arXiv:2302.00237v1 [cs.LG] 1 Feb 2023

融合物理信息神经网络与强化学习：哈密顿-雅可比-贝尔曼近端策略优化（HJBPPO）

Amartya Mukherjee[1]*，Jun Liu[1]滑铁卢大学应用数学系，滑铁卢，安大略省 N2L 3G1，加拿大

摘要

本文将哈密顿-雅可比-贝尔曼近端策略优化（HJBPPO）算法引入强化学习。哈密顿-雅可比-贝尔曼（HJB）方程用于控制理论中，以评估价值函数的最优性。我们的工作将 HJB 方程与连续状态与动作空间中的强化学习相结合，以改进价值网络的训练。我们将价值网络视为物理信息神经网络（PINN），通过精确计算其相对于输入的导数来求解 HJB 方程。我们基于此实现改进了近端策略优化（PPO）裁剪算法，因为该算法使用价值网络来计算其策略网络的目标函数。与 MuJoCo 环境中的 PPO 相比，HJBPPO 算法表现更佳。

关键词：连续时间强化学习、物理信息神经网络、近端策略优化、Hamilton-Jacobi-Bellman方程

*通讯作者。电子邮箱：a29mukhe@uwaterlooo.ca

1 引言

近年来，人们对用于连续控制问题的强化学习（RL）兴趣日益浓厚。通过在环境中平衡探索与利用已学习的策略，RL在动力学未知的环境中取得了令人鼓舞的成果。自带基线的REINFORCE问世以来，RL算法中的价值网络已被证明有助于寻找最优策略，并充当评论家网络 [21]。如今，这种价值网络仍应用于最先进的RL算法中。

在离散时间RL中，价值函数将给定状态的回报估计为各时间步回报之和。该价值函数通过求解Bellman最优性方程得到。另一方面，在连续时间RL中，价值函数将给定状态的回报估计为对时间的积分。该价值函数通过求解一种称为Hamilton-Jacobi-Bellman（HJB）方程的偏微分方程（PDE）得到 [12]。这两个方程都难以通过解析或数值方法求解，因此强化学习智能体必须探索环境并进行连续估计。

当前强化学习文献中已有的算法，例如近端策略优化（近端策略优化），旨在利用贝尔曼最优方程更新价值函数，使其估计每个状态的离散时间回报。然而，我们发现，该价值函数在 MuJoCo 环境中训练时，并未表现出向

HJB 方程所描述的最优价值函数收敛的趋势（见图 4）。这表明，使用离散时间步而非连续时间训练价值函数会导致信息丢失。

自 [17] 提出物理信息神经网络（PINNs）以来，科学机器学习取得了显著进展。物理信息神经网络利用自动微分，精确计算神经网络相对于其输入和模型参数的导数。这使得支配相关数据集的物理定律（由常微分方程或偏微分方程描述）能够作为神经网络的正则化项。因此，物理信息神经网络能够利用数据所蕴含的物理规律，在此类数据集上优于常规神经网络。

据我们所知，本文首次研究物理信息神经网络与强化学习的交叉领域。为了促使近端策略优化中价值函数的收敛结果趋近于 HJB方程的解，我们利用物理信息神经网络对该偏微分方程进行编码，并弥合离散时间与连续时间下计算的回报之间的信息差距。这使我们的算法能够利用自动微分，消除梯度计算和时间离散化所带来的误差。我们提出了 Hamilton-Jacobi-Bellman 近端策略优化（ HJBPPO）算法。与 MuJoCo 环境中的近端策略优化相比，该算法在奖励更高、收敛更快、稳定性更强等方面表现优异，是一项显著改进。

2 预备知识

考虑一个由下列方程建模的受控动力系统：

$$\dot{x} = f(x,u), \quad x(t_0) = x_0, \tag{1}$$

其中，$x(t)$ 是状态，$u(t)$ 是控制输入。在控制理论中，最优价值函数$V^*(x)$ 有助于求解控制问题[8]:

$$V^*(x) = \sup_u \int_{t_0}^{\infty} \gamma^t R(x(\tau;t_0,x_0,u(\cdot)),u(\tau))d\tau, \tag{2}$$

其中，$R(x,a)$ 是奖励函数，γ 是折扣因子。以下定理提出了一项用于评估价值函数 [[11], [13]]最优性的准则。

定理 2.1。 函数$V(x)$ 是最优值函数，当且仅当：

1. $V \in C^1(\mathbb{R}^n)$ *and* V *satisfies the Hamilton-Jacobi-Bellman (HJB) Equation*

$$V(x)\ln\gamma + \sup_{u\in U}\{R(x,u) + \nabla_x V^T(x) f(x,u)\} = 0 \tag{3}$$

for all $x \in \mathbb{R}^n$.

