随机演员-评论家：通过时间偶然不确定性缓解高估

预印本

乌尔詹·厄扎尔普土耳其航
空航天公司 安卡拉，土耳其
ugurcanozalp06@gmail.com
ugurcan.ozalp2@tai.com.tr

2026年1月5日

摘要

强化学习中的离策略演员-评论家方法通过时序差分更新训练评论家，并将其作为策略（演员）的学习信号。这种设计通常比纯粹的同策略方法具有更高的样本效率。然而，评论家网络往往会系统性地高估价值估计。对此，常见做法是根据不确定性估计引入悲观偏差。当前方法采用集成来量化评论家的认知不确定性——即由数据有限和模型歧义导致的不确定性——并据此调整悲观更新。在本研究中，我们提出一种名为随机演员-评论家（STAC）的新算法，它将时间（单步）偶然不确定性——由随机转移、奖励以及策略导致的贝尔曼目标变化所引起的不确定性——融入时序差分更新，以此调整悲观偏差，而不依赖认知不确定性。STAC 使用单个分布式评论家网络对时序回报不确定性进行建模，并对评论家和演员网络都采用丢弃法进行正则化。我们的结果表明，仅基于分布式评论家的悲观机制就足以缓解高估，并能在随机环境中自然地产生风险规避行为。进一步引入丢弃法，通过正则化提高了训练稳定性和性能。凭借这一设计，STAC 使用单个分布式评论家网络，实现了更高的计算效率。

关键词 强化学习 · 不确定性 ·高估 · 悲观主义 · 演员-评论家 · 丢弃法

arXiv:2601.00737v1 [cs.LG] 2 Jan 2026

1 引言

演员-评论家方法利用离策略样本训练评论家网络，相较于同策略算法，有望实现更高的样本效率。尽管具有这一优势，演员-评论家智能体在价值函数学习中仍常因函数逼近、时序差分学习和离策略采样的共同作用而面临高估偏差 [Thrun and Schwartz, 2014, Sutton and Barto, 2018, Van Hasselt 等人， 2018], 这可能会使训练不稳定。在这项工作中，我们从一个全新的视角重新审视这一问题，重点关注在演员-评论家架构中使用不确定性建模。

认知不确定性表示训练数据不足，当前的演员-评论家方法利用它来调整悲观更新 [Fujimoto 等人，2018, Haarnoja 等人，2018, Kuznetsov 等人，2020, Moskovitz 等人，2021, Chen 等，2021, Hiraoka 等，2021, Zhang 等人， 2024]。这些方法使用评论家集成来对此类不确定性进行建模。然而，基于评论家网络认知不确定性的悲观更新可能会阻碍对状态-动作空间的探索，这与面对不确定性时的乐观原则 [Kocsis and Szepesvári, 2006, Audibert 等人，2007, Azizzadenesheli 等人，2018, Ciosek 等人，2019, O'Donoghue, 2023, Wu 等人， 2023]的原则相悖。

另一方面，偶然不确定性用于衡量训练数据中的噪声。一些研究提出，基于回报（折扣累积奖励）的认知不确定性和偶然不确定性进行悲观更新，以缓解过高估计 [Kuznetsov 等人， 2020]。最初，对偶然不确定性的建模是分布式

强化学习的研究主题 [Bellemare 等，2017，Dabney 等人，2018a，Kim 和 Oh，2022，Duan 等人，2021，2025，Ma 等人， 2025], ，其主要用途是调整智能体的风险敏感性 [Tang 等，2019，Yang 等，2021，Théate 和 Ernst，2023，Ma 等人， 2025]。悲观的策略更新会导致风险规避行为，而乐观更新则会导致风险寻求行为。最近，在 Q学习设定中，Achab 等人 [2023] 提出对回报的时间偶然分布进行建模，其中不确定性仅由环境的单步动力学引起，而非由所有时间步共同引起。从理论上证明了这种方法具有收敛保证。

在离策略演员-评论家方法中，时序差分目标会随着策略更新而不断变化。尽管给定评论家网络时，策略改进是确定性的，但高维函数逼近会导致小的评论家误差被以非平稳的方式选择性利用，从而使价值目标产生不可消除的变异性，即使在确定性环境中也是如此；这是造成高估的一个主要来源。我们认为，最好将这种变异性视为评论家输出中的时间偶然不确定性，因为它汇总了导致高估的所有不确定性来源：策略诱导效应以及转移或奖励的随机性。

在这项工作中，我们表明，时间偶然不确定性可以通过分布式评论家建模；仅针对这种不确定性施加悲观主义，就足以控制高估，而无需使用评论家网络集成。基于这一想法，我们提出了一种新颖的离策略分布式演员-评论家算法——随机演员-评论家（STAC），专门对评论家的一步时间回报不确定性进行建模，并基于该分布进行悲观更新。这样，STAC 无需双评论家或集成评论家，从而降低计算和内存成本。换言之，STAC 通过面对时间偶然不确定性时采取悲观主义 来缓解高估。

我们的方法不同于其他利用偶然不确定性进行风险规避的方法：Tang等人 [2019], 、Yang等人[2021], 、Kim和Oh [2022], 、Ma 等人 [2025]。这些方法仅在演员更新中针对总回报不确定性引入悲观主义，而 STAC 则在评论家和演员更新中都针对时间回报不确定性引入悲观主义。其主要目的在于缓解高估，而非进行风险规避学习；不过，STAC中的悲观主义仍会带来风险规避效果，与完全分布式方法类似。

Nauman等人 [2024] 近期的研究也表明，网络正则化能够通过减少过拟合对高估产生积极影响。集成方法通过使用多个评论家网络，天然具备正则化效果。由于 STAC 使用单个评论家网络，因此容易受到学习误差和过拟合的影响。为此，演员和评论家网络均采用丢弃法 [（Srivastava等人，2014]）和层归一化 [（Ba等人， 2016] ）作为正则化工具。

实现非常简单，只需在网络中引入丢弃法、对分布式评论家建模，并在软演员-评论家算法 [Haarnoja 等人， 2018] 的基础上定义一个悲观学习目标即可。我们在标准强化学习基准上开展实验，以评估 STAC 相较于现有方法的性能。结果表明，STAC 在所需计算资源更少（单评论家）的情况下，仍能达到与最先进方法相当的性能，证明了其有效性，也使其成为现实世界中的强化学习应用的一种有前景的方法。

2 预备知识

本节介绍分析离策略演员-评论家方法中由不确定性引起的高估现象所需的最低限度背景知识。全文中， $\mathcal{P}(\Omega)$ 表示 Ω 上所有可能的概率分布构成的集合。$A \overset{\mathrm{D}}{=} B$ 表示两个随机变量 A 和 B 具有相同的概率律。

2.1 偶然不确定性与认知不确定性

估计中的不确定性主要分为偶然不确定性或认知不确定性 [Der Kiureghian 和 Ditlevsen, 2009年, Kendall 和 Gal, 2017年, Gal 等人，2016]。认知不确定性指的是由于训练数据不足而导致的模型参数不确定性（θ）。给定训练数据 $\mathcal{D}$ 和模型参数 θ 上的先验分布 $p(\theta)$，我们可以使用贝叶斯推断计算参数的后验分布 $p(\theta \mid \mathcal{D})$。

另一方面，偶然不确定性源于数据内在的随机性。即使拥有无限量的数据，它仍无法避免，也无法降低，因为它是所建模过程的固有组成部分。在深度学习中，我们可以通过构建一个分布式网络来对此建模：该网络根据输入x [输出概率分布 $p_\theta(y|x)$Lakshminarayanan 等人，2017年，Kendall 和 Gal， 2017]。给定数据集 $\mathcal{D}$，用于训练网络的损失函数可以由负对数似然推导得出： $\mathcal{L}_\theta(\mathcal{D}) = \mathbb{E}_{\{(x_i,y_i)\}\sim\mathcal{D}}[-\log\ p_\theta(y_i|x_i)]$。在强化学习中，偶然不确定性自然源于随机奖励、转移动力学，以及时序差分目标中的策略诱导的随机性。

2.2 最大熵演员-评论家强化学习

用强化学习的术语来说，智能体处于一个马尔可夫决策过程（MDP）中，该过程由元组 $\mathcal{M} = (\mathcal{S}, \mathcal{A}, d_0, \tau, R)$表示，其中 $\mathcal{S}$是状态空间， $\mathcal{A}$是动作空间， $d_0 \in \mathcal{P}(\mathcal{S})$是初始状态分布，$\tau\colon \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$是转移核， $R\colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 是奖励函数。

