/TemplateVersion (2026.1)

基于 TLE 的 A2C 代理用于地面覆盖轨道路径规划

Anantha Narayanan[1], Battu Bhanu Teja[2], Pruthwik Mishra[1]

[1] Sardar Vallabhbhai National Institute of Technology Surat, [2]CAMS
u23cs088@coed.svnit.ac.in, bhanubattu@alum.iisc.ac.in, pruthwikmishra@aid.svnit.ac.in

Abstract

低地球轨道（LEO）日益拥堵给地球观测卫星的高效部署和安全运行带来了持续的挑战。任务规划者现在不仅需要考虑任务特定的要求，还需要考虑与现役卫星和太空碎片碰撞风险的增加。本文提出了一种使用优势演员评论家（A2C）算法的强化学习框架，以优化卫星轨道参数，实现预定义地表半径内的精确地面覆盖。通过在定制的 OpenAI Gymnasium 环境中将问题表述为马尔可夫决策过程（MDP），我们的方法使用经典的开普勒元素模拟轨道动力学。代理逐步学习调整五个轨道参数——半长轴、偏心率、倾角、升交点赤经和近地点幅角——以实现目标地面覆盖。与近端策略优化（PPO）的比较评估显示，A2C 的性能更佳，累计奖励提高 5.8 倍（10.0 对 9.263025），且收敛时间步减少 31.5 倍（2,000 对 63,000）。A2C 代理在不同目标坐标上始终满足任务目标，同时保持适合实时任务规划应用的计算效率。关键贡献包括：（1）一个基于 TLE 的轨道仿真环境，结合物理约束，（2）验证演员评论家方法在连续轨道控制中优于信任区域方法的卓越性，（3）展示快速收敛以实现自适应卫星部署。这种方法确立了强化学习作为可扩展和智能 LEO 任务规划的计算高效替代方案。

we will release the data and code soon.

近年来，空间兴趣呈多方面增长，发射了数百个航天器以满足监视、地球观测、连接和其他需求。虽然这开启了新的前沿，但也产生了一个独特的问题，即轨道拥挤。任务规划现在必须考虑现有的轨道、碰撞路线和太空垃圾，因此需要能够适应动态需求和多样化目标的复杂系统。传统的轨道设计方法依赖于计算开销高的优化技术 (Song et al. 2018) 、解析近似 (Savitri et al. 2017) 或难以应对动态环境和实时约束的启发式规划方法 (Mok et al. 2019) 。因此，优化轨道参数的新方法对于解决这些挑战是至关重要的。

强化学习（RL），尤其是在解决复杂的序列问题方面，已经成为一种范式，处于传统方法局限性的前沿。将深度强化学习应用于轨道力学 (Kyuroson et al. 2024) 为开发自适应和智能卫星控制系统提供了独特的机会。

这项工作通过将其表述为一个马尔可夫决策过程 (MDP) 问题，解决了低地球轨道 (LEO) 中的目标观测卫星的轨道规划挑战，并采用优势演员评论算法，预测最佳参数。这种方法展示了现代强化学习算法在复杂轨道动力学问题和卫星任务规划中的实用性，强调：

- 一种用于卫星轨道优化的新型 MDP 公式
- 一个使用开普勒元素的自定义轨道力学模拟环境，以及
- 展示 A2C 在学习稳定轨道策略方面的有效性。

问题表述

轨道要素

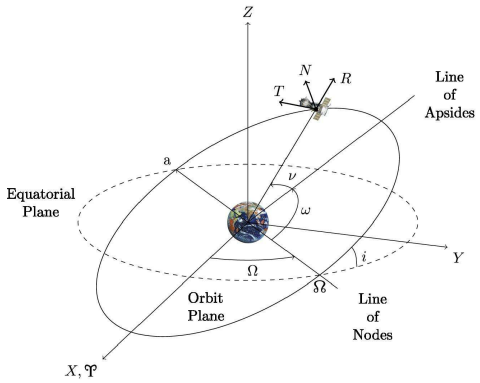

Figure 1: 本研究中使用的轨道要素示意。 (Tafanidis et al. 2025)