2. 对于所有 $x \in \mathbb{R}^n$，都存在一个控制器 $u^*(\cdot)$ 使得：

$$\begin{aligned} &V(x)\ln\gamma + R(x,u^*(x)) + \nabla_x V^T(x) f(x,u^*(x)) \\ &= V(x)\ln\gamma + \sup_{\hat{u}\in U}\{R(x,\hat{u}) + \nabla_x V^T(x) f(x,\hat{u})\}. \end{aligned} \tag{4}$$

目前已有的强化学习算法并未着重于求解 HJB 方程，以最大化每个回合的总奖励。例如，在近端策略优化中，对 MuJoCo环境进行测试时，HJB方程似乎并未得到满足。

为说明这一点，我们将每个回合的 HJB损失定义如下：

$$\begin{aligned} &MSE_f \\ &= \frac{1}{T}\sum_{t=0}^{T-1} |V(x_t)\ln\gamma + R(x_t,a_t) + \nabla_x V^T(x_t) f(x_t,a_t)|^2, \end{aligned} \tag{5}$$

其中，T 是回合中的时间步数，x_t是时间步t时环境的状态，a_t 是时间步t采取的动作。 $\nabla_x V^T(x_t)$通过自动微分精确计算。我们使用有限差分近似$f(x_t, a_t)$：

$$\begin{aligned} MSE_f = \frac{1}{T}\sum_{t=0}^{T-1} |&V(x_t)\ln\gamma + R(x_t,a_t) \\ &+ \nabla_x V^T(x_t)\left(\frac{x_{t+1}-x_t}{\Delta t}\right)|^2, \end{aligned} \tag{6}$$

其中， Δx 是环境中使用的时间步长。我们绘制了使用近端策略优化在各个环境中的HJB损失，如图4所示。各环境的平均HJB损失值都极高，并且在10个环境中有6个未能收敛，这表明价值函数无法像HJB方程所示那样收敛到最优价值函数。

作为对比，我们在图5中绘制了价值网络损失曲线。Bellman最优性损失在10个环境中的8个环境里呈现收敛。这表明，与连续时间价值函数相比，求解离散时间价值函数的Bellman最优性方程时会丢失信息。这也表明，价值函数收敛并不一定会带来HJB损失的收敛。

为解决这一问题（如图4所示），我们将价值网络视为物理信息神经网络，并使用基于梯度的方法来降低HJB损失。利普希茨连续性要求，并最大限度地降低了有限差分方法所带来的计算误差。

利用神经网络求解HJB方程一直是多个研究项目关注的领域。 [7] 使用结构化循环神经网络求解HJB方程，并为Dubins汽车问题实现最优控制。 [22] 使用Pineda架构 [15]估计价值函数相对于其输入的偏导数。他们使用迭代最小二乘法求解HJB方程。该算法在多个控制问题中均实现了收敛，且无需初始稳定策略。 [20] 开发了DGM算法来求解偏微分方程。他们使用自动微分计算一阶导数，并使用蒙特卡洛方法估计高阶导数。该算法曾用于求解HJB方程，以实现随机偏微分方程的最优控制，误差达到0.1%。我们希望在这些研究的基础上进一步推进，并使用物理信息神经网络求解HJB方程。

3 相关工作

近年来，强化学习社区和控制理论社区对将HJB方程用于连续强化学习产生了浓厚兴趣，并推动了若干前景良好的研究。 [10] 为Q网络提出了一种替代性HJB方程，并据此推导出一个在时间上利普希茨连续的控制器。在测试的四种MuJoCo环境中，有三种环境的表现优于深度确定性策略梯度（DDPG），且无需演员网络。 [24]提出了一种分布式HJB方程，用于训练FD-WGF Q学习算法。在粒子控制任务中，与分位数回归时序差分（QTD）相比，该算法能更准确地建模回报分布。研究使用有限差分方法对该HJB方程进行数值求解。我们计划在这些工作的基础上加入策略网络，并采用物理信息神经网络求解HJB方程。此外，作者在结论中提到，使用自动微分提高分布式HJB方程的精度，是未来研究的一个潜在方向。我们的工作放宽了对控制器的

在最优反馈控制问题的设定中， [14] 使用物理信息神经网络求解 HJB 方程，以训练价值网络。这避免了使用有限差分计算价值函数的导数。该论文在高维问题中取得了与真实最优控制函数相近的结果。我们计划在此基础上，将物理信息神经网络应用于动力学未知且需要探索的强化学习场景。

4 HJBPPO

据我们所知，我们的工作首次将 HJB 方程与现有的强化学习算法近端策略优化相结合，也首次在强化学习场景中使用物理信息神经网络求解 HJB 方程。

PPO裁剪算法在此实现中得到改进，因为它使用价值网络来计算优势函数，并据此更新其策略网 [18]。近端策略优化是一种行动者-评论家方法，它会在每次迭代中将策略网络的更新限制在一个较小的信赖域内。这确保策略网络的目标函数能够良好地近似真实目标函

数，同时也促使价值网络进行小幅更新。因此，近端策略优化通过确保每次迭代都进行小幅且稳健的更新，在确定性强化学习环境中展现出最先进的性能。

一项由 [9] 开展的研究，对受控噪声下的双时间尺度随机逼近进行了收敛性分析。在演员-评论家方法中，使用随机梯度下降或 Adam 等优化器更新神经网络参数，可视为随机常微分方程的数值解。

因此， [5] 利用这项工作对演员-评论家方法进行了收敛性分析。此外，这些结果还被用于证明近端策略优化和 RUDDER 的渐近收敛性 [1]。作者提出了模型假设和损失函数假设，满足这些假设后，近端策略优化和 RUDDER 中的参数便可能收敛到初始值附近的某个局部最小值。研究表明，近端策略优化中的策略网络和价值网络参数可能收敛到初始值附近的某个局部最小值。

另一项由 [6]开展的理论研究得出结论：包括近端策略优化在内的策略优化方法，通过将问题表述为非凸优化问题并采用基于梯度的优化，在 线性二次型调节任务 上能够保证收敛。该研究还表明，这些方法在状态反馈控制问题上表现可靠。论文指出，先进的正则化技术可能有助于提升鲁棒性。这进一步证明了引入 HJB损失 作为正则化项的合理性。

