$

策略网络的更新过程中，算法严格遵循先优化均值网络，再调整标准差网络的顺序。

对于均值网络的优化，首先通过交叉熵方法从当前样本均值和标准差的正态分布中采样候选均值参数样本 $\{\mu_1,\cdots,\mu_{N_{\mathrm{cem}}}\}$，接着对每个候选均值参数样本进行评估：

$$F(\mu_i)=E_{s_t\sim D}\left[\alpha\log N(a_t;\mu_i,\pi_{\phi_2}^{\sigma}(s_t))-\min_{j=1,2}Q_{\theta_j}(s_t,a_t)\right],\quad i=1,\cdots,N_{\mathrm{cem}} \tag{18}$$

式中：$F(\mu_i)$ 为每个均值样本参数的评估值；$N(a_t;\mu_i,\pi_{\phi_2}^{\sigma}(s_t))$ 为动作 a_t 服从均值为 μ_i、标准差为 $\pi_{\phi_2}^{\sigma}(s_t)$ 的正态分布，其中动作 a_t 由重参数化生成：

$$a_t=\mu_i+\varepsilon_t\mathrm{e}^{\pi_{\phi_2}^{\sigma}(s_t)},\quad \varepsilon_t\sim N(0,1) \tag{19}$$

将评估结果 $F(\mu_i)$ 降序排序，从中选取前 $N_{\mathrm{cem}}\times\rho_{\mathrm{cem}}$ 个作为精英均值参数样本 $\{\mu_1^{*},\cdots,\mu_{N_{\mathrm{cem}}\times\rho_{\mathrm{cem}}}^{*}\}$，计算其均值和标准差并更新高斯分布。通过上述过程不断迭代优化得到最佳均值 μ^{*}。使用最小化当前均值网络输出的均值 $\pi_{\phi_1}^{\mu}(s_t)$ 与 μ^{*} 之间的均方误差来更新均值网络的参数。均值网络损失函数 $L_{\mu}(\phi_1)$ 可表示为

$$L_{\mu}(\phi_1)=E_{s_t\sim D}\left[\frac{1}{2}(\pi^{\mu}_{\phi_1}(s_t)-\mu^{*})^2\right] \tag{20}$$

在均值网络更新后，对于标准差网络的更新则通过梯度下降法调整网络参数 ϕ_2，目标是最小化 KL 散度(Kullback-Leibler Divergence)，其标准差网络损失函数 $L_{\sigma}(\phi_2)$ 可表示为

$$L_{\sigma}(\phi_2)=E_{s_t\sim D}\left[\alpha\log N(a_t;\pi^{\mu}_{\phi_1},\pi^{\sigma}_{\phi_2})-\min_{j=1,2}Q_{\theta_j}(s_t,a_t)\right] \tag{21}$$

此外，为增强 SAC-CEPO 算法鲁棒性，对于最终动作的输出通过 tanh 函数限制在[−1,1]有限区间内：

$$a_t=\pi^{\mu}_{\phi_1}+\varepsilon_t e^{\pi^{\sigma}_{\phi_2}(s_t)},\quad \varepsilon_t\sim N(0,1) \tag{22}$$

3.2 算法步骤

智能体首先通过与油藏流场环境持续交互收集历史调控经验数据，并保存到经验回放缓冲区 D 中。然后在训练过程中，SAC-CEPO 通过小批量随机采样得到的历史调控经验分别对均值网络、标准差网络、Q 值网络和目标 Q 值网络进行迭代更新。