空间中的每一个轨道都由其一组经典开普勒轨道元件唯一定义。这些参数描述了轨道的大小、形状和方向，以及在给定时间内卫星沿轨道的位置信息。为了进行轨道优化，在此工作中，我们关注以下五个元素，它们足以确定相对于中心天体（如地球）的轨道路径：半长轴、偏心率、倾角、升交点赤经和近地点幅角。真实近点角（υ）不纳入优化范围，因为它表示的是卫星的瞬时位置而不是固定的轨道特性。

对于 LEO 任务，这些参数受到操作约束（例如，高度范围）和物理可行性（例如，避免碰撞，满足重访要求）的限制。在这项工作中，轨道要素被视为可以优化的行动参数，以最大化卫星通过预定义阈值到达地面目标的能力。我们采用基于学习的方法来动态调整这些元素，而不是依赖于固定值或手动调整。这使卫星能够发现满足覆盖目标且符合物理可行性的轨道配置。

为了促进这种基于学习的轨道优化，我们设计了一个基于 OpenAI Gymnasium (Towers et al. 2024) 的自定义仿真环境。

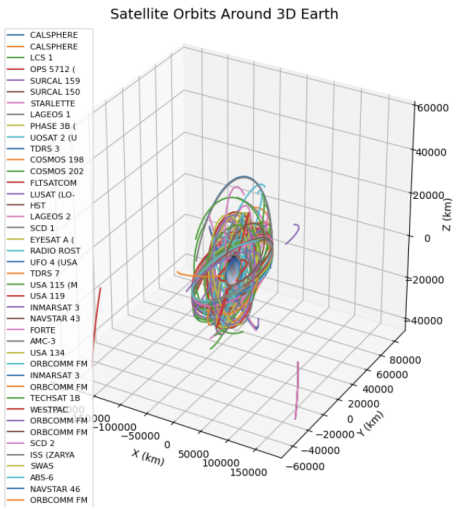

Figure 2: 基于 TLE 的 OpenAI Gymnasium 环境示例。

该环境允许使用其动作和观测空间进行动态评估和测试，结合了轨道力学。OpenAI Gymnasium API 规定了功能标准，以实现与 Stable Baselines3 的兼容性和跨平台使用。

为了确保现实性、可扩展性和任务特定的相关性，我们使用受限于公共获取的地球轨道人造卫星双行元数据（TLE）的参数来初始化轨道要素。TLE 是一种标准化格式，用于编码卫星的轨道参数，由像 Celestrak 和 NORAD 这样的组织维护并定期更新。这些数据反映了正在运行的卫星的实际状态，因此非常适合用来定义我们的仿真环境中合理的轨道构型。

虽然 TLE 更新仅反映了卫星在其轨道路径中的瞬时位置，但我们的仿真框架侧重于整个轨道配置，而不是卫星在其轨道内的实时相对位置。这减少了每个仿真步骤中持续检索和验证 TLE 的需求。相反，我们提取一次轨道参数，重建完整轨道，并使用这种静态表示来实施任务约束。此方法在保持实时任务相关性的同时，将更新的开销降到最低。

TLE 格式与解读 一个典型的 TLE 由两行字母数字字符组成，这些字符编码了各种开普勒元素和卫星识别数据。以下是 ISS (ZARYA) 的 TLE 数据的示例：

```
ISS (ZARYA)
1 25544U 98067A 24146.63752315 .00009537 00000+0 17465-3
0 9994
2  25544  51.6422  41.9330  0005197  351.2436  8.8447
15.50954063448027
```