4.1 用于 HJB 方程的物理信息神经网络

我们的工作将HJB方程与连续状态与动作空间中的强化学习相结合，以改进价值网络的训练。对于具有无限时间范围的随机系统，HJB方程是一个二阶椭圆方程 [2]。 [19] 的理论研究表明，物理信息神经网络能够一致收敛到二阶线性椭圆方程的解，从而证明了使用物理信息神经网络求解HJB方程的合理性。

我们将价值网络视为物理信息神经网络，通过精确计算其相对于输入的导数来求解HJB方程。

请注意，该项

$$\sup_{\hat{u}\in U}\{R(x,\hat{u}) + \nabla_x V^T(x) f(x,\hat{u})\}$$

在 HJB 方程中无法确定，除非智能体在其环境中进行探索。根据定理 2.1，最优策略 $\pi^*(a|x)$和最优控制器$u^*(x) = \text{argmax}_a \pi^*(a|x)$满足公式（4）。最优策略由策略网络 π_θ 建模，其参数由 θ 确定；最优控制器则可用$u(x) = \text{argmax}_a \pi_\theta(a|x)$近似。因此，我们可以使用以下近似：

$$\begin{aligned} &V(x)\ln\gamma + R(x,u(x)) + \nabla_x V^T(x) f(x,u(x)) \\ &\approx V(x)\ln\gamma + \sup_{\hat{u}\in U}\{R(x,\hat{u}) + \nabla_x V^T(x) f(x,\hat{u})\}. \end{aligned} \quad (7)$$

因此，这证明了可以将公式 6 用作 HJB损失，并在每个回合更新价值函数。

损失函数的计算方式为

$$J(\varphi) = 0.5 MSE_u + \lambda_{HJB} MSE_f,$$

其中，MSE_f 在公式（6）中定义，MSE_u是近端策略优化中价值网络使用的标准损失函数，用于改进价值函数回报的离散时间估计：

$$MSE_u = \frac{1}{T}\sum_{t=0}^{T-1} |V(x_t) - (R(x_t,a_t) + \gamma V(x_{t+1}))|^2, \quad (8)$$

其中，$\{x_t\}^T{}_{t=1}$ 表示在单个回合中探索的一批状态，而$\{a_t = \text{argmax}_a \pi_\theta(a|x_t)\}^T{}_{t=1}$ 表示在时间步t 按照策略π_θ执行的一批动作。根据HJB损失曲线相对于贝尔曼最优性损失曲线的量级确定超参数 λ_{HJB} ，以便为两个损失函数赋予相近的权重。

4.2 算法

HJBPPO算法见算法1。策略更新以及均方误差$_u$的最小化

算法1 HJBPPO

1: 初始化策略网络参数 θ 和价值网络参数 φ

2: **对于**迭代 $= 1, 2, \ldots$ **执行**

3: Run the 策略 π_θ in the 环境，持续 T 时间步 and 观察 环境-样本 $\{(s_t, a_t, R_t, s_{t+1})\}_{t=1}^{T}$.

4: 计算优势 A_t

5: 计算 $r_t(\theta) = \frac{\pi_\theta(-_t|\text{秒}_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$

6: 计算策略的目标函数

网络的目标函数：

$$L(\theta) = \frac{1}{T}\sum_{t=0}^{T-1} \min[r_t(\theta)A_t,\ \text{clip}(r_t(\theta),\ 1-\varepsilon, 1+\varepsilon)A_t]$$

7: U更新 $\theta \leftarrow \theta + \alpha_1 \nabla_\theta L(\theta)$

8: 将价值网络损失计算为：$J(\varphi) = 0.5MSE_u + \lambda_{HJB} MSE_f$，如公式（8）所述和（6）

9: 更新 $\varphi \leftarrow \varphi - \alpha_2 \nabla_\varphi J(\varphi)$

10: **结束循环**

价值网络与近端策略优化相同。为了满足定理 2.1 中所述的$V(x) \in C^1(\mathbb{R}^n)$，我们使用无穷可微的双曲正切 激活函数作为价值网络的激活函数。

算法中的第3–7行与PPO算法相同。优势项A_t 的计算方式如下：$A_t = \sum_{=}^{T-1} {}_{nt} (\gamma\lambda)^{n-t}\delta_n$，其中 $\delta_t = R_t + \gamma V\varphi(s_{t+1}) - V\varphi(s_t)$，$\gamma$ 为折扣因子，λ 为广义优势估计（GAE）参数。α_1 和 α_2 分别是策略网络优化器和价值网络优化器的学习率。ε 为裁剪参数。

算法中的第8–9行是我们对PPO算法所做的修改。我们将价值网络视为物理信息神经网络（PINN），并在其损失函数中加入MSE_f 项。这样，HJB方程就被用作价值网络的正则化项。