基于 SAC-CEPO 的油藏注采优化的算法优化过程如图 3 所示。

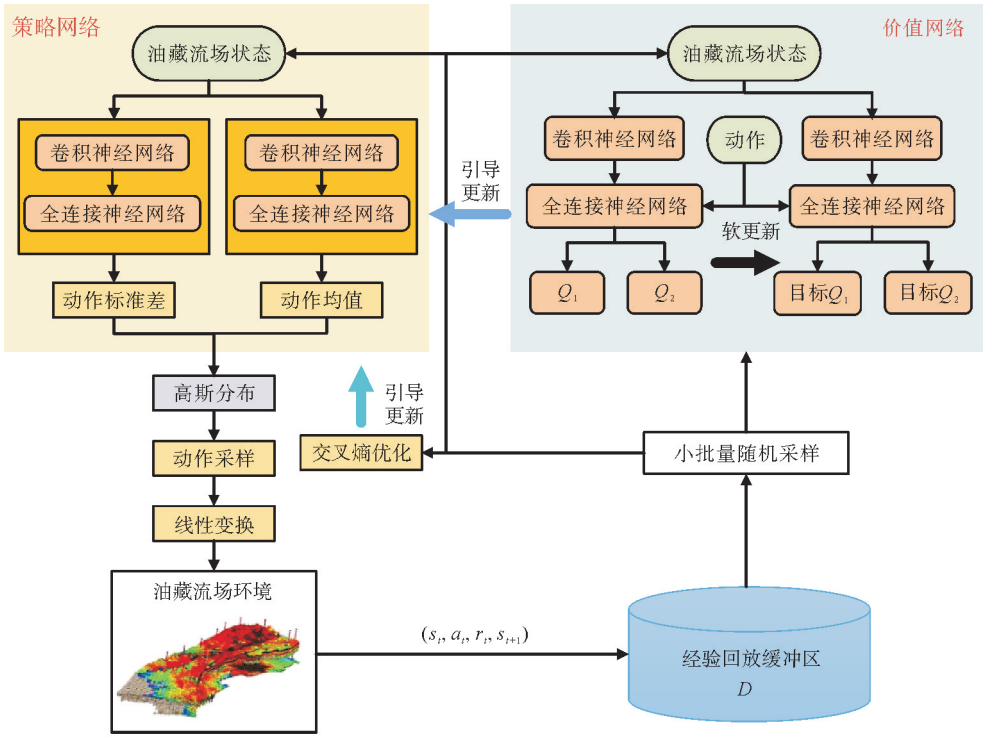


图 3 基于 SAC-CEPO 算法的注采优化流程

4 实例测试

为了测试所提出的基于 SAC-CEPO 的油藏注采优化方法的性能，将其与粒子群算法(Particle Swarm Optimization，PSO)、进化算法(Differential Evolution Algorithm，DE)以及强化学习主流的 SAC 方法在二维油藏模型实例中进行性能对比测试。PSO 方法的种群大小 N 和进化代数 G 分别设置为 40 和 100，初始惯性权重 ω_{ini} 和结束惯性权重 ω_{end} 分别设为 0.7 和 0.1。DE 方法的种群大小 N 和进化代数 G 分别设置为 100 和 40，交叉概率 p_{c} 和变异概率 p_{m} 分别设为 0.1 和 0.5。SAC 和 SAC-CEPO 方法的学习率设为 3×10^{-4}，批量大小为 128。在优化过程中，使用油藏数值模拟器 eclipse2020 进行数值模拟交互。

本案例的二维油藏模型是一个具有 3 个渗透通道的二维单层油藏流场模型，该模型由 25×25×1 个

网格组成，其中包含 4 口注水井和 9 口生产井并以五点法井网形式开发生产。该油藏模型的渗透率场和井位分布如图 4 所示。

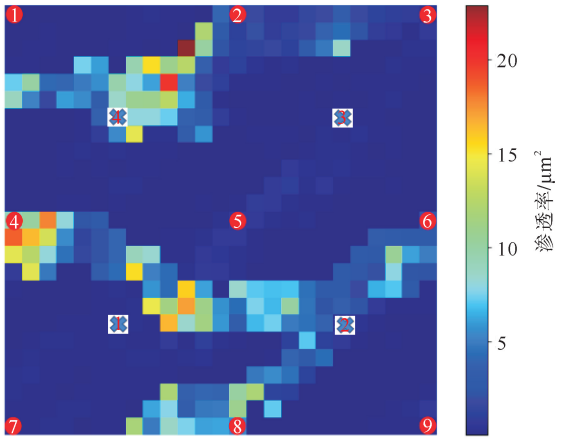


图 4　二维油藏模型渗透率场和井位分布

●生产井(数字表示井号)；✖注水井(数字表示井号)