首先对初始状态进行采样， $s_0 \sim d_0(\cdot)$。在时间步 t，处于状态 s_t；根据所采取的动作 $a_t \sim \pi(\cdot \mid s_t)$，从环境中获得下一个状态，$s_{t+1} \sim \tau(\cdot \mid s_t, a_t)$。最后，根据奖励函数 R获得奖励， $r_t = R(s_t, a_t)$。状态回报和状态-动作回报分别定义如下：

$$G^\pi(s) \stackrel{\mathrm{D}}{=} \sum_{t=0}^{\infty} \gamma^t \big(R(s_t, a_t) - \alpha \log \pi(a_t|s_t)\big), \quad s_0 = s, \tag{1}$$

$$Z^\pi(s, a) \stackrel{\mathrm{D}}{=} R(s, a) + \sum_{t=1}^{\infty} \gamma^t \big(R(s_t, a_t) - \alpha \log \pi(a_t|s_t)\big), \quad s_0 = s, \quad a_0 = a, \tag{2}$$

智能体的最终目标是推导出策略 $\pi\colon \mathcal{S} \to \mathcal{P}(\mathcal{A})$，以最大化带有熵奖励 [的折扣回报 G^π Haarnoja 等人， 2017, 2018]。然而，回报是随机变量，不能用作目标。因此，将值（V）和动作值（Q）函数定义为关于策略和转移动力学的回报期望，$V^\pi(s) = \mathbb{E}_{\pi,\tau}\big[G^\pi(s)\big]$, $Q^\pi(s, a) = \mathbb{E}_{\pi,\tau}\big[Z^\pi(s, a)\big]$。演员-评论家方法对状态-动作价值函数 Q 进行建模，学习过程则在策略评估和策略改进之间迭代。
策略评估通过梯度步最小化时序差分： $\big\|\mathcal{T}^\pi Q(s,\ a) - Q(s,\ a)\big\|$ ，其中 $\mathcal{T}^\pi Q$ 是贝尔曼备份算子，定义如下：

$$\mathcal{T}^\pi Q(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim \tau(\cdot|s,a)}\big[Q(s', a') - \alpha \log \pi(a' \mid s')\big]. \tag{3}$$

策略改进通过梯度步最大化价值函数 $V^\pi(s) = \mathbb{E}_{a \sim \pi}\Big[Q^\pi(s,\ a) - \alpha \log\ \pi(a|s)\Big]$。在第 k 步进行策略改进和策略评估后，第 $k+1$ 步更新后的状态-动作价值函数为 $Q^{k+1}(s,\ a) = \mathcal{T}^* Q^k(s,\ a)$，其中 $\mathcal{T}^*$ 是贝尔曼最优算子（Haarnoja 等人公式 5 中的公式 5），

$$\mathcal{T}^* Q(s, a) = R(s, a) + \gamma \mathbb{E}_{s' \sim \tau(\cdot|s,a)}\Big[\tilde{\alpha} \log\Big(\int_{\mathcal{A}} \exp(\tilde{\alpha}^{-1} Q(s', a'))da'\Big)\Big]. \tag{4}$$

其中 $\tilde{\alpha} \in [\alpha, \infty)$ 的定义旨在展示策略改进的连续性，也就是说，策略在每次迭代中都会部分利用状态-动作价值函数进行更新。当 $\tilde{\alpha} \to \infty$时， $\mathcal{T}^* Q$ 归结为 $\mathcal{T}^{\mathcal{U}} Q$，它表示均匀策略下的状态-动作价值。另一方面， $\tilde{\alpha} = \alpha$ 表示完全利用后更新的状态-动作价值。因此，在深度学习场景中，演员网络的学习率越高， $\tilde{\alpha}$就越低，也越接近 α。

3 分布式强化学习与高估

3.1 分布式最大熵演员-评论家

除了学习状态-动作价值函数 Q之外，还可以将状态-动作回报建模为随机变量，以识别给定策略 [Bellemare 等人， 2017]可能产生的结果。这也构成了分布式最大熵演员-评论家方法 [Duan 等人，2021年，Ma 等人， 2025] 的方法论。为此，定义状态-动作回报分布 $\mathcal{Z} \in \mathcal{P}(\mathbb{R}^{\mathcal{S}\times\mathcal{A}})$，其中状态-动作回报从该分布中采样得到， $Z(s,\ a)$ $\sim \mathcal{Z}(s,\ a)$。回顾可知， $Q(s,\ a) = \mathbb{E}_{Z,\ \tau}[Z(s,\ a)]$。相应的贝尔曼备份算子 $\mathcal{T}^\pi$ 定义如下：

$$\mathcal{T}^\pi Z(s, a) \stackrel{\mathrm{D}}{=} R(s, a) + \gamma\big(Z(s', a') - \alpha \log \pi(a' \mid s')\big), \quad s' \sim \tau(\cdot \mid s, a), \quad a' \sim \pi(\cdot \mid s'), \tag{5}$$

为便于表示，定义分布式贝尔曼备份 $\mathcal{T}_D^\pi \mathcal{Z}$，其中，从该分布中采样得到贝尔曼备份的采样值，即 $\mathcal{T}^\pi Z(s,$ $a) \sim \mathcal{T}_D^\pi \mathcal{Z}(s,\ a)$。在每一步 k，通过一个梯度步最小化与分布式贝尔曼备份之间的 Kullback-Leibler 散度，以更新状态-动作回报分布： $D_{\mathrm{KL}}\big(\mathcal{T}_D^\pi \mathcal{Z}(s,\ a)||\mathcal{Z}(s,\ a)\big)$。

单步分布不确定性 与传统的分布建模不同，也可以只对截至第一步的随机性进行建模，从而表示状态-动作回报 Achab [2020],Achab等人 [2023]。在这种方法中，贝尔曼备份使用下一状态和动作的回报期望（确定性值）。这样，环境随机性和演员引起的不确定性（由策略改进导致）仅针对一步进行建模，从而直接针对评论家高估的来源。该分布记为 $\mathcal{Q}$，而非 $\mathcal{Z}$。因此，与先前的表述不同，$Q \sim \mathcal{Q}$ 是一个随机变量。贝尔曼备份的定义与公式 5 类似，但使用下一状态值的期望，

$$\mathcal{T}^{\pi} Q(s,a) \overset{\mathrm{D}}{=} R(s,a) + \gamma\big(\mathbb{E}_{Q\sim\mathcal{Q}}[Q(s',a')] - \alpha \log \pi(a' \mid s')\big), \quad s' \sim \tau(\cdot \mid s,a), \quad a' \sim \pi(\cdot \mid s'). \tag{6}$$

在每一步 k，通过一个梯度步最小化状态-动作价值分布与分布式贝尔曼备份之间的 Kullback-Leibler 散度，从而更新状态-动作价值分布：$D_{\mathrm{KL}}\big(\mathcal{T}_D^{\pi}\mathcal{Q}(s,\ a)||\mathcal{Q}(s,\ a)\big)$。相应的贝尔曼最优算子如下：

$$\mathcal{T}^{*} Q(s,a) \overset{\mathrm{D}}{=} R(s,a) + \gamma \mathbb{E}_{Q\sim\mathcal{Q}}\Big[\tilde{\alpha} \log\Big(\int_{\mathcal{A}} \exp(\tilde{\alpha}^{-1} Q(s',a'))da'\Big)\Big], \quad s' \sim \tau(\cdot \mid s,a). \tag{7}$$

Achab等人 [2023], 证明，该算子在 Q学习 中是一个压缩映射，其中使用 max 替代 logsumexp 算子。这保证了表格设置中的学习收敛。然而，在函数逼近中，还存在高估和过拟合等其他现象。作为本文的主要关注点，我们将讨论单步贝尔曼最优算子的高估界，并在下一节提出一种缓解高估的方法。

3.2 次高斯评论家分布的高估量化

由于策略改进，$\mu(s,\ a)$均值的贝尔曼最优性备份不等于备份的均值$= \mathbb{E}_{\mathcal{Q},\ \tau}[Q(s,\ a)],$，两者之差即为高估偏差。在本节中，我们仿照 Chen 等人 [2021] 和 Lan 等人 [2020] 分析高估偏差，但采用的是软学习框架，而非离散动作。我们将高估误差定义为 $\mathbb{E}_{\mathcal{Q},\ \tau}[\mathcal{T}^{*}Q(s,\ a)]$ 与平均值$\mathcal{T}^{*}\mu(s,\ a)$之差，记为 ϵ，

$$\epsilon(s,a) = \mathbb{E}_{\mathcal{Q},\tau}[\mathcal{T}^{*}Q(s,a)] - \mathcal{T}^{*}\mu(s,a). \tag{8}$$