我们将比较 HJBPPO 与近端策略优化在MuJoCo环境中的奖励表现，以及 HJBloss。

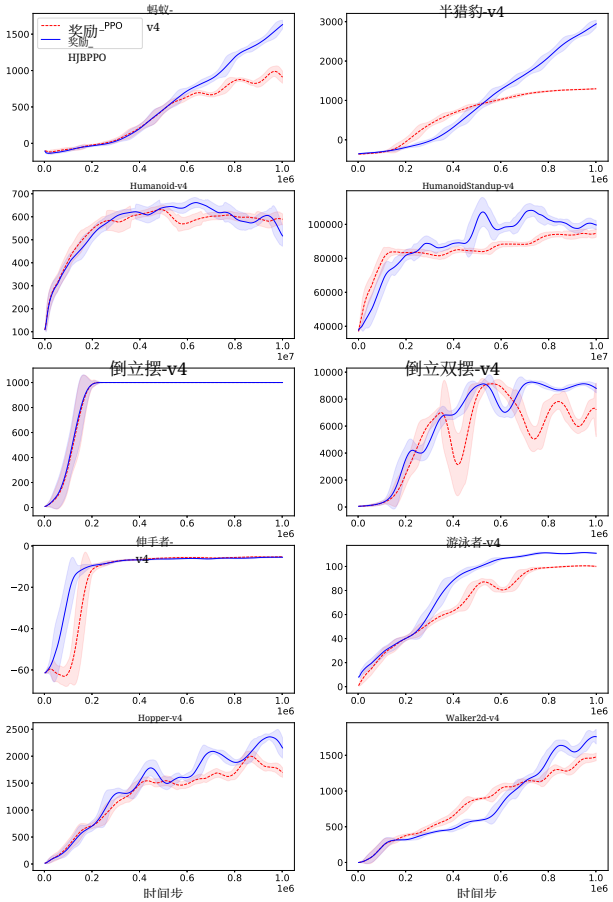

图1：PPO（红色，虚线）与HJBPPO（蓝色，平滑曲线）学习曲线的比较

5 结果

5.1 训练

HJBPPO算法由 [16] 修改 Stable Baselines 3 库中的近端策略优化代码实现。为确保结果可复现，我们已将代码发布于（GitHub 仓库已隐去，代码见补充材料），并在附录 A 的表 1 和表 2 中提供了超参数。