在油藏生产优化过程中，优化时长为 1800 d，每 180 d 调控 1 次，共 10 个调控时间步。在每个时间步，对生产井的产液速率和注水井的注水速率进行优化调控，因此调控变量总数为 10×13＝130 个。设定每口生产井的产液速率上限为 300 m^3/d，下限为 0 m^3/d。每口注水井的注水速率上限为 700 m^3/d，下限为 0 m^3/d。油价为 80 元/m^3，产出水处理成本费用为 5 元/m^3，注水成本费用为 0 元/m^3，年折现率设为 0。

采用 DE、PSO、SAC 以及 SAC-CEPO 算法对该模型进行优化并对优化结果进行对比分析。为了保证实验结果的准确性，所有方法均在 5 个相同随机种子下独立运行 4000 个回合，1 个回合指的是一个完整的油藏生产周期。图 5 展示了各方法在 5 次独立运行后平均经济净现值随数值模拟运行回合的变化情况。从图 5 中可以看出，SAC-CEPO 算法可以在较少的数值模拟回合内快速达到较高的经济净现值，并在后期保持高效稳定，其全局优化能力和收敛效率在各方法中表现最优。而 SAC 也表现出较强的优化能力，但结果稍逊于 SAC-CEPO 方法。相比之下，DE 和 PSO 方法通过随机搜索的方法寻找最优解，但由于无法利用历史调控经验，其优化效率显著偏低，最终陷入局部最优。

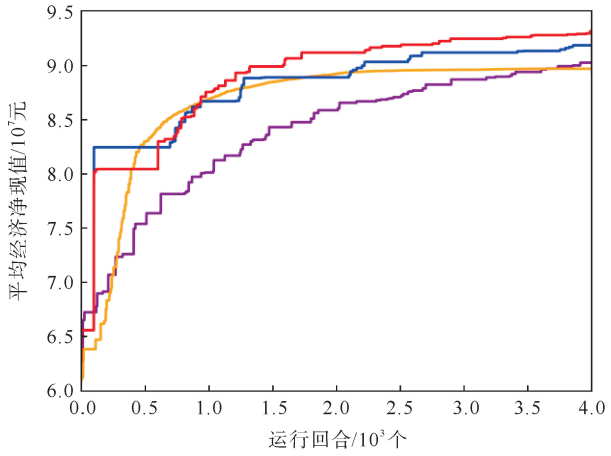
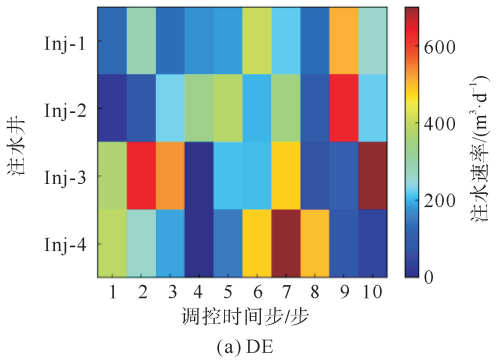