TLE 格式对轨道参数的编码如下：
第 2 行：

- 倾角：$i = 51.6422°$
- 升交点赤经 (RAAN): $\Omega = 41.9330°$
- 偏心率：$e = 0.0005197$
- 近地点幅角：$\omega = 351.2436°$
- 平均近点角：$M = 8.8447°$
- 平均运动：$n = 15.50954063$ 转/天
- 历元时的圈数：448027

半长轴（a）是通过使用公式

$$a = \sqrt[3]{\frac{\mu}{n^2}}$$

由平均运动（n）得出的，其中 μ 是地球的引力参数（约为 398,600 km^3/s^2）。对于国际空间站 TLE，$n = 15.50954063\ rev/day$ 产生 $a \approx 6,792\ km$，对应于地表上方约 414 公里的高度。这些从 TLE 导出的界限为六个经典开普勒轨道要素的允许范围提供了信息：半长轴、偏心率、倾角、升交点赤经（RAAN）、近地点幅角和平近点角。通过将初始化限制在这些范围内，我们确保物理上可行且有效的学习状态。

在训练期间，每个轨道配置在一次情节开始或策略重置时随机初始化在这些范围内。用于 LEO 的 5 个开普勒元素的典型范围是：这种初始化确保了训练样本的多样性，防止过拟合到狭窄的轨道范围，并避免学习过程中次优收敛或停滞。此外，它使得可以在更广泛的轨道配置场景中进行泛化，而不是固定的轨迹。这种方法使得学习到的策略在实际部署和现实任务规划中更具适应性和准备性。

动作空间 基于 PPO 的智能体对修改五个轨道参数以实现最佳配置具有决定性控制。

- 半长轴（a）
- 偏心率（e）
- 倾角（i）
- 升交点赤经（Ω）
- 近地点幅角（ω）

这些参数构成一个连续的动作空间，其中每个元素都受限于低地球轨道（LEO）的物理有效限制。在训练过程中，PPO 智能体对这些参数进行调整，然后将其传递到轨道模拟中。在训练过程中，这些参数由 PPO 智能体不断进行调整，并在环境中进行测试。动作空间的连续性允许探索更广泛的轨道配置，以最大化卫星到预定地面目标的接近程度。将真近点角（υ）排除在动作空间之外是因为它的作用是瞬时位置参数，而不是固定的轨道特征，如第 3.1 节所述。

观测空间 Gymnasium 环境返回观测以帮助调试和学习。观测空间为强化学习（RL）代理提供关于卫星轨道状态及其相对于任务目标的性能的反馈。

观测空间实现为一个 Dict 空间，结合了 Box 和 Discrete 组件，其结构如下：

- 轨道要素（框空间）：
  - 半长轴（a）
  - 离心率（e）
  - 倾角（i）
  - 升交点赤经（Ω）
  - 近地点幅角（ω）
- 地面目标有效性（离散）：一个二值（0 或 1），表示卫星当前轨道是否经过地面目标的预定半径范围内。值为 1 表示成功覆盖，而 0 表示未能达到接近阈值。
- 覆盖误差（离散）：一个二进制值（0 或 1），指示卫星的轨道高度是否满足任务的标准（即在预定义的上限和下限高度范围内）。值为 1 表示符合标准；否则为 0。

- 安全缓冲距离（离散）：一个二进制值（0 或 1），指示轨道配置是否满足任务预先设定的安全阈值。值为 1 表示轨道维持了最低安全距离；否则为 0。这对在近地轨道（LEO）中的操作尤为关键。

这种结构化观察格式提供了一个全面的状态表示，有助于稳健的学习和高效的调试。

奖励函数设计

奖励函数在引导强化学习代理朝向最佳轨道配置方面至关重要。在这项工作中，我们设计了一个综合的奖励结构，平衡了三个标准：地面目标覆盖率、轨道安全距离和高度有效性，同时对偏心率和倾角等特定参数设定软约束。