代码在 Compute Canada 的 B´eluga 集群上运行。该集群提供了 MuJoCo 环

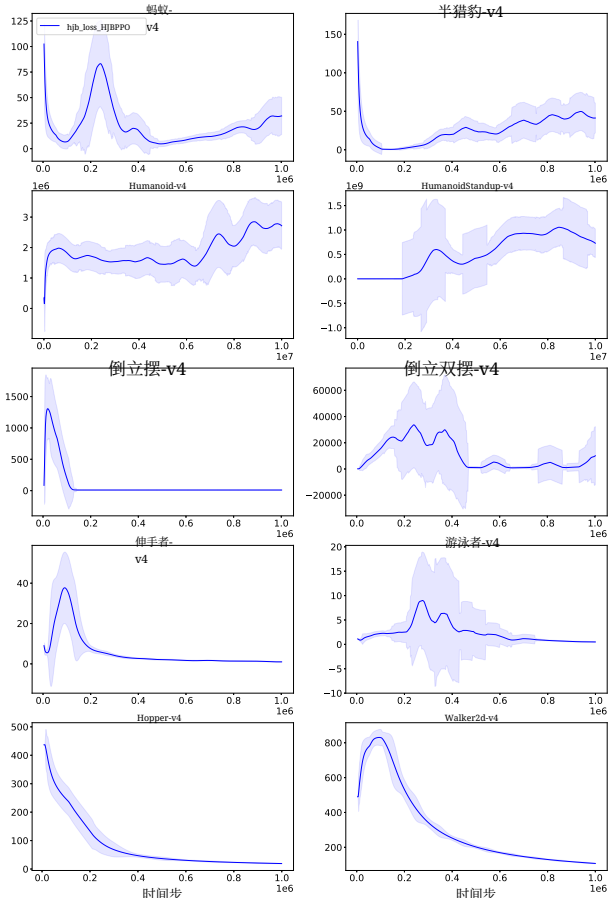

图2：HJBPPO在MuJoCo环境中的HJB损失曲线

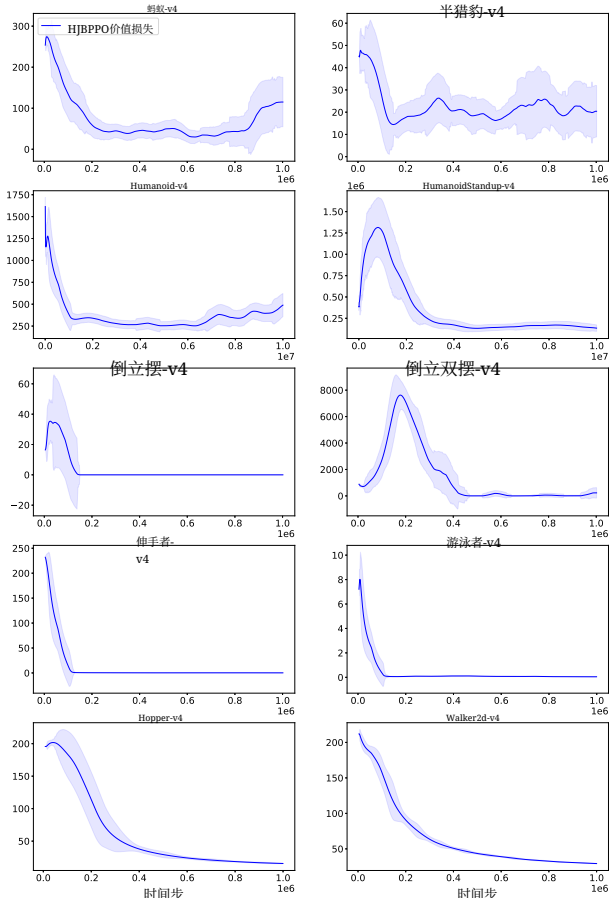

图3：HJBPPO在MuJoCo环境中的贝尔曼最优性损失曲线

境用于训练。每个算法训练 100 万个时间步需要七小时，训练 1000 万个时间步需要三天。我们使用 Python 的多进程库，同时在多个环境中训练每个算法。

5.2 奖励曲线

图1绘制了奖励曲线，对比了 HJBPPO 与近端策略优化在所有 MuJoCo 环境中经过一百万或一千万个时间步后的表现。曲线表示总奖励，是连续 50 个回合的平均值；阴影区域表示连续 50 个回合总奖励的标准差。HJBPPO 在 Ant-v4 和半猎豹-v4 中表现出显著提升；在伸手者-v4、游泳者-v4 和倒立双摆-v4 中收敛更快、稳定性更好；在 HumanoidStandup-v4、Hopper-v4 和 Walker2d-v4 中略有提升。在其余两个环境（Humanoid-v4 和倒立摆-v4）中，HJBPPO 的表现与近端策略优化相当。

因此，图表显示，将连续时间HJB方程纳入近端策略优化算

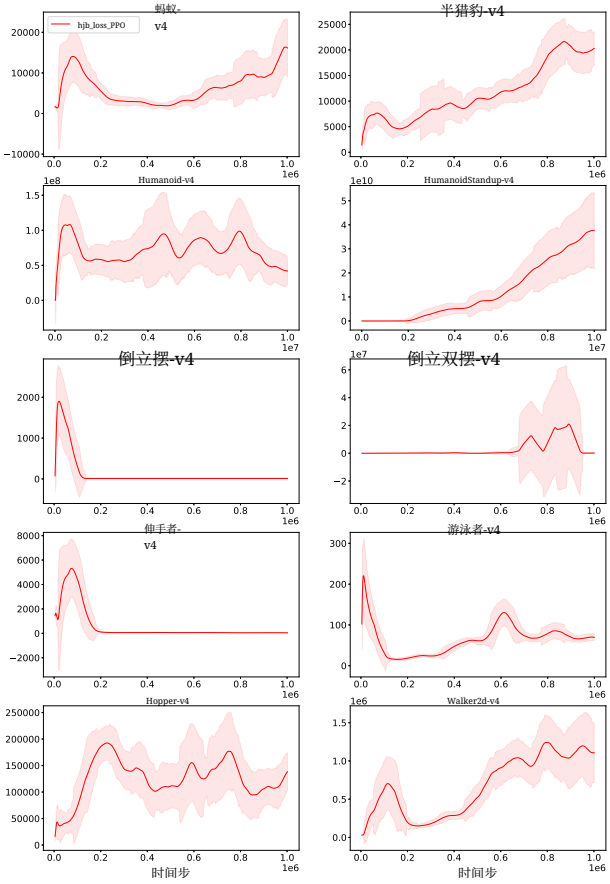

图4：MuJoCo环境中近端策略优化的HJB损失曲线

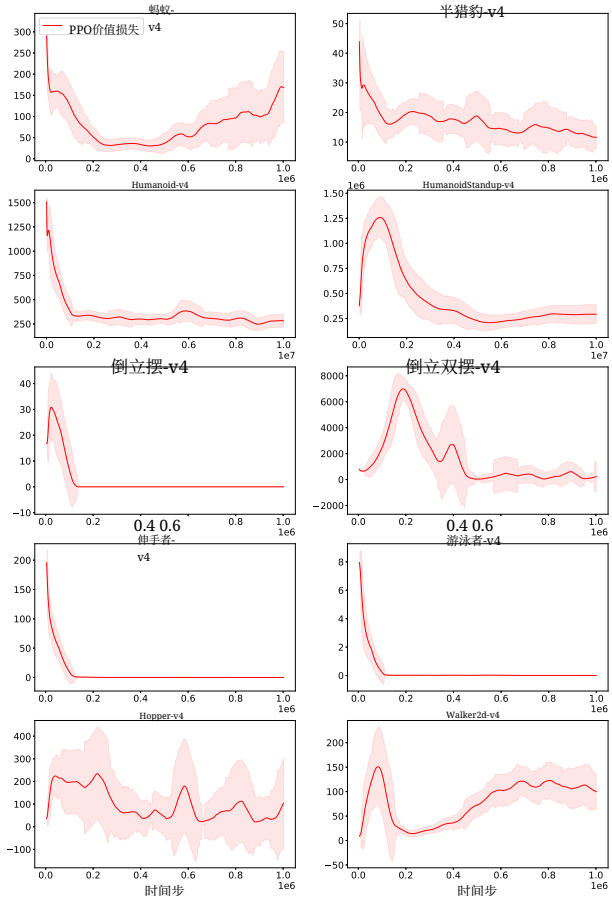

图5：近端策略优化在MuJoCo环境中的Bellman最优性损失曲线

法中用于训练价值函数，可以改善智能体的学习曲线。这是因为HJBPPO使用物理信息神经网络来利用环境中的物理规律。它使用有限差分近似环境的底层控制方程$f(x, u)$，并使用自动微分求解HJB方程，从而实现最优控制。