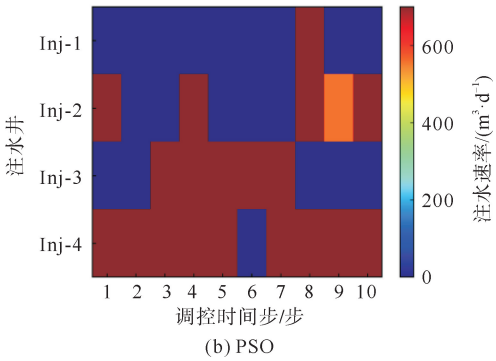
图 5　平均经济净现值随运行回合的优化曲线

DE；PSO；SAC；SAC-CEPO

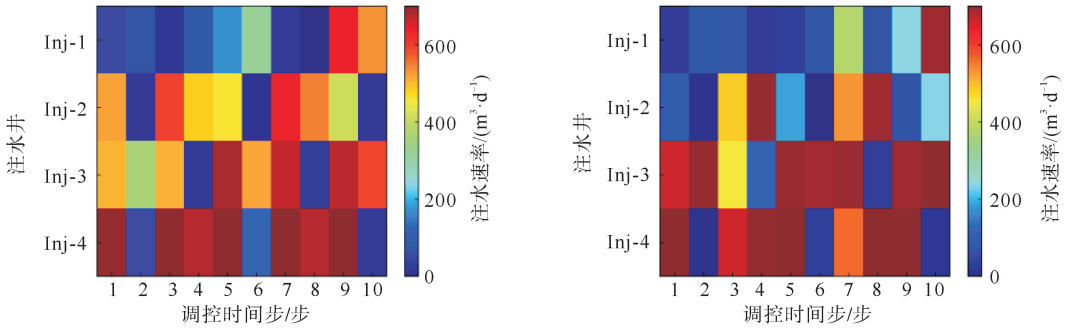
图 6 和图 7 分别显示了 DE、PSO、SAC 以及 SAC-CEPO 方法经过优化后得到的注水井和生产井的调控方案。图 8 显示了各方法经过调控方案运行后，累积产油量和累积产水量随时间变化的曲线，可以看出，SAC-CEPO 方法实现了更好的增油控水效果。图 9 显示了各方法优化后的剩余油分布，可见 SAC-CEPO 驱油效果更佳。