在理想情况下，如果不存在高估，$\epsilon(s,\ a)$应为零；然而，由于评论家不确定性，实际并非如此。为对此进行量化，我们假设评论家分布 $\mathcal{Q}(s,\ a)$是次高斯分布，其方差代理为 $\sigma^2(s,\ a)$，均值为 $\mu(s,\ a)$。我们采用次高斯假设，因为它提供了一种温和且广泛使用的集中模型，能够在不限制评论家必须服从特定参数分布的前提下，推导出解析形式的高估界。

定义 3.1。 均值为 $\mu = \mathbb{E}[X]$ 的随机变量 $X \in \mathbb{R}$，若其矩生成函数满足下式，则称其为方差代理为 σ^2 的次高斯随机变量

$$\mathbb{E}[\exp(\lambda X)] \le \exp\big(\lambda\mu + \frac{1}{2}\lambda^2\sigma^2\big), \quad \forall \lambda \in \mathbb{R}. \tag{9}$$

在次高斯假设下，定理 3.1 量化了评论家不确定性如何通过 softBellman 最优算子传播，而推论 3.1.1 则为依赖方差的悲观修正提供了依据。

定理 3.1 (次高斯评论家的高估量化)。 设给定的状态-动作价值分布 $\mathcal{Q} \in \mathcal{P}(\mathbb{R}^{\mathcal{S}\times\mathcal{A}})$ 对于所有状态-动作对均为次高斯分布，其均值为 $\mu(s,a)$，方差代理为 $\sigma^2(s,a)$，且具有有界支撑。对于每个样本 $Q \sim \mathcal{Q}$,

$$\mathcal{T}^{*}Q(s,a) \le R(s,a) + \gamma\tilde{\alpha} \log\Big(\int_{\mathcal{A}} \exp(\tilde{\alpha}^{-1}\mu(s',a') + \frac{1}{2}\tilde{\alpha}^{-2}\sigma^2(s',a'))da'\Big). \tag{10}$$

Co论 3.1.1 (悲观评论家目标)。 对于策略评估，使用悲观贝尔曼回 *up,*

$$\mathcal{T}_{\beta}^{\pi}Q(s,\ a) \overset{\mathrm{D}}{=} R(s,a) + \gamma\big(\mu(s',a') - \beta\sigma(s',a') - \alpha\log\pi(a' \mid s')\big), \quad s' \sim \tau(\cdot \mid s,a), \quad a' \sim \pi(\cdot \mid s'), \tag{11}$$

只要 $\beta \ge$ 最大$_{(s,a)}\ \frac{1}{2}\tilde{\alpha}^{-1}\sigma(s,\ a)$，就能防止高估。

此外，定理 3.2 证明了最大高估界。

定理 3.2 (高估界)。 由分布不确定性 $\mathcal{Q}$导致的高估（记为 ϵ）在贝尔曼更新中的上界为，

$$\epsilon(s,a) \leq \frac{\gamma}{2\tilde{\alpha}} \mathbb{E}_{s' \sim \tau}\Big[\max_{a'} \sigma^2(s',a')\Big]. \tag{12}$$

作为补充符号，我们引入悲观分布式贝尔曼备份 $\mathcal{T}^{\pi}_{\beta,\ D}\mathcal{Q}$ 其中，悲观贝尔曼备份中的采样值取自该分布，即 $\mathcal{T}^{\pi}_{\beta}Q(s,\ a) \sim \mathcal{T}^{\pi}_{\beta,\ D}\mathcal{Q}(s,\ a)$。$\mathcal{T}^{\pi}_{\beta,\ D}\mathcal{Q}(s,\ a)$ 与 $\mathcal{T}^{\pi}_{D}\mathcal{Q}(s,\ a)$ 之间的差异如图 1 所示。

综合来看，这些结果表明，离策略演员-评论家方法中的高估源于贝尔曼目标中的时间不确定性，而简单的基于方差代理的悲观偏移便足以对其进行控制。这一观察促成了下文介绍的随机演员-评论家算法。

4 随机演员-评论家

本节将讨论实现计算高效且样本高效的演员-评论家学习所需的关键机制，并提出我们的算法随机演员-评论家。该算法采用一个单一分布式 评论家网络 $\mathcal{Q}_\theta$，其参数集为 θ，用于捕获时间上的（单步）偶然不确定性；同时采用丢弃法和层归一化进行网络正则化。演员网络 π_ϕ 输出经双曲正切变换的正态分布，将动作限制在$[-1,1]$，其参数化形式为参数集 ϕ。不同于其他方法，STAC 利用时间偶然不确定性，以悲观方式使用评论家预测进行策略评估和策略改进。

这一论点的依据是：由于环境随机性和持续进行的演员更新（策略改进）而产生的时序差分目标随机性，大多会从评论家一侧以内在的偶然不确定性形式体现出来，而高估正是这种随机性的结果。尽管给定评论家时，策略改进本身是确定性的，但策略更新会以非平稳的方式选择性地利用评论家的微小近似误差，从而在实践中导致贝尔曼目标出现不可预测的变化。与分布式方法不同， [Bellemare 等人， 2017], 我们不对完整回报的偶然不确定性建模，因为它与高估没有直接关系。

尽管回报不确定性仅针对单步进行建模，STAC 中的悲观主义也会带来风险规避学习，因为评论家也通过悲观更新进行学习。传统的分布式算法将完整回报建模为一个分布。为了实现风险规避学习，通常只以悲观方式更新演员，而不更新评论家，方法是采用条件风险价值（CVaR）度量 [Dabney 等，2018b，Tang 等，2019，Yang 等，2021，Kim 和 Oh，2022，Ma 等人， 2025] 或采用标准差 [Ma 等人， 2025]。

认识论悲观主义通过降低评论家误差被策略改进利用的概率，自然会降低高估程度。然而，分布外状态和动作可通过较高的认知不确定性检测到，不应受到悲观主义的抑制。相反，这些状态-动作对应当通过探索加以访问。因此，我们认为，不应将认知不确定性用作缓解高估的悲观信号。

[Nauman等人， 2024] 证明了网络正则化方法能够提升性能。由于 STAC 采用单个评论家网络，因此比基于集成的方法更容易过拟合。为防止过拟合，STAC 使用丢弃法 [Srivastava等人， 2014]，与 Hiraoka 等人 [2021]的方法类似。虽然丢弃法也可能通过向策略和价值估计中注入随机性来促进探索，但我们在此不试图单独验证或证明这一效果。丢弃法还能够捕捉网络的概率特性，从而体现贝叶斯神经网络 [Gal 和 Ghahramani， 2016]。不过，值得注意的是，STAC 对评论家网络和行动者网络使用丢弃法，仅仅是出于正则化目的，而非用于认知不确定性建模。

4.1 策略评估

为简化起见，STAC 将评论家建模为正态分布。这与定理 3.1 中有界分布的假设相矛盾，但这样可以得到简单的损失函数，也便于解释。不过，对于具有有界奖励函数的有限时域或折扣马尔可夫决策过程（$\gamma < 1$），假设评论家分布有界仍是合理的。

使用经验回放中的转移元组作为批次， $\mathcal{D}_b = \{(s_i,\ a_i,\ r_i,\ s'_i,\ \texttt{done}_i)\}_{i=1}^{N_b}$，表示贝尔曼备份的时序差分目 (TD) 标 Q_i^{TD} 是 β-悲观的，

$$Q_i^{TD} = r_i + \gamma(\mu_{\bar{\theta}}(s'_i,\tilde{a}'_i) - \beta\sigma_{\bar{\theta}}(s'_i,\tilde{a}'_i) - \alpha\log\pi_\phi(s'_i,\tilde{a}'_i))(\neg\texttt{done}_i), \quad \tilde{a}'_i \sim \pi_\phi(\cdot \mid s'_i). \tag{13}$$