5.3 HJB损失曲线

图2展示了各个环境中的HJBloss。与近端策略优化相比，HJBPPO的HJB损失显著降低。而HJB损失

在10个环境中的8个环境里表现出收敛，包括蚂蚁-v4和半猎豹-v4；在这两个环境中，其奖励表现显著优于近端策略优化，从而表明价值函数收敛到了HJB方程所示的最优价值函数。

对于 Humanoid-v4 和 HumanoidStandup-v4，HJB损失并未收敛，尽管这些环境中的 HJB损失显著低于图4所示的数值。在这两个环境中，HJBPPO 的奖励曲线与近端策略优化表现相似。这表明，如果 HJB损失未显示出收敛性，HJBPPO 的性能总体上并不会显著优于

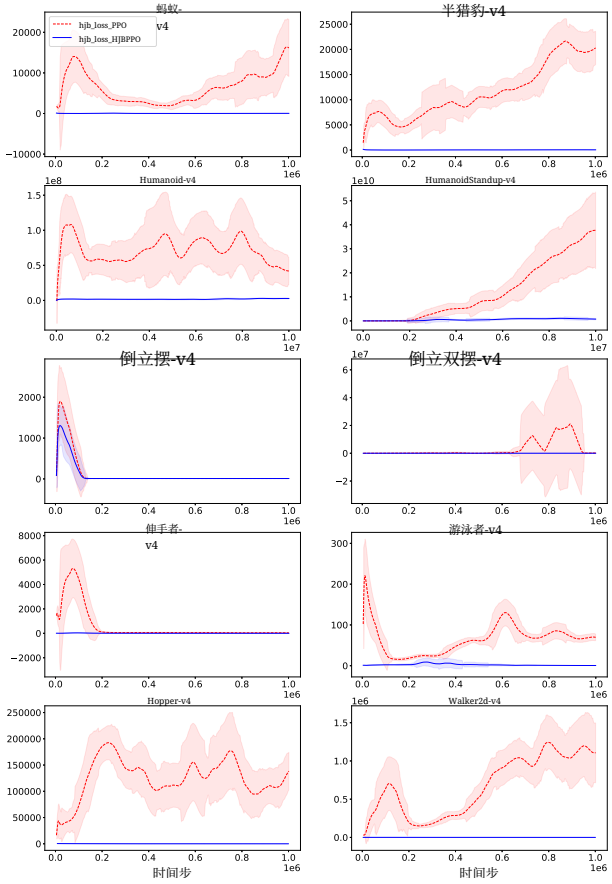

图6：PPO（红色，虚线）与 HJBPPO（蓝色，平滑）的 HJB损失曲线对比

近端策略优化。

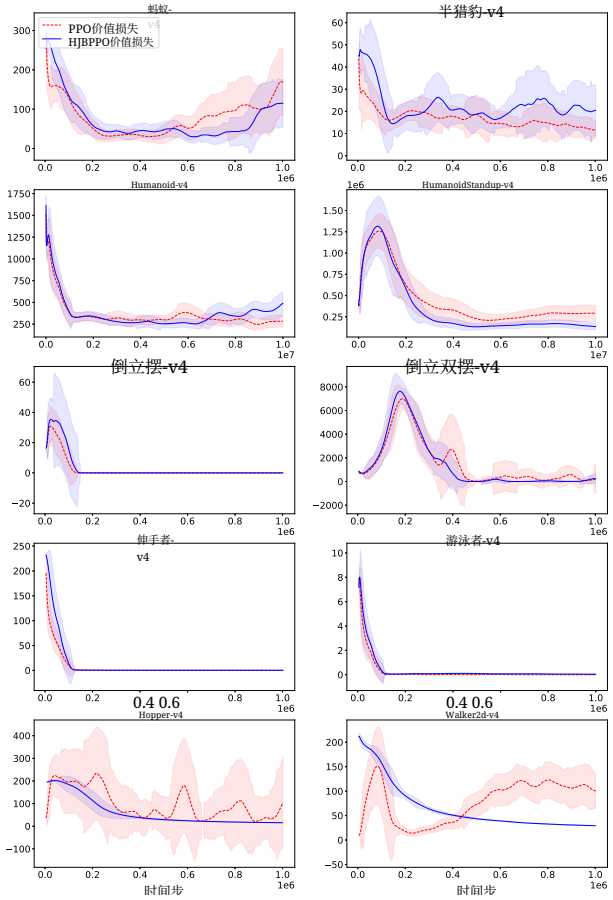

图7：PPO（红色，虚线）与 HJBPPO（蓝色，平滑）的贝尔曼最优性损失曲线对比

5.4 贝尔曼最优性损失曲线

每个环境的Bellman最优性损失都已绘制在图3中。价值网络在每个环境中都表现出收敛。这表明，连续时间HJB方程中的价值函数收敛，也能改善其在离散时间Bellman最优性方程中的收敛；反之则不一定成立。

5.5 与近端策略优化的比较

作为比较，下面的图4和图5展示了近端策略优化（PPO）的HJB损失和Bellman最优性损失曲线。一个显著的区别是，与PPO相比，HJBPPO的HJB损失值明显更低。这是因为在HJBPPO的训练过程中，HJB损失会被主动最小化。为了更清楚地展示这种差异，我们在图6中将损失曲线绘制在同一张图上。显然，与PPO相比，HJBPPO的HJB损失值极低。

如图4所示，在10个环境中，PPO的HJB损失仅在4个环境中呈现收敛：倒立摆-v4、InvertedDoublePendulum-v4、伸手者-v4和游泳者-v4。对于其余环境，HJB损失总体呈上升趋势。