(a) DE

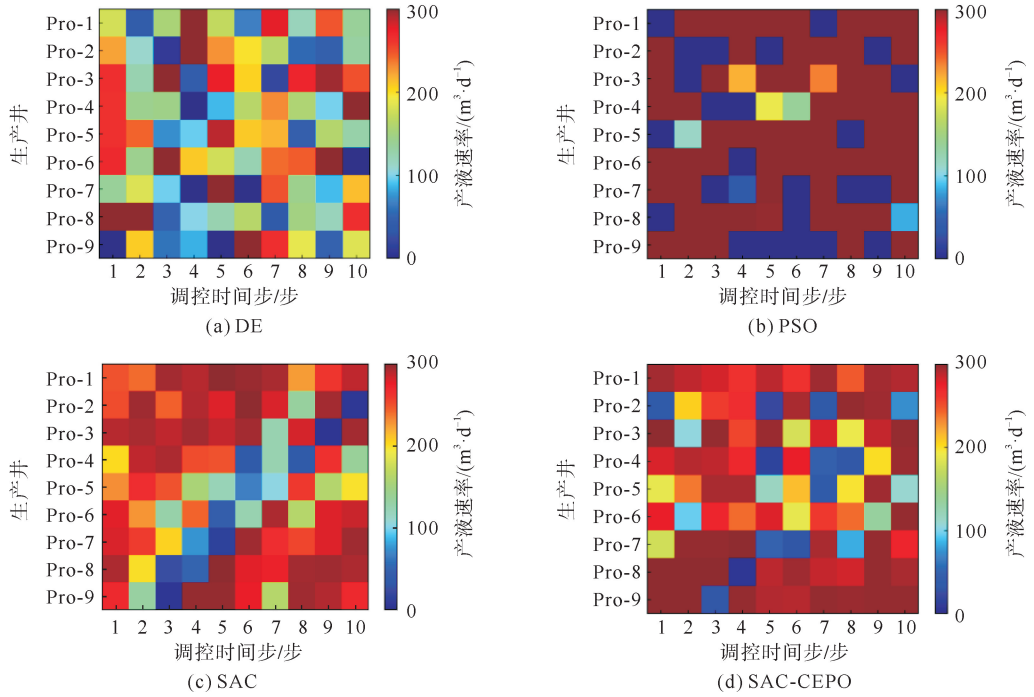


(b) PSO



(c) SAC (d) SAC-CEPO

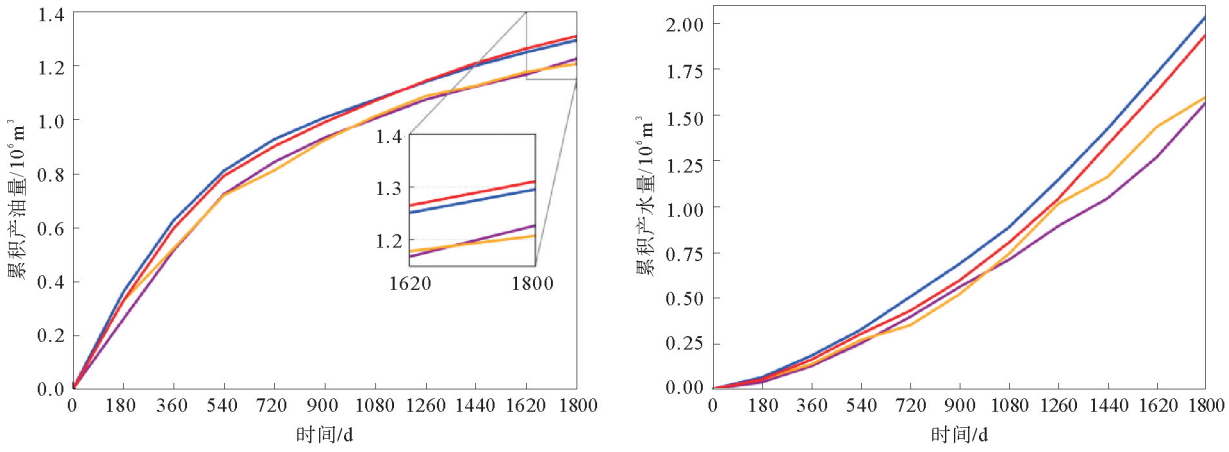
图 6 各种方法的注水井最优调控方案



(a) DE (b) PSO

(c) SAC (d) SAC-CEPO

图 7 各种方法的生产井最优调控方案



(a) 累积产油量随时间变化曲线 (b) 累积产水量随时间变化曲线

图 8 各种方法提供的优化结果对比

DE; PSO; SAC; SAC-CEPO

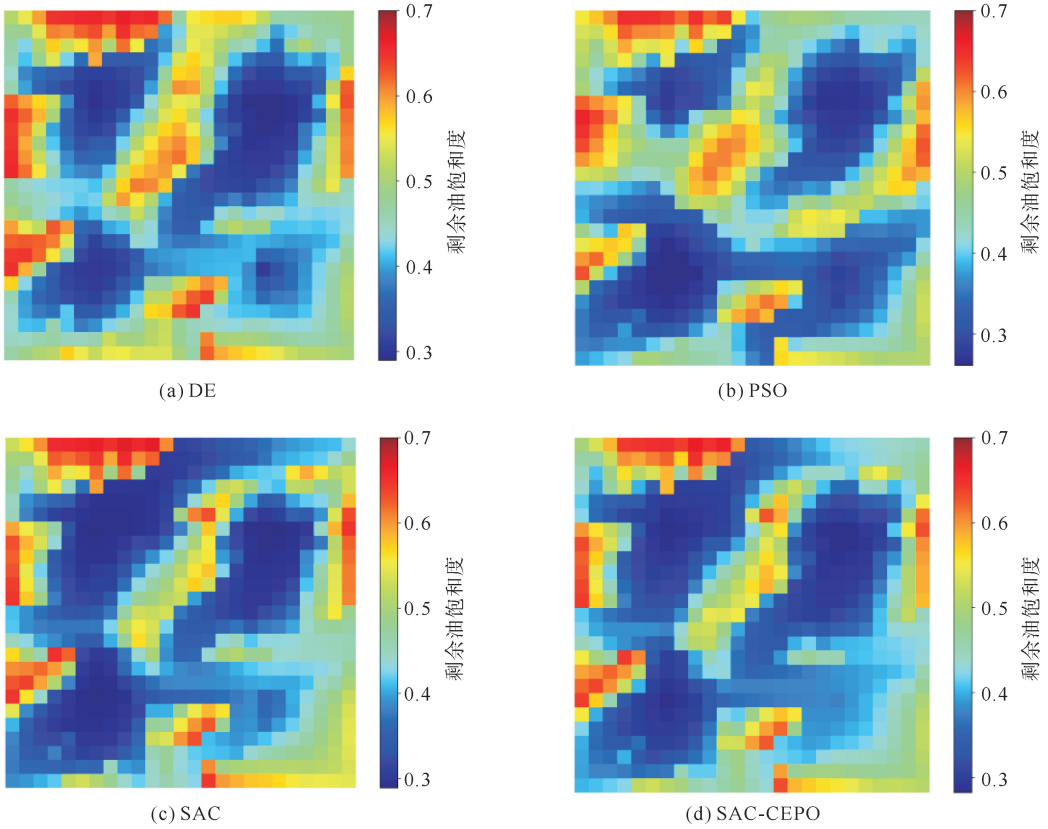


图 9　各优化方案的剩余油分布

5　结论

针对现有方法中策略探索能力不足和历史经验样本利用效率不高等问题，提出了一种基于交叉熵策略优化的强化学习油藏注采优化方法。

1）引入最大熵策略探索机制，提高了策略探索的多样性以避免过早陷入局部最优。

2）通过融合交叉熵方法对策略采样过程进行优化，减少无效探索，从而快速适应油藏流场的动态变化。同时利用与环境交互获得历史经验数据，不断迭代优化直至学会最优的注采优化策略。训练后的策略具备目标油藏流场各个开发过程中的流场调控优化经验，可以直接离线应用于目标油藏流场模型中，能根据目标油藏流场状态变化实时提供相应的注采优化方案，无需重新训练。

3）将所提出的方法应用到二维油藏流场模型中，并与现有的优化方法 DE、PSO、SAC 进行对比分析。实验结果表明，所提出的方法策略探索能力更强，能够获得更高的经济净现值和较好的增油控水效果。

参考文献(References)：

[1] 刘合，裴晓含，罗凯，等. 中国油气田开发分层注水工艺技术现状与发展趋势[J]. 石油勘探与开发，2013，40(6)：733-737.
LIU He，PEI Xiaohan，LUO Kai，et al. Current status and trends of separated layer water flooding in China[J]. Petroleum Exploration and Development，2013，40(6)：733-737.

[2] 韩大匡. 中国油气田开发现状，面临的挑战和技术发展方向[J]. 中国工程科学，2010，12(5)：51-57.
HAN Dakuang. Status and challenges for oil and gas field development in China and directions for the development of corresponding technologies[J]. Strategic Study of CAE，2010，12(5)：51-57.