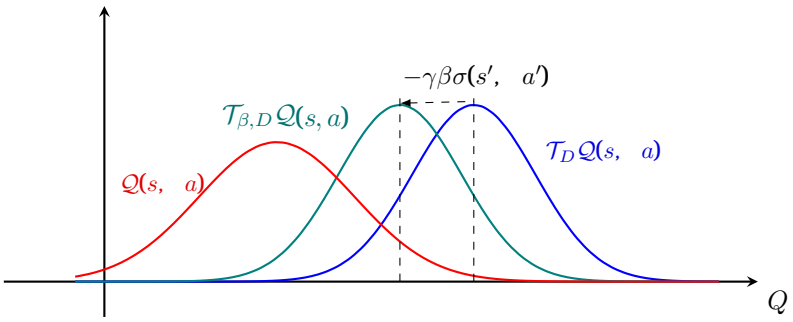


图 1：悲观分布式贝尔曼备份的演变。虽然 $\mathcal{T}_D^{\pi}\mathcal{Q}(s,\ a)$ 被高估，但经过校正的备份 $\mathcal{T}_{\beta,\ D}^{\pi}\mathcal{Q}(s,\ a)$ 更接近 $\mathcal{Q}(s,\ a)$。

学习目标是交叉熵损失（对数损失），

$$\mathcal{L}_\theta(\mathcal{D}_b) = \frac{1}{N_b}\sum_{i=1}^{N_b} -\log \mathcal{Q}_\theta(Q_i^{TD} \mid s_i, a_i) = \frac{1}{2}\log 2\pi + \frac{1}{2N_b}\sum_{i=1}^{N_b}\left(\log \sigma_\theta^2(s_i,a_i) + \frac{(Q_i^{TD} - \mu_\theta(s_i,a_i))^2}{\sigma_\theta^2(s_i,a_i)}\right). \quad (14)$$

4.2 策略改进

根据推论 3.1.1，应使用悲观的时序差分目标来训练评论家网络。然而，这项分析没有考虑策略学习，因为它假设策略是对状态-动作值应用 Softmax 得到的。在演员-评论家框架中，策略改进也应使用相同的悲观目标，因为策略评估和策略改进的目标应当相同。换言之，在学习步骤 k，贝尔曼最优性必须等于采用新策略的贝尔曼备份，$\mathcal{T}^*Q^k \overset{\text{D}}{=} \mathcal{T}^{\pi^{k+1}}Q^k$。只有在策略改进时使用相同的悲观评论家值，才能做到这一点。

4.3 其他细节与算法总结

层归一化STAC 在评论家和演员网络的所有隐藏层激活之后应用层归一化 [Ba等人， 2016]，以稳定学习并防止可能出现的数值不稳定性。这已被证明能显著提升深度强化学习的性能 [Nauman 等人，2024，Hiraoka 等人， 2021]。

用于时序差分目标的滞后评论家当训练好的评论家网络也用于计算目标值时，评论家训练容易发散 [Li等人， 2023]。为此，一种常见做法是使用另一个评论家网络来评估时序差分目标 [Mnih等人， 2013]。与 Lillicrap 等人、 [2015], Fujimoto 等人和 [2018], Haarnoja 等人 [2018], 类似，我们使用延迟版本的评论家网络来评估时序差分目标，如公式 13 所示。目标¯ ¯评论家的参数仅通过在学习过程中对主评论家网络权重进行 Polyak 平均来更新；$\theta \leftarrow \rho\theta + (1-\rho)\theta$。这一策略对于确保时序差分学习的稳定性至关重要。

自动温度调节 使用恒定温度时，奖励幅度的变化会导致策略发生变化。为缓解这一问题，Haarnoja 等人 [2018] 提出了策略熵约束，并将温度表示为¯该约束的拉格朗日乘子。给定作为超参数的目标熵 $\mathcal{H}$，与该约束相关的损失函数如下；

$$\mathcal{L}_\alpha(\mathcal{D}_b) = -\alpha\bar{\mathcal{H}} + \alpha\sum_{i=1}^{N_b}\mathbb{E}_{a\sim\pi_\phi(\cdot|s_i)}\big[-\log \pi_\phi(a \mid s_i)\big]. \quad (15)$$

网络示意图见附录 D。请注意，上划线记号表示带有不可训练参数的滞后网络。STAC 在算法 1 中通过梯度下降进行总结，但在我们的实验中使用 Adam 优化器 [Kingma 和 Ba， 2014]。

算法 1随机演员-评论家

要求： 环境 env
要求： 经验缓冲区 $\mathcal{D}$
要求： 评论家 $\mathcal{Q}_\theta$、滞后评论家 $\mathcal{Q}_{\bar{\theta}}$、演员 π_ϕ，均采用丢弃法
要求： 初始温度 α，目标熵 $\bar{\mathcal{H}}$
要求： 悲观主义 β
要求： 学习率 $\eta_Q, \eta_\pi, \eta_\alpha$，Polyak参数 ρ
要求： 总训练步数 N，批量大小 N_b

$s \sim$env.reset() ▷ 重置环境
对于 N 个时间步**执行**
  $a \sim \pi_\phi(\cdot \mid s)$ ▷ 采样动作
  r， s'，完成 $\sim$ 环境步进(a) ▷ 对环境执行动作
  $\mathcal{D} \leftarrow \mathcal{D} \cup$ （s, a, r, s', 完成） ▷ 记录转移元组
  如果完成**则** $s \leftarrow s'$ **否则** $s \sim$ env.reset() ▷状态转移或重置
  进行 G 梯度步数**次**
    $\mathcal{D}_b = \{$（s_i, a_i, r_i, s'_i,完成$_i$）$\}_{i=1}^{N_b} \sim \mathcal{D}$ ▷ 采样用于训练的小批量
    $\tilde{a}'_i \sim \pi_\phi(\cdot \mid s'_i)$ ▷ 采样下一步动作
    $Q_i^{TD} = r_i + \gamma(\mu_{\bar{\theta}}(s'_i, \tilde{a}'_i) - \beta\sigma_{\bar{\theta}}(s'_i, \tilde{a}'_i) - \alpha$ 对数 $\pi_\phi(\tilde{a}'_i \mid s'_i))(\neg$完成$_i)$ ▷ 构建时序差分目标
    $\theta \leftarrow \theta - \eta_Q \nabla_\theta \Big(\frac{1}{N_b} \sum_{i=1}^{N_b} -$ 日志 $\mathcal{Q}_\theta(Q_i^{TD} \mid s_i, a_i) \Big)$ ▷ 更新评论家
    $\phi \leftarrow \phi - \eta_\pi \nabla_\phi \Big(\frac{1}{N_b} \sum_{i=1}^{N_b} \mathbb{E}_{a \sim \pi_\phi(\cdot \mid s_i)} \big[\mu_\theta(s_i, a) - \beta\sigma_\theta(s_i, a) - \alpha$ 对数 $\pi_\phi(a \mid s_i) \big] \Big)$ ▷ 更新演员
    $\alpha \leftarrow \alpha - \eta_\alpha \nabla_\alpha \Big(-\alpha\bar{\mathcal{H}} + \alpha \sum_{i=1}^{N_b} \mathbb{E}_{a \sim \pi_\phi(\cdot \mid s_i)} \big[-$ 对数 $\pi_\phi(a \mid s_i) \big] \Big)$ ▷ 更新温度
    $\bar{\theta} \leftarrow \rho\bar{\theta} + (1-\rho)\theta$ ▷ 更新目标评论家网络
  结束循环
结束循环

5 实验

使用 Gymnasium API [Towersetal., 2023], ，我们选用六种 MuJoCo 环境和三种 Box2D 环境进行评估，因为它们是文献中的标准基准。Box2D 中的 BipedalWalker-v3 和 BipedalWalkerHardcore-v3 对应在随机生成的地形上行走的双足机器人，而 LunarLander-v3 则模拟在风和湍流影响下着陆的宇宙飞船；我们的实验将风和湍流的影响最大化。因此，与 MuJoCo 环境不同，这些环境本质上具有随机性。

在评估过程中，每经过 1000个时间步，我们便使用在线策略执行一次测试回合，并通过计算该回合累积的总奖励来衡量其表现。在测试回合中，按照文献中的最佳实践，将丢弃法和策略温度设为零。指定的环境经过固定次数的环境交互进行训练，并使用 5 个随机种子重复实验，以评估算法的稳定性。更多实验细节请见附录 C。

各环境对应的超参数见附录 C 的表 3。所有实验均使用 PyTorch（2.7.1版本）[Paszke 等人， 2019] 。代码库请参见附录 E。

5.1 与其他算法的比较

我们的实验旨在探究，在离策略演员-评论家方法中引入 STAC，能否提升其在连续控制基准上的样本效率和计算效率。为此，我们将 STAC 与其他具有竞争力的类似算法进行比较：分布式软演员-评论家（DSAC） [Ma 等人， 2025], 以及软演员-评论家（SAC）[Haarnoja 等人， 2018]。这两种方法在更新时均采用双评论家网络中的较小值。为确保公平比较，所有算法结果均使用内部代码获得，并采用相同的网络架构（包括层归一化，但不包括丢弃法）。