如图2所示，对于PPO中HJB损失收敛的环境，HJBPPO中的HJB损失也呈现收敛。尽管HJBPPO在HumanoidStandup-v4中未能实现HJB损失收敛，但与PPO相比，其损失曲线有所改善；PPO的损失曲线则呈现显著上升趋势。

图5显示，使用PPO时，10个MuJoCo环境中有8个的Bellman最优性损失实现了收敛。PPO通过训练价值函数来求解Bellman最优性方程，从而使Bellman最优性损失收敛。然而，尽管采用了这一损失函数，在蚂蚁-v4和Walker2d-v4中，Bellman最优性损失并未收敛，而是在较大的时间步数下呈现上升趋势。

如图3所示，HJBPPO解决了这一问题。在包括蚂蚁-v4和Walker2d-v4在内的所有环境中，Bellman最优性损失总体呈下降趋势。为了更清楚地展示差异，我们在图7中将Bellman最优性损失曲线绘制在同一张图上。总体而言，HJBPPO的Bellman最优性损失曲线表现与PPO相当，并且在蚂蚁-v4、HumanoidStandup-v4、倒立双摆-v4、Hopper-v4和Walker2d-v4中收敛更快。

总而言之，HJBPPO 展现出了优于PPO 的性能。其奖励曲线、HJBloss 曲线以及Bellman 最优性损失曲线均有所改善。这是因为HJBPPO 引入了 HJB损失正则化项，并利用最优控制中的方法来改进价值函数的学习，从而提高算法的收敛性。

6 结论

本文介绍了HJBPPO算法，该算法改进了PPO算法，以求解HJB方程。据我们所知，本文首次将物理信息神经网络与强化学习相结合。我们将价值函数视为物理信息神经网络，在强化学习场景中求解HJB方程。与PPO相比，HJBPPO算法的整体表现有所提升，这是因为它能够利用环境中的物理规律以及最优控制，从而改善智能体的学习曲线。本文还表明，价值函数在连续时间HJB方程中的收敛，也能改善其在离散时间贝尔曼最优性方程中的收敛。

7 未来研究

尽管奖励曲线整体有所改善，HJBPPO仍为利用物理信息神经网络改进强化学习算法留有探索空间。

如图2所示，HJBPPO算法的一个局限是，尽管与近端策略优化相比有显著改进，但HJB损失在环境中并非总能收敛。未来研究的一个潜在方向是探索新的物理信息神经网络优化方法，以改善HJB损失的收敛性。

训练价值网络时使用的损失函数 MSE_f 是根据方程 7 中采用的近似推导得出的。因此，这并不能保证策略网络收敛到满足以下条件的最优策略：$\hat{u}(x) = \sup_{\hat{u}\in U} \{R(x, \hat{u}) V(x) f(x, \hat{u})\}$，其中控制器 $u(x)$ 是由策略 $\pi_\theta(a|x)$ 推导得出的。[5]证明了近端策略优化中策略网络参数收敛到局部最优，

但并不保证全局收敛。因此，一个值得进一步研究的潜在方向是，为价值网络探索其他 HJB损失函数，以放宽这一近似条件。

本文探讨并比较了两种确定性强化学习算法——HJBPPO和PPO。值得研究的是，如何将该算法扩展到随机环境。在SAC论文中，[4]提出了一种熵正则化随机策略，这种策略不太容易过拟合或陷入局部最优。由于方程（7）中所采用的近似，HJBPPO算法也存在这样的风险：价值函数的HJB损失可能导致策略过拟合，最终得到次优策略。可以通过引入另一种能够促进探索并融入熵最大化的HJB方程来降低这一风险。因此，利用物理信息神经网络和该HJB方程改进SAC算法，是一个值得进一步探索的方向。

在MuJoCo环境中，HJBPPO算法的表现优于PPO。但这是因为可以通过有限差分估计$f(x, u)$，从而能够利用环境中的物理规律。这些环境提供了选择动作所需的全部状态信息。HJBPPO的一项局限是，在部分可观测环境中可能表现不佳，因为$f(x, u)$的估计可能不准确。 [3] 提出了深度Transformer Q网络（DTQN），并在许多部分可观测环境中取得了最先进的结果。一个值得进一步研究的方向是引入另一种能够适应部分可观测性的HJB方程。通过结合物理信息神经网络使用该HJB方程，可以改进DTQN算法。

此外，在高维环境中，有限差分近似的准确性会降低 [20]。这使得 HJB 损失在 Humanoid-v4 和 HumanoidStandup-v4 等环境中不太可靠，因为这些环境的状态是一个 376 维向量。有限差分近似无法精确计算 $f(x, u)$，因为环境采用半隐式欧拉积分步，而非欧拉方法 [23]。因此，未来研究的一个潜在方向是将 HJBPPO 与基于模型的强化学习相结合，从而以更小的误差估计 $f(x, u)$。

致谢

作者谨向滑铁卢大学计算机科学系的 Pascal Poupart、Ashish Gaurav 和 Yanting Miao 致谢，感谢他们为本文提供反馈。

参考文献

[1] Arjona-Medina, J.A., Gillhofer, M., Widrich, M., Unterthiner, T., Brandstetter, J., and Hochreiter, S. RUDDER：延迟奖励的回报分解。神经信息处理系统进展，32，2019年。