[3] 李阳. 陆相高含水油藏提高水驱采收率实践[J]. 石油学报，2009，30(3)：396-399.
LI Yang. Study on enhancing oil recovery of continental reservoir by water drive technology[J]. Acta Petrolei Sinica，2009，30(3)：396-399.

[4] CHEN Y，OLIVER D S，ZHANG D. Efficient ensemble-based closed-loop production optimization[J]. SPE Journal，2009，14(4)：634-

645.

[5] REGIS R G. Evolutionary programming for high-dimensional constrained expensive black-box optimization using radial basis functions[J]. IEEE Transactions on Evolutionary Computation,2014,18(3):326-347.

[6] BEYKAL B,BOUKOUVALA F,FLOUDAS C A,et al. Global optimization of grey-box computational systems using surrogate functions and application to highly constrained oil-field operations[J]. Computers & Chemical Engineering,2018,114:99-110.

[7] FOROUD T,BARADARAN A,SEIFI A. A comparative evaluation of global search algorithms in black box optimization of oil production:A case study on Brugge field[J]. Journal of Petroleum Science and Engineering,2018,167:131-151.

[8] OGUNTOLA M B,LORENTZEN R J. Ensemble-based constrained optimization using an exterior penalty method[J]. Journal of Petroleum Science and Engineering,2021,207:109165.