我们还加入了另一种算法，称为认知随机演员-评论家（ESTAC），它采用与 STAC 类似的双分布式评论家和单步不确定性。对于高估，与 SAC 和 DSAC 类似，该算法使用两个评论家中的较小值进行学习，而不是采用时间 aleatoric 悲观主义。其目的是通过以认知过估计缓解取代时间 aleatoric 悲观主义，隔离 STAC 改进的来源。

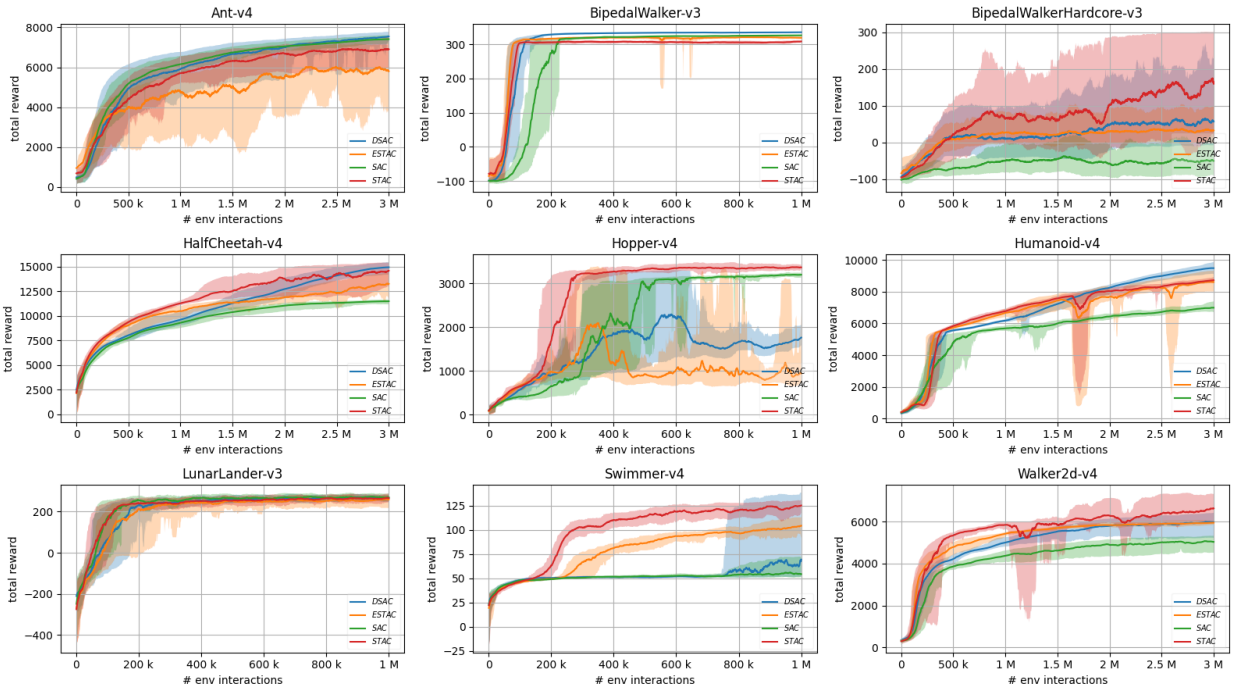


图2：STAC 与其他算法的回合得分曲线。

用于比较的悲观率和丢弃配置是通过消融研究得出的，并将在下一节中说明。最优配置汇总于表 4。

学习曲线 图2展示了 STAC 与前文提到的最先进算法的性能对比。此外，图3呈现了价值估计误差。价值估计误差是通过评估回合期间收集的评论家预测与折扣回报之差来衡量的。粗线表示跨随机种子的四分位距均值，阴影区域表示不同随机种子下总奖励对应的四分位数范围。为便于观察，对曲线进行了时间平滑处理。

F学习过程在随机种子上的均值最终性能总结于表 1.

表1：评估回合最后 %1 的四分位距均值结果，STAC 的 β 和丢弃法参数选择以获得最佳性能。

Env	# 步数	DSAC	ESTAC	SAC	STAC
Ant-v4	3M	**7543.1**	5798.9	7407.1	6907.3
BipedalWalker-v3	1M	**335.3**	319.2	326.3	308.5
BipedalWalkerHardcore-v3	3M	56.8	32.3	-50.8	**159.9**
HalfCheetah-v4	3M	**14921.5**	13241.6	11475.6	14569.3
Hopper-v4	1M	1762.9	944.3	3196.4	**3363.7**
Humanoid-v4	3M	**9484.1**	8608.9	6990.4	8726.2
LunarLander-v3	1M	265.4	258.3	**268.9**	265.7
Swimmer-v4	1M	69.3	104.1	54.2	**125.0**
Walker2d-v4	3M	5981.4	5952.5	5036.4	**6634.7**

样本效率 如图 2 和表 1 所示，在部分环境中，STAC 的样本效率优于其他算法（参见 BipedalWalkerHardcore-v3、Hopper-v4、Swimmer-v4 和 Walker2d-v4），但在其他环境中表现较差。不过，值得注意的是，性能会受到许多因素的影响，例如分布建模方法、评论家网络的数量、正则化方法、网络规模和学习率等。我们不仅关注绝对性能，还表明，STAC 在使用单评论家的同时，无收敛问题地实现了有竞争力甚至更优的性能；相比之下，其他方法则依赖双评论家。

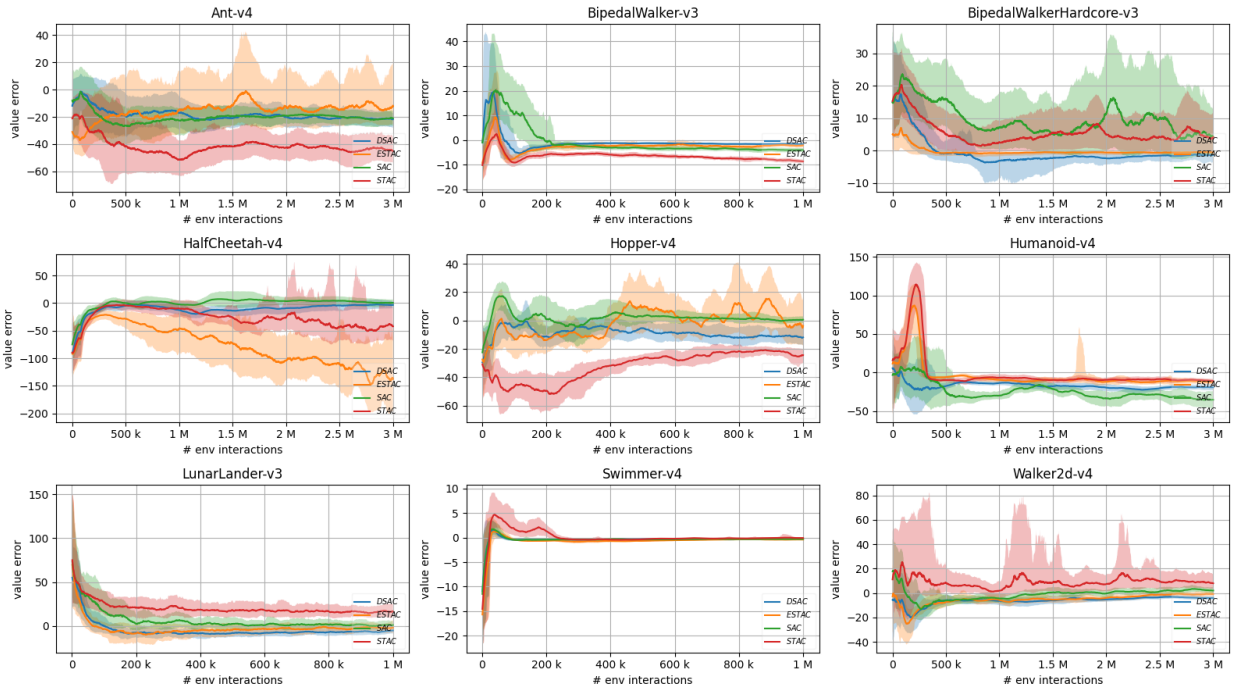


图3：STAC 与其他算法的回合平均价值估计误差。

ESTAC 与 STAC 的比较 除 BipedalWalker-v3 外，STAC 在所有环境中的表现都始终优于 ESTAC。这表明，至少在我们的实验设置中，时间 aleatoric 悲观主义通常比以最小值裁剪形式实现的认识论悲观主义更有效。