[2] Chang, F. R. 连续时间随机优化。载于剑桥大学出版社，2004年。[3]Esslinger, K., Platt, R. 和 Amato, C. 用于部分可观测强化学习的深度 Transformer Q 网络。预印本 *arXiv:2206.01078*，2022年。[4] Haarnoja, T., Zhou, A., Abbeel, P. 和 Levine, S. 软演员-评论家：采用随机演员的离策略最大熵深度强化学习。载于第 *35* 届国际机器学习会议论文集，第 80 卷《机器学习研究论文集》，第 1861–1870 页。PMLR，2018年。[5]Holzleitner, M., Gruber, L., Arjona-Medina, J., Brandstetter, J. 和 Hochreiter, S. 应用于近端策略优化和 RUDDER 的演员-评论家方法的收敛性证明。载于大型数据与知识中心系统汇刊 *XLVIII*，第 105–130 页。施普林格，2021年。[6] Hu, B., Zhang, K., Li, N., Mesbahi, M., Fazel, M. 和 Bas¸ar, T. 迈向学习控制策略的策略优化理论基础。载于《控制、机器人与自主系统年度评论》，2022年。

[7] Jiang, F., Chou, G., Chen, M. 和 Tomlin, C. J. 使用神经网络进行快速可达集计算。预印本 *arXiv:611.03158*，2016年。

[8] Kamalapurkar, R., Rosenfeld, J., Dixon, W. E., and Walters, P.基于李雅普诺夫方法的最优反馈控制强化学习。施普林格·查姆，2018年。

[9]Karmakar, P. 和 Bhatnagar, S. 带受控马尔可夫噪声的双时间尺度随机逼近与离策略时序差分学习。数学运筹学，43 (1):130–151，2018年2月。[10] Kim, J.、Shin, J. 和 Yang, I. 面向具有利普希茨连续控制的确定性连续时间系统的哈密顿–雅可比深度 Q 学习。《机器学习研究杂志》，22:1–34，2021年。[11] Liberzon, D. 变分法与最优控制理论：简明导论。普林斯顿大学出版社，2012年。[12] Munos, R. 通过黏性解研究连续情形下的强化学习。《机器学习》，40:265–299，1999年。[13] Munos, R.、Baird, L. 和 Moore, A. 基于梯度下降的方法求解基于神经网络的哈密顿-雅可比-贝尔曼方程。在 IJCNN'99：国际联合神经网络会议论文集（目录号 99CH36339）中，第 3 卷，第 2152–2157 页，第 3 卷，1999年。[14] Nakamura-Zimmerer, T.、Gong, Q. 和 Kang, W. 一种用于高维哈密顿-雅可比-贝尔曼方程的无因果关系神经网络方法。载于*2020*年美国控制会议（*ACC*），第 787–793 页，2020年。[15] Pineda, F. 将反向传播推广至循环神经网络和高阶神经网络。载于神经信息处理系统，1987年。

[16]Raffin, A.、Hill, A.、Gleave, A.、Kanervisto, A.、Ernestus, M. 和 Dormann, N. Stable-Baselines3：可靠的强化学习实现。《机器学习研究杂志》，22(268):1–8，2021年。

[17] Raissi, M., Perdikaris, P., and Karniadakis, G. 物理信息神经网络：用于求解涉及非线性偏微分方程的正问题与反问题的深度学习框架。《计算物理学杂志》，378:686–707，2019年。[18] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. 近端策略优化算法。预印本 *arXiv:1707.06347*，2017年。[19]Shin, Y., Darbon, J., and Karniadakis, G. E. 关于物理信息神经网络在线性二阶椭圆型和抛物型偏微分方程中的收敛性。《计算物理学通讯》，28(5):2042–2074，2020年。ISSN 1991-7120。[20] Sirignano, J. and Spiliopoulos, K. DGM：用于求解偏微分方程的深度学习算法。《计算物理学杂志》，375:1339–1364，2018年。ISSN 0021-9991。[21] Sutton, R. S. and Barto, A. G.《强化学习：导论》（第二版）。美国马萨诸塞州剑桥：麻省理工学院出版社，2018年。[22] Tassa, Y. and Erez, T. 利用神经网络价值函数逼近器求解 HJB方程的最小二乘解。《*IEEE*神经网络汇刊》，18(4):1031–1041，2007年。[23] Tassa, Y., Erez, T., and Todorov, E. 通过在线轨迹优化实现复杂行为的综合与稳定化。载于*2012*年 *IEEE*/*RSJ*智能机器人与系统国际会议，第4906–4913页，2012年。[24] Wiltzer, H. E., Meger, D., and Bellemare, M. G. 面向连续时间强化学习的分布式哈密顿-雅可比-贝尔曼方程。载于第*39*届国际机器学习会议论文集，第162卷，第23832–23856页。PMLR，2022年。

A 超参数

超参数	值
时域 (T)	2048
Adam 步长	3e-04
训练轮数	10
小批量大小	64
折扣因子 (γ)	0.99
广义优势估计参数 (λ)	0.95

表 1：HJBPPO 超参数

环境	值
蚂蚁-v4	0.1
半猎豹-v4	0.1
Humanoid-v4	1e-04
HumanoidStandup-v4	1.0
倒立摆-v4	1e-04
倒立双摆-v4	1e-03
伸手者-v4	1.0
游泳者-v4	1e-04
Hopper-v4	0.1
Walker2d-v4	0.1

表 2： λ_{HJB} 每个环境的超参数