结构 总奖励计算为个体子奖励和与以下相关的惩罚的加权（根据任务需求可调）和。

- 地面目标有效性：表明轨道在目标地面坐标上的位置。
- 轨道安全距离：指示轨道到最近轨道的最小距离。
- 覆盖高度：表示轨道高度。

无效或极端的参数值会因观测空间的边界而失效。所有的奖励都经过标准化和剪裁，以避免数值不稳定，并确保使用 PPO 的平滑收敛。

覆盖高度奖励 卫星的平均高度（由半长轴推导出）预计在预定义的高度范围 $[h_{\min}, h_{\max}]$ 内。偏差通过归一化覆盖误差受到惩罚：

$$\text{normalized error}_{\text{coverage}} = \frac{\text{coverage error}}{\max\left(10^{-6},\ h_{\max} - h_{\min}\right)} \tag{1}$$

覆盖奖励 R_c 的计算公式如下：

$$R_c = \max\left(0.0,\ 1.0 - \text{normalized error}_{\text{coverage}}\right) \tag{2}$$

一个惩罚 P_c 也被定义用于反映违反的程度：

$$P_c = \min\left(1.0,\ \text{normalized error}_{\text{coverage}}\right) \tag{3}$$

安全缓冲距离奖励 为了与其它运行的卫星保持最小安全距离，使用双曲正切函数根据接近边界来设计奖励。使用双曲正切函数将奖励标准化在 $[-1, 1]$ 的范围内，同时提供一个逐渐增加的奖励斜率。

安全裕度值如下初始化最小距离范围：

$$\text{safe margin} = d_{\min} - d_{\text{safe}} \tag{4}$$

$$\text{normalized margin}_{\text{safety}} = \text{clip}\left(\frac{\text{safe margin}}{d_{\text{safe}}},\ -1,\ 1\right) \tag{5}$$

$$R_s = \frac{1}{2} \cdot \left[\tanh\left(\text{normalized margin}_{\text{safety}}\right) + 1\right] \tag{6}$$

$$P_s = 1 - R_s \tag{7}$$

这种平滑的塑造避免了突然的奖励过渡，促进了稳定的策略学习。

地面目标有效性奖励 任务约束定义了一个地面目标坐标和一个验证阈值 σ，这是轨道在任何时候必须经过的半径。当卫星在指定地面目标范围内经过时，代理会获得高奖励。随着距离增加，奖励按指数递减：

目标的距离被归一化为：

$$\text{normalized distance} = \frac{d_{\text{target}}}{\sigma} \tag{8}$$

$$R_t = e^{-3 \cdot \text{normalized distance}} \tag{9}$$

$$P_t = 1 - R_t \tag{10}$$

这种指数衰减鼓励精确覆盖并对远距离传递进行严厉惩罚。

个体元素约束 为了提高学习效率和加速收敛，对偏心率和倾角应用了个体奖励塑造。

偏心率：对于低地轨道（LEO），偏心率通常约为 0.025，从而形成近似圆形的轨道。对偏离这一数值的行为给予奖励或惩罚，有助于代理保持轨道的稳定性和可预测性。

倾角：倾角决定了轨道的纬度覆盖范围，对于确保可以到达不同位置的地面目标来说至关重要。通过根据倾角设计奖励，可以鼓励代理选择在任务约束内最大化覆盖范围的轨道。

$$R_{e,i} = R_e + R_i \tag{11}$$

$$P_{e,i} = P_e + P_i \tag{12}$$

这种针对性的奖励策略通过将优化过程与关键的轨道动力学相联系，改善了策略学习，从而在训练过程中实现更快和更稳定的收敛。

最终奖励和目标奖励 最终奖励由一个加权和构成，可以根据任务要求动态调整：

$$R = w_c R_c + w_s R_s + w_t R_t + R_{e,i} \tag{13}$$

每个目标的柔性乘法奖励构成了加快收敛速度的奖励：

$$\text{bonus} = 3 \cdot \left(\frac{R_s + R_t + R_c}{3}\right)^3 \tag{14}$$

根据达到的目标给予严厉处罚：

$$\text{penalty} = \left(1 - \frac{R_c + R_s + R_t}{3}\right)^2 \cdot \frac{P_s + P_r + P_t + P_{e,i}}{5} \tag{15}$$