5.2 扫描悲观程度与丢弃法

在不同的悲观程度 β 和丢弃率下对 STAC 进行实验，以分别考察悲观主义和丢弃法的影响。我们使用了 5 个悲观程度和 4 种丢弃配置：不使用丢弃法、演员丢弃、评论家丢弃，以及演员与评论家丢弃。

表2：不同 β 设置下，最后 %1 个评估回合的四分位距均值分数。评论家和演员的丢弃率均为 0.01。

Env	$\beta = 0.0$	$\beta = 0.125$	$\beta = 0.25$	$\beta = 0.375$	$\beta = 0.5$
Ant-v4	5264.4	**7082.7**	6907.3	6246.4	4834.4
BipedalWalker-v3	171.9	125.4	262.6	**308.0**	305.1
BipedalWalkerHardcore-v3	61.8	**118.8**	43.4	-0.0	-59.8
HalfCheetah-v4	**14489.5**	12299.6	13385.4	11042.1	7602.7
Hopper-v4	1132.4	1308.8	2643.4	3318.2	**3363.7**
Humanoid-v4	5623.9	**8407.2**	8277.9	7587.2	6930.8
LunarLander-v3	**265.7**	263.8	157.2	14.3	-77.7
Swimmer-v4	**103.7**	93.5	70.0	54.2	45.9
Walker2d-v4	4299.2	**6306.4**	5842.1	5795.2	5539.1

学习曲线 在实验中，我们开展了受控研究，以隔离丢弃法的作用：分布式悲观主义能够消除高估偏差，而丢弃法主要提升训练稳定性和性能。两条学习曲线均可见于附录 B。在演员/评论家联合丢弃法下，不同悲观主义程度对应的回合得分及相关价值估计误差汇总于图 5 和图 6。此外，在相同悲观主义程度下，不同丢弃配置对应的相同曲线汇总于图 7 和图 8，其中 β 值见表 4。

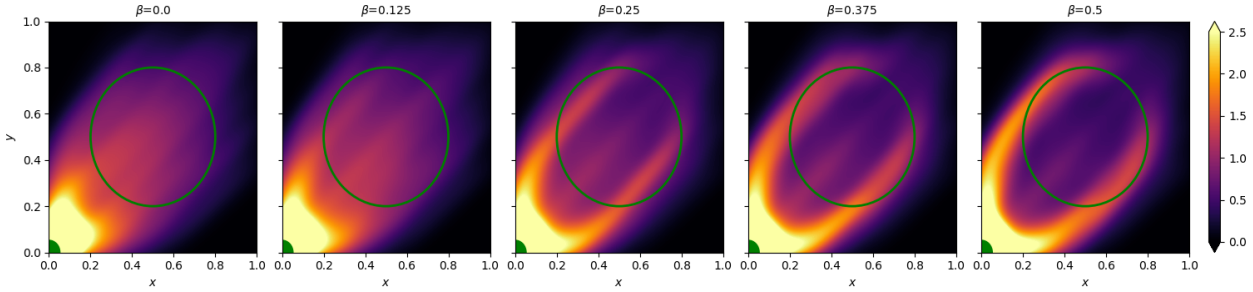


图 4：在不同 β 值下，STAC 在 RiskyPointMass-v0 上的位置出现密度热力图。中间的绿色圆圈表示危险区域的边界。当点质量进入左下方的绿色区域时，回合即告终止。

阴影区域表示四分位数界限，实线表示在不同随机种子以及 %1 个回合评估平滑窗口下，随环境步数变化的四分位距均值。

悲观主义 表2总结了最后 %1 个评估回合的四分位距均值得分表现。结果表明，β 是一个敏感参数；β 越高，误差越趋于负值（见图 6），这与预期效果一致。此外，如果值估计趋于正值，得分曲线会更差，说明评论家高估未得到充分缓解（见 Ant-v4、BipedalWalker-v3、Hopper-v4、Humanoid-v4、Walker-v4）。另一方面，当误差曲线比预期更负时，得分曲线也会再次变差，这意味着智能体受困于高悲观主义导致的评论家低估（见 BipedalWalkerHardcore-v3、HalfCheetah-v4、LunarLander-v3、Swimmer-v4）。总之，这种敏感性因环境而异，可能是由于奖励稀疏性和随机性的差异。该参数是 STAC 的主要瓶颈，目前只能通过这种启发式方法确定。

丢弃法 丢弃法在 STAC 中的主要作用是作为一种正则化工具，以提升优化稳定性以及对噪声数据的鲁棒性。在图 7 中，无论是否采用评论家丢弃，演员丢弃都能通过稳定策略更新来持续提升性能。根据图 8，采用和不采用丢弃法时观察到的高估趋势在定性上相似，这证实了我们的核心贡献——通过偶然不确定性缓解高估——不受是否采用丢弃法的影响，但 Hopper-v4 除外。然而，在该环境中，若不使用评论家丢弃，学习甚至无法收敛。因此，在某些环境中，丢弃法是确保训练稳定所必需的，而非用于缓解高估。双重/集成评论家方法能够自然地缓解训练不稳定，而 STAC 在某些情况下则需要使用适当的丢弃法。

环境随机性下的悲观主义 基于偶然不确定性的悲观学习可能与探索相互作用，尤其是在具有复杂动力学的环境中。比较 BipedalWalker-v3 和 BipedalWalkerHardcore-v3 后，我们发现，在较难的环境中，较低的悲观程度效果更好；而在较简单的环境中，适度的悲观程度能够提升性能（见图 5 和图 6）。一种可能的解释是，在较简单的任务中，不确定性主要源于近似和自举效应，而悲观主义有助于减少高估。相比之下，在更具挑战性的环境中，不确定性可能更多地由环境本身驱动，过度悲观则可能因阻碍探索而减缓学习。这些观察表明，合适的悲观程度取决于任务复杂性与学习动态之间的相互作用。

5.3 悲观主义带来的风险敏感性

在 STAC 中，演员更新针对单步回报不确定性采取悲观策略。然而，评论家目标对于连续步骤也同样采取悲观策略。因此，悲观主义也会与奖励一同进行自举，预期随着 β 增加，风险规避行为也会增强。为此，我们使用一个玩具问题来展示这一行为。

环境详情 按照 Ma 等人 [2021] 和 Ma 等人 [2025], 的设定，我们采用风险导航任务作为一个简单的随机问题。该环境名为 RiskyPointMass-v0。此任务的目标是在二维空间中操控一个点质量前往目标点，同时避开危险区域。危险区域是一个圆形区域，半径为 $r_0 = 0.3$，中心位于点 (0.5,0.5)。初始状态的两个坐标均从 $\mathcal{U}$(0.3,1) 中随机选取，并排除危险区域。若智能体与目标点 (0,0) 的欧氏距离小于 0.05，则回合终止。每一步中，智能体获得三个奖励分量之和：到目标点的负欧氏距离

、用于促使回合终止的固定 -0.1 点奖励，以及当智能体位于危险区域内时的 -10 点惩罚，该情况发生的概率为 0.1 $\exp(-4r^2/r_0^2)$，其中 r 是到危险区域中心的距离。智能体在两个维度上的移动幅度最多为 0.1 个单位，另加最多 0.02 个单位的随机滑移。奖励函数经过设计，使得最优风险中性解是沿直线前往目标点，而最优风险规避解则是绕过危险区域的圆形边界。

结果 对于每个 β 值，STAC 均使用一个固定的随机种子运行，因为本实验旨在定性展示风险敏感行为，而非比较性能。在图 4 中，展示了针对不同 β 值训练的智能体的位置密度。较大的绿色圆圈表示危险区域边界，左下方的绿色四分之一圆表示目标点。本演示使用随机初始点运行了 500 个评估步。显然，随着 β 值增大，智能体更倾向于绕过圆圈，而不是像使用较低 β值训练的智能体那样行动。这一效果符合预期，因为悲观性通过时序差分目标进行自举，从而在滚动展开过程中有效诱导出保守行为。总而言之，STAC 不仅缓解了高估，还通过对时间不确定性持悲观态度，促成了风险规避学习。