[9] ZHANG K,ZHAO X G,CHEN G D,et al. A double-model differential evolution for constrained waterflooding production optimization[J]. Journal of Petroleum Science and Engineering,2021,207:109059.

[10] 李培宇,黄世军,柴世超,等.断块油藏水平井提液参数优化[J].科学技术与工程,2022,22(2):524-531.
LI Peiyu,HUANG Shijun,CHAI Shichao,et al. Optimization of horizontal well's fluid extraction parameters in fault block reservoirs[J]. Science Technology and Engineering,2022,22(2):524-531.

[11] 张凯,陈国栋,薛小明,等.基于主成分分析和代理模型的油藏生产注采优化方法[J].中国石油大学学报(自然科学版),2020,44(3):90-97.
ZHANG Kai,CHEN Guodong,XUE Xiaoming,et al. A reservoir production optimization method based on principal component analysis and surrogate model[J]. Journal of China University of Petroleum(Edition of Natural Sciences),2020,44(3):90-97.

[12] 张凯,赵兴刚,张黎明,等.智能油田开发中的大数据及智能优化理论和方法研究现状及展望[J].中国石油大学学报(自然科学版),2020,44(4):28-38.
ZHANG Kai,ZHAO Xinggang,ZHANG Liming,et al. Current status and prospects for the research and application of big data and intelligent optimization methods in oilfield development[J]. Journal of China University of Petroleum(Edition of Natural Sciences),2020,44(4):28-38.

[13] 奚雪峰,周国栋.面向自然语言处理的深度学习研究[J].自动化学报,2016,42(10):1445-1465.
XI Xuefeng,ZHOU Guodong. A survey on deep learning for natural language processing[J]. Acta Automatica Sinica,2016,42(10):1445-1465.

[14] 兰欣,卫荣,蔡宏伟,等.机器学习算法在医疗领域中的应用[J].医疗卫生装备,2019,40(3):93-97.
LAN Xin,WEI Rong,CAI Hongwei,et al. Application of machine learning algorithms in the medical field[J]. Chinese Medical Equipment Journal,2019,40(3):93-97.

[15] 王文东,石梦翩,庄新宇,等.基于机器学习的井位及注采参数联合优化方法[J].深圳大学学报(理工版),2022,39(2):126-133.
WANG Wendong,SHI Menghe,ZHUANG Xinyu,et al. Joint optimization method of well location and injection-production parameters based on machine learning[J]. Journal of Shenzhen University(Science & Engineering),2022,39(2):126-133.

[16] TANG X L,HUANG B,LIU T,et al. Highway decision-making and motion planning for autonomous driving via soft actor-critic[J]. IEEE Transactions on Vehicular Technology,2022,71(5):4706-4717.

[17] 李茹杨,彭慧民,李仁刚,等.强化学习算法与应用综述[J].计算机系统应用,2020,29(12):13-25.
LI Ruyang,PENG Huimin,LI Rengang,et al. Overview on algorithms and applications for reinforcement learning[J]. Computer Systems & Applications,2020,29(12):13-25.

[18] HAARNOJA T,ZHOU A,HARTIKAINEN K,et al. Soft actor-critic algorithms and applications[J]. arXiv Preprint,2018:1812.05905. DOI:10.48550/arXiv.1812.05905.

[19] ZHANG K,WANG Z Z,CHEN G D,et al. Training effective deep reinforcement learning agents for real-time life-cycle production optimization[J]. Journal of Petroleum Science and Engineering,2022,208:109766.

[20] BOER P T D,KROESE D P,MANNOR S,et al. A tutorial on the cross-entropy method[J]. Annals of Operations Research,2005,134:19-67.

[21] PUTERMAN M L. Markov decision processes[J]. Handbooks in Operations Research and Management Science,1990,2:331-434.

[22] SHI Z,SINGH S P N. Soft actor-critic with cross-entropy policy optimization[J]. arXiv Preprint,2021:2112.11115. DOI:10.48550/arXiv.2112.11115.

(责任编辑 赵金环;英文校审 徐 飞)