最终奖励计算如下：

$$R_{\text{final}} = R + \text{bonus} - \text{penalty} \tag{16}$$

剪裁 为了稳定训练并防止异常值，总奖励会被裁剪：

$$R_{\text{final}} \in [-10,\ 10] \tag{17}$$

为了修改和学习轨道配置，我们采用同步 Advantage Actor-Critic (A2C) 算法，该算法使用 Stable-Baselines3 库实现，并配置了自定义回调和定制的超参数，以确保任务的奖励最大化和快速收敛。

X0 策略（行动者）和价值（评论者）网络均实现为全连接前馈神经网络，并以正交方式初始化，其结构如下：

- 输入层：维度与展平的观测字典的大小相匹配，包括轨道要素和离散二进制标志。
- 隐藏层：三个层的大小分别是 512、256、128。所有层都在 LeakyReLU 激活函数之前。
- 输出层：演员输出多元高斯分布的均值和对数标准差。
  - 动作网络：输出轨道元素的动作均值。
  - 评论网络：输出一个表示估计状态价值函数的单一标量值。

LeakyRelu 通过为负输入引入一个小的非零梯度来避免 ReLU 激活的“死亡神经元”问题，从而允许网络中的连续权重更新。而这种持续的变化减轻了消失梯度的风险，这是使用 Tanh 激活时常遇到的挑战，尤其是在深层架构中。

超参数是根据要求选择和调整的。

Hyperparameter	Value
gamma	0.99
gae_lambda	0.98
learning_rate	0.0001
ent_coef	0.03
vf_coef	0.75
max_grad_norm	0.4
use_rms_prop	True
rms_prop_eps	1e-5
n_steps	32
normalize_advantage	True
use_sde	True
sde_sample_freq	75
policy	MultiInputPolicy

Table 1: A2C 训练期间使用的超参数

PPO **超参数配置** 超参数是根据需求进行选择和调整的。

向量化环境 对于并行训练，环境使用 DummyVecEnv 进行矢量化，并使用 VecNormalize 进行归一化，通过减少输入和奖励之间的尺度不匹配来提高训练的稳定性。

Hyperparameter	Value
gamma	0.99
gae_lambda	0.98
learning_rate	0.0001
ent_coef	0.03
vf_coef	0.75
max_grad_norm	0.4
batch_size	1024 (default)
n_steps	2048 (default)
n_epochs	8
normalize_advantage	True
use_sde	True
sde_sample_freq	75
target_kl	0.3
policy	MultiInputPolicy

Table 2: 训练期间使用的 PPO 超参数

自定义回调 自定义回调使得在训练过程中可以进行自适应干预，以解决学习平台期和局部最优的问题。这个回调从 Stable Baselines3 框架中的 BaseCallback 扩展而来，并在检测到连续运行中的性能停滞时，通过强制重置训练环境来引入代理的动态重新定位。

一个自定义回调可以在训练过程中进行干预，以解决学习停滞和局部最优的问题。这个回调扩展了来自 Stable Baselines3 框架的 BaseCallback ，并在检测到连续的回合停滞时，通过强制重置训练环境来重新定位代理。

特性：

- 平台检测：该回调函数使用一个窗口大小为 patience 的移动窗口，在每次展开后监测智能体的平均每集奖励。如果在此窗口中的所有条目中奖励的变化保持在阈值以下，则认为智能体处于平台状态。
- 强制探索：一旦检测到平台期，回调函数会强制重新设置所有向量化环境中的轨道配置。这要求代理探索替代轨道轨迹，并逃离次优的局部奖励最大值。
- 与规范化环境的兼容性：特别注意访问和重置在 VecNormalize 、DummyVecEnv 或嵌套向量化设置下封装的环境。
- 评估感知：该回调在一个独立的评估环境中，使用确定性策略在固定数量的回合（n_eval_episodes ）中定期评估模型的策略。这确保了干预决策是基于实际的策略质量，而不是噪声较大的回合。