6 相关工作

通过认知不确定性降低高估 在相关文献中，高估问题通过基于评论家网络认知不确定性 的悲观学习来解决 [Chen 等，2021，Hiraoka 等人， 2021]。基于模型的强化学习方法也采用了相同的方法 [Janner等，2019，Chua等，2018，Depeweg 等人， 2016]。近年来，对于离策略无模型演员-评论家算法，认知不确定性通过双网络 [Fujimoto等，2018，Haarnoja 等人， 2018], 或集成网络 [Chen等，2021，Moskovitz 等人， 2021], 或丢弃法 [Hiraoka 等人，2021]进行估计。由于需要优化多个参数，集成评论家网络的计算开销较大。上述方法在策略评估和改进中采用固定程度的悲观性，而 Moskovitz 等人[2021] 则将悲观性作为多臂老虎机问题处理，着重于动态调整 悲观性，而非将其固定。类似地，Cetin 和 Celiktutan[2023] 通过将悲观性视为对偶变量来在线调整它，并强制使预期的离策略动作价值偏差为零。

通过偶然不确定性降低高估 Kuznetsov 等人 [2020] 声称，偶然不确定性 也是造成高估的原因，因为采用贝尔曼最优算子（$\mathcal{T}^*$）时，任何随机性都会被利用。为此，他们使用集成网络来建模认知不确定性，其中每个网络都是用于偶然不确定性建模的分布式网络（以分位数建模），并利用这两种不确定性来校正高估。同样，STAC 使用分布式评论家，但只对时序（单步）偶然不确定性建模，而不对完整回报分布建模，因为前者是高估的主要来源；该方法不进行认知不确定性建模。

风险敏感强化学习 价值函数的偶然不确定性表征至关重要，尤其是在存在近似的情况下 [Bellemare 等人， 2017]。这可以通过原子[Bellemare 等人， 2017], 分位数 [Dabney 等人，2018a，Ma 等人， 2025], 或正态分布 [Tang 等，2019，Yang 等， 2021] 来实现，以对累积回报行为建模。对于安全关键型强化学习应用，为避免灾难性情形，会采用偶然不确定性悲观主义 [Tang 等，2019，Yang 等，2021，Lim 和 Malik，2022，Greenberg 等，2022，Ma 等， 2025]。这些方法通过对完整回报不确定性建模，仅在演员更新时采用悲观主义来调节风险敏感性；而 STAC 则通过针对时序回报不确定性的悲观更新，同时调节演员和评论家，以缓解高估。

7 结论与未来方向

本文介绍了一种新颖的离策略演员-评论家算法——随机演员-评论家（STAC）。其核心思想是利用时序偶然评论家不确定性，将悲观学习目标纳入算法，以缓解高估问题，从而实现更快、更稳健的学习。为此，我们将评论家建模为分布式神经网络。尽管 STAC 使用正态分布，但我们的分析适用于所有次高斯评论家表示。

我们推导出了高估的上界，证明适度的悲观主义能够缓解高估而不致造成低估，从而促进计算与样本高效学习。此外，这种方法还带来了风险规避学习过程。最后，我们提出在演员和评论家网络中使用丢弃法来缓解过拟合。演员丢弃法主要能提升性能，而在评论家中使用丢弃法则仅是确保学习收敛所必需的。

自适应悲观与正则化 虽然 STAC 证明，仅凭时间 aleatoric 悲观主义就足以通过单评论家缓解高估，但仍有若干方向有待探索。具体而言，悲观系数 β 目前是通过经验选取的，其最优值因环境而异，因此未来可以研究自适应或有原则的调参策略，类似于 Moskovitz 等人 [2021], CetinandCeliktutan [2023]。此外，尽管 STAC 通过避免使用双评论家降低了计算开销，但在某些环境中，单评论家设置可能需要显式正则化（例如丢弃法）以实现稳定优化。未来工作可以考虑其他正则化方法（包括集成），以实现更稳定的训练。

分布建模 悲观主义对高度随机环境的影响也是一个重要的研究课题。如图所示，悲观更新会导致风险规避行为。我们的结果仅适用于简单的随机任务，值得在更复杂的任务中对 STAC 进行实证研究。对于更先进的建模，可以使用分位数来建模评论家的偶然不确定性 [Dabney 等人，2018a，Ma 等人， 2025]或使用流匹配 [Chen 等人，2025年，Dong 等人， 2025], 而不是正态分布。

探索 最后，当前方法使用评论家的认知不确定性来解决高估问题，这与面对不确定性时的乐观原则 相矛盾。未来的研究或可考虑利用评论家和演员的认知不确定性进行探索，同时保留时间偶然不确定性悲观主义。

更广泛的影响 STAC 对于理解离策略演员-评论家学习中的高估现象作出了重要贡献。它表明，评论家高估也是分布式学习中的一个研究课题，而解决方案是为演员和评论家设计风险规避目标。此外，它还证明，对于这一问题，完全没有必要对认知不确定性进行建模，从而直接提高了计算效率。

致谢
本研究未获得外部资助。

参考文献

Sebastian Thrun 和 Anton Schwartz。《在强化学习中使用函数逼近的问题》。载于*1993*年联结主义模型暑期学校论文集，第255–263页。Psychology Press，2014。Richard S Sutton 和 Andrew G Barto。《强化学习：导论》。麻省理工学院出版社，2018。Hado Van Hasselt、Yotam Doron、Florian Strub、Matteo Hessel、Nicolas Sonnerat 和 Joseph Modayil。《深度强化学习与致命三元组》。*arXiv*预印本 *arXiv:1812.02648*，2018。Scott Fujimoto、Herke Hoof 和 David Meger。《解决演员-评论家方法中的函数逼近误差》。载于国际机器学习会议，第1587–1596页。PMLR，2018。Tuomas Haarnoja、Aurick Zhou、Kristian Hartikainen、George Tucker、Sehoon Ha、Jie Tan、Vikash Kumar、Henry Zhu、Abhishek Gupta、Pieter Abbeel 等。《软演员-评论家算法及其应用》。*arXiv*预印本 *arXiv:1812.05905*，2018。Arsenii Kuznetsov、Pavel Shvechikov、Alexander Grishin 和 Dmitry Vetrov。《利用连续分布分位数评论家的截断混合控制高估偏差》。载于国际机器学习会议，第5556–5566页。PMLR，2020。Ted Moskovitz、Jack Parker-Holder、Aldo Pacchiano、Michael Arbel 和 Michael Jordan。《深度强化学习中的战术性乐观与悲观》。神经信息处理系统进展，34：12849–12863，2021。Xinyue Chen、Che Wang、Zijian Zhou 和 Keith Ross。《随机集成双Q学习：无需模型的快速学习》。*arXiv*预印本 *arXiv:2101.05982*，2021。Takuya Hiraoka、Takahisa Imagawa、Taisei Hashimoto、Takashi Onishi 和 Yoshimasa Tsuruoka。《用于双重高效强化学习的随机失活Q函数》。*arXiv*预印本 *arXiv:2110.02034*，2021。Junwei Zhang、Shuai Han、Xi Xiong、Sheng Zhu 和 Shuai Lü。《Explorer-actor-critic：更优的深度强化学习演员》。信息科学，662：120255，2024。Levente Kocsis 和 Csaba Szepesvári。《基于老虎机的蒙特卡洛规划》。载于欧洲机器学习会议，第282–293页。施普林格，2006。Jean-Yves Audibert、Rémi Munos 和 Csaba Szepesvári。《随机环境中的老虎机算法调优》。载于国际算法学习理论会议，第150–165页。施普林格，2007。