回调参数 自定义回调的参数

目的 在轨道环境中的强化学习由于几何约束和不稳定的奖励梯度，常常会早期收敛到次优策略。此回调功能在进展停滞时重置环境，帮助智能体探索更好的策略。它与熵、归一化、梯度剪裁以及状态依赖探索

Parameter	Value
threshold	0.25
patience	3

Table 3: 用于平台检测和干预的回调参数。

(SDE) 等技术协同工作，使训练更稳健并防止停滞和次优收敛。

我们通过与其他已实现的 RL 算法进行比较，评估这一优化框架的奖励最大化。

训练

训练是使用两个有记录的强化学习算法进行的：Stable-Baselines3 ： Advantage Actor-Critic(A2C) 和 Proximal Policy Optimization(PPO)。PPO 智能体训练到大约 62 ，000 个时间步，而 A2C 智能体训练到 2 ，500 个时间步，两者均使用 SDE 和自定义回调。在通过指数衰减和软惩罚的奖励塑造下，促进了早期约束学习。并行化的虚拟矢量环境（ DummyVecEnv ）加速了收敛并提高了样本效率。

图

政策优化动态

3 和 4 展示了与 PPO 代理的策略和价值函数更新相关的训练损失。

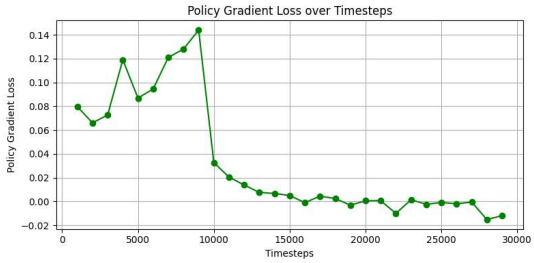

Figure 3: 策略梯度损失对时间步长的影响

一般的策略优化方法定义策略梯度损失为：

$$L_\theta^{\mathrm{PG}} = \hat{\mathbb{E}}_t \left[\log \pi_\theta(a_t \mid s_t) \cdot \hat{A}_t \right]$$

，其中 π_θ 是随机策略，$\hat{A}_t$ 是时间步 t 时优势函数的估计，定义为：

$$\hat{A}_t = \mathit{Discounted\ Rewards} - \mathit{Baseline\ Estimate}$$

。PPO 目标方法 (Schulman et al. 2017b) ，与 TRPO 并没有太大区别，其截断的替代目标定义为

$$L^{CLIP}(\theta) = \hat{E}_t \left[min(r_t(\theta)\hat{A}_t,\ clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right]$$

，其中 ϵ 是一个超参数。TRPO 惩罚 KL 散度 (Joyce 2011) 以使政策更新保持在可信区域内。虽然这种方法保持了稳定的策略梯度，但可能阻碍探索。然而，由于物理限制的环境，如本研究中讨论的那些，需要足够的探索以实现最佳收敛。根据 (Schulman et al. 2017a) 的信任区域策略优化工作，优化技术需要在训练中使用显著更多的时间步。

PPO 通过采用裁剪代理目标并引入 KL 散度作为约束来缓解这种限制，强制执行裁剪机制。当与强制搬迁（参见小节自定义回调）等策略以及更精确的超参数调整——特别是针对熵、梯度归一化、KL 收敛和策略损失相结合时，PPO 代理能够实现更稳健的探索。

相比之下，如 (Mnih et al. 2016) 所展示，所采用的同步（A2C ）以及异步演员批评方法（A3C ）采取了一种从根本上不同的方法，这种方法被证明更适合物理约束的轨道环境。

A2C 利用向量化的并行环境而不是经验回放存储器，从而能够在多个轨道场景中同时探索。这种方法消除了与存储大量经验回放数据相关的内存开销，同时确保所有收集的数据反映当前的策略，保持了策略内学习的一致性，这对需要在信任区域之外扩展探索以发现最优解决方案的轨道动力学是有利的。

并行向量化环境更可能探索环境的不同部分，从而减少对显式熵正则化超参数的依赖。因此，从多个并行环境（n_env ）收集的整体经验比从单个环境顺序生成数据的经验更不容易相关。这种去相关性改进了梯度估计，从而导致更稳定的策略更新——这是轨道力学中一个关键因素，因为小的参数变化可能导致显著不同的轨道。

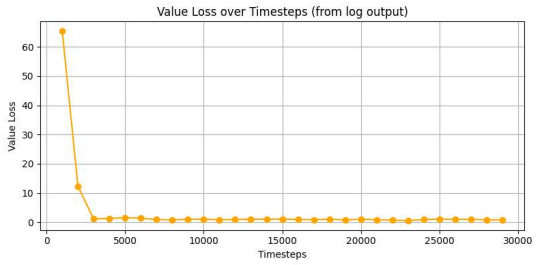

Figure 4: 价值函数损失与时间步骤

预测 & 奖励曲线分析

A2C 的不受约束的策略更新能够产生更高的一致性奖励，特别是在需要积极探索以最大化奖励的环境中非常有用。不同于 PPO 的信任区域，当 A2C 发现有利的轨道配置能够产生高奖励时，它会立即采取大幅度的策略步骤朝向这些最大化奖励的行为。

A2C 的并行向量环境能够同时探索和即时知识传播，因此比 PPO 的顺序方法更频繁地增加高奖赏状态的概率。它的同步更新机制确保在更新的情况下立即产生全球策略影响，创建一个更紧密的反馈循环以进行策略调整。这些因素使 A2C 能够快速传播成功的策略，从而

使 A2C 能够识别产生更高奖励的轨道配置。

图 5 和 6 展示了代理的平均每集奖励随时间步骤的演变。

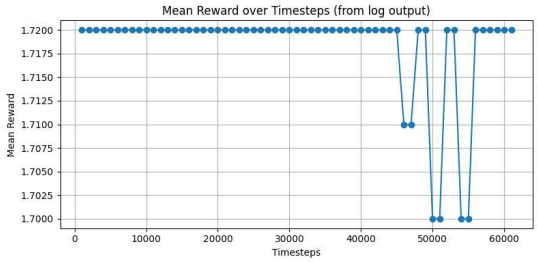

Figure 5: PPO 代理的平均情节奖励随时间步的变化

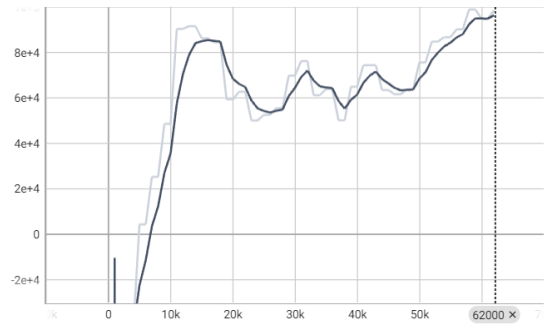

Figure 6: A2C 智能体的平均回合奖励随时间步长的变化

尽管相比经过超过 61000 时间步的 PPO 模型（参见表 4），仅经过 2300 时间步训练的 A2C 模型（参见表 4）的预测结果更好。

Table 4: 模型预测和经过训练的 A2C 智能体输出

Parameter	Value
Semi-major axis (km)	7527.649
Eccentricity	0.049
Inclination (rad)	1.618
RAAN (rad)	3.127
Argument of periapsis (rad)	3.085
Cumulative Reward	10.0
Objectives Met	True

Table 5: 模型预测和训练过的 PPO 智能体的输出

Parameter	Value
Semi-major axis (km)	7078.137
Eccentricity	0.100
Inclination (rad)	3.142
RAAN (rad)	0.009
Argument of periapsis (rad)	6.283
Cumulative Reward	9.263025
Objectives Met	True

这种性能差距因此突出了 A2C 算法在给定 MDP 表述中的优势，表明 actor-critic 方法非常适合需要广泛探索、奖励利用和策略稳定性的轨道优化任务。

本研究成功开发并验证了一种基于 TLE 的强化学习框架，用于自主卫星轨道优化。通过优势演员-评论家（A2C）算法，相较于传统方法，实现了卓越的性能。实验结果证实，A2C 获得了 73.6% 的更高累积奖励（10.0 对 9.263025），同时相比于邻近政策优化（PPO），所需训练时间步数减少了 27.4 倍（2,240 对 61,440），确立 A2C 作为物理约束轨道规划任务的最佳算法。自定义奖励函数的使用，结合依赖状态的探索和自定义回调干预，使代理能够克服稀疏奖励挑战并收敛至可行解决方案。策略网络和量身定制的超参数成功地指导轨道要素在最大化目标覆盖的同时保持安全约束。关键技术细节包括：（1）一个与 Gymnasium 兼容的仿真环境，整合了真实 TLE 数据、轨道力学和任务约束；（2）混合奖励函数，旨在指导探索并满足目标；（3）通过依赖状态的探索和自定义回调干预成功应对稀疏奖励挑战；（4）验证了演员-评论家方法在连续轨道域中相较于信任区域方法的优势。该方法展示了强化学习作为轨道动力学传统优化方法的可行替代方案的潜力，尤其适用于低地球轨道（LEO）任务中的响应性和自适应轨道规划。

References

Joyce, J. M. 2011. Kullback-Leibler Divergence, 720–722. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-04898-2.

Kyuroson, A.; Banerjee, A.; Tafanidis, N. A.; Satpute, S.; and Nikolakopoulos, G. 2024. Towards fully autonomous orbit management for low-earth orbit satellites based on neuro-evolutionary algorithms and deep reinforcement learning. European Journal of Control, 80: 101052. 2024 European Control Conference Special Issue.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T. P.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous Methods for Deep Reinforcement Learning. arXiv:1602.01783.

Mok, S.-H.; Jo, S.; Bang, H.; and Leeghim, H. 2019. Heuristic-Based Mission Planning for an Agile Earth Observation Satellite. International Journal of Aeronautical and Space Sciences, 20(3): 781–791.

Savitri, T.; Kim, Y.; Jo, S.; and Bang, H. 2017. Satellite Constellation Orbit Design Optimization with Combined Genetic Algorithm and Semianalytical Approach. International Journal of Aerospace Engineering, 2017(1): 1235692.

Schulman, J.; Levine, S.; Moritz, P.; Jordan, M. I.; and Abbeel, P. 2017a. Trust Region Policy Optimization. arXiv:1502.05477.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017b. Proximal Policy Optimization Algorithms. arXiv:1707.06347.

Song, Z.; Chen, X.; Luo, X.; Wang, M.; and Dai, G. 2018. Multi-objective optimization of agile satellite orbit design. Advances in Space Research, 62(11): 3053–3064.

Tafanidis, N. A.; Banerjee, A.; Satpute, S.; and Nikolakopoulos, G. 2025. Reinforcement learning-based station keeping using relative orbital elements. Advances in Space Research, 76(2): 750–763.

Towers, M.; Kwiatkowski, A.; Terry, J.; Balis, J. U.; Cola, G. D.; Deleu, T.; Goulão, M.; Kallinteris, A.; Krimmel, M.; KG, A.; Perez-Vicente, R.; Pierré, A.; Schulhoff, S.; Tai, J. J.; Tan, H.; and Younis, O. G. 2024. Gymnasium: A Standard Interface for Reinforcement Learning Environments. arXiv:2407.17032.