Kamyar Azizzadenesheli、Emma Brunskill 和 Animashree Anandkumar。《通过贝叶斯深度Q网络实现高效探索》。载于*2018*年信息论与应用研讨会（*ITA*），第1–9页。IEEE，2018。Kamil Ciosek、Quan Vuong、Robert Loftin 和 Katja Hofmann。《通过乐观演员-评论家实现更优探索》。神经信息处理系统进展，32，2019。Brendan O'Donoghue。《通过认知风险偏好策略优化实现高效探索》。*arXiv* 预印本 *arXiv:2302.09339*，2023。Xinyang Wu、Mohamed El-Shamouty、Christof Nitsche 和 Marco F Huber。《利用贝叶斯神经网络实现不确定性引导的主动强化学习》。载于*2023*年*IEEE*机器人与自动化国际会议（*ICRA*），第5751–5757页。IEEE，2023。Marc G Bellemare、Will Dabney 和 Ré mi Munos。《从分布视角看强化学习》。载于国际机器学习会议，第449–458页。PMLR，2017。Will Dabney、Mark Rowland、Marc Bellemare 和 Rémi Munos。《基于分位数回归的分布式强化学习》。载于*AAAI*人工智能会议论文集，第32卷，2018a。Dohyeong Kim 和 Songhwai Oh。《利用信赖域条件风险价值实现高效的离策略安全强化学习》。*IEEE*机器人与自动化快报，7(3)：7644–7651，2022。Jingliang Duan、Yang Guan、Shengbo Eben Li、Yangang Ren、Qi Sun 和 Bo Cheng。《分布式软演员-评论家：用于解决价值估计误差的离策略强化学习》。*IEEE*神经网络与学习系统汇刊，33(11)：6584–6598，2021。Jingliang Duan、Wenxuan Wang、Liming Xiao、Jiaxin Gao、Shengbo Eben Li、Chang Liu、Ya-Qin Zhang、Bo Cheng 和 Keqiang Li。《具有三项改进的分布式软演员-评论家》。*IEEE*模式分析与机器智能汇刊，2025。Xiaoteng Ma、Junyao Chen、Li Xia、Jun Yang、Qianchuan Zhao 和 Zhengyuan Zhou。《DSAC：用于风险敏感强化学习的分布式软演员-评论家》。人工智能研究杂志，83，2025。Yichuan Charlie Tang、Jian Zhang 和 Ruslan Salakhutdinov。《最坏情形策略梯度》。*arXiv*预印本 *arXiv:1911.03618*，2019。Qisong Yang、Thiago D Simão、Simon H Tindemans 和 Matthijs TJ Spaan。《WCSAC：面向安全约束强化学习的最坏情形软演员-评论家》。载于*AAAI*人工智能会议论文集，第35卷，第10639–10646页，2021。Thibaut Théate 和 Damien Ernst。《采用分布式强化学习的风险敏感策略》。算法，16(7)：325，2023。Mastane Achab、Reda Alami、Yasser Abdelaziz Dahou Djilali、Kirill Fedyanin 和 Eric Moulines。《单步分布式强化学习》。*arXiv*预印本 *arXiv:2304.14421*，2023。Michal Nauman、Michał Bortkiewicz、Piotr Miło´s、Tomasz Trzci´ nski、Mateusz Ostaszewski 和 Marek Cygan。《演员-评论家中的高估、过拟合与可塑性：强化学习的苦涩教训》。*arXiv*预印本 *arXiv:2403.00514*，2024。Nitish Srivastava、Geoffrey Hinton、Alex Krizhevsky、Ilya Sutskever 和 Ruslan Salakhutdinov。《丢弃法：一种防止神经网络过拟合的简单方法》。机器学习研究杂志，15(1)：1929–1958，2014。Jimmy Lei Ba、Jamie Ryan Kiros 和 Geoffrey E Hinton。《层归一化》。*arXiv*预印本 *arXiv:1607.06450*，2016。Armen Der Kiureghian 和 Ove Ditlevsen。《偶然不确定性还是认知不确定性？这重要吗？》。结构安全，31(2)：105–112，2009。Alex Kendall 和 Yarin Gal。《计算机视觉的贝叶斯深度学习需要哪些不确定性？》。神经信息处理系统进展，30，2017。Yarin Gal 等人。《深度学习中的不确定性》。2016。Balaji Lakshminarayanan、Alexander Pritzel 和 Charles Blundell。《利用深度集成进行简单、可扩展的预测不确定性估计》。神经信息处理系统进展，30，2017。Tuomas Haarnoja、Haoran Tang、Pieter Abbeel 和 Sergey Levine。《基于深度能量策略的强化学习》。载于国际机器学习会议，第1352–1361页。PMLR，2017。Mastane Achab。《排序与风险感知强化学习》。博士论文，巴黎综合理工学院，2020。Qingfeng Lan、Yangchen Pan、Alona Fyshe 和 Martha White。《最大最小Q学习：控制Q学习的估计偏差》。*arXiv*预印本 *arXiv:2002.06487*，2020。

Will Dabney、Georg Ostrovski、David Silver 和 Rémi Munos。用于分布式强化学习的隐式分位数网络。载于国际机器学习会议，第 1096–1105 页。PMLR，2018b。Yarin Gal 和 Zoubin Ghahramani。将丢弃法作为贝叶斯近似：在深度学习中表示模型不确定性。载于 第 *33* 届国际机器学习会议论文集（*ICML-16*），2016年。Sicen Li、Qinyun Tang、Yiming Pang、Xinmeng Ma 和 Gang Wang。现实演员-评论家：实现价值高估与低估之间平衡的框架。《style id='5'>神经机器人学前沿》，16:1081242，2023年。Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Alex Graves、Ioannis Antonoglou、Daan Wierstra 和 Martin Riedmiller。用深度强化学习玩雅达利游戏。*arXiv* 预印本 *arXiv:1312.5602*，2013年。Timothy P Lillicrap、Jonathan J Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。基于深度强化学习的连续控制。*arXiv* 预印本 *arXiv:1509.02971*，2015年。Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*arXiv* 预印本 *arXiv:1412.6980*，2014。Mark Towers、Jordan K. Terry、Ariel Kwiatkowski、John U. Balis、Gianluca de Cola、Tristan Deleu、Manuel Goulão、Andreas Kallinteris、Arjun KG、Markus Krimmel、Rodrigo Perez-Vicente、Andrea Pierré、Sander Schulhoff、Jun Jet Tai、Andrew Tan Jin Shen 和 Omar G. Younis。Gymnasium，2023年3月。网址：https://zenodo.org/record/ 8127025。Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、Zeming Lin、Natalia Gimelshein、Luca Antiga 等。PyTorch：一种命令式风格的高性能深度学习库。神经信息处理系统进展，32，2019年。Yecheng Ma、Dinesh Jayaraman 和 Osbert Bastani。保守离线分布式强化学习。神经信息处理系统进展，34:19235–19247，2021年。Michael Janner、Justin Fu、Marvin Zhang 和 Sergey Levine。何时信任你的模型：基于模型的策略优化。神经信息处理系统进展，32，2019年。Kurtland Chua、Roberto Calandra、Rowan McAllister 和 Sergey Levine。利用概率动力学模型，通过少量试验进行深度强化学习。神经信息处理系统进展，31，2018年。Stefan Depeweg、José Miguel Hernández-Lobato、Finale Doshi-Velez 和 Steffen Udluft。利用贝叶斯神经网络，在随机动力系统中进行学习与策略搜索。*arXiv* 预印本 *arXiv:1605.07127*，2016年。Edoardo Cetin 和 Oya Celiktutan。为强化学习学习悲观主义。载于*AAAI*人工智能会议论文集，第 37 卷，第 6971–6979 页，2023年。Shiau Hong Lim 和 Ilyas Malik。面向风险敏感策略的分布式强化学习。神经信息处理系统进展，35:30977–30989，2022年。Ido Greenberg、Yinlam Chow、Mohammad Ghavamzadeh 和 Shie Mannor。高效的风险规避强化学习。神经信息处理系统进展，35:32639–32652，2022年。Deshu Chen、Yuchen Liu、Zhijian Zhou、Chao Qu 和 Yuan Qi。借助分布式评论家释放流策略。*arXiv* 预印本 *arXiv:2509.23087*，2025年。Perry Dong、Chongyi Zheng、Chelsea Finn、Dorsa Sadigh 和 Benjamin Eysenbach。价值流。*arXiv* 预印本 *arXiv:2510.07650*，2025年。

附录 A 证明

定理 *3.1* 的证明。分析贝尔曼更新 $\mathcal{T}^*Q(s,\ a)$,

$$\begin{aligned}\mathcal{T}^*Q(s,a) &= R(s,a) + \gamma\mathbb{E}_{Q\sim\mathcal{Q}}\Big[\tilde{\alpha}\log\Big(\int_{\mathcal{A}}\exp(\tilde{\alpha}^{-1}Q(s',a'))da'\Big)\Big] \\ &\leq R(s,a) + \gamma\tilde{\alpha}\log\Big(\mathbb{E}_{Q\sim\mathcal{Q}}\Big[\int_{\mathcal{A}}\exp(\tilde{\alpha}^{-1}Q(s',a'))da'\Big]\Big) \\ &= R(s,a) + \gamma\tilde{\alpha}\log\Big(\int_{\mathcal{A}}\mathbb{E}_{Q\sim\mathcal{Q}}\Big[\exp(\tilde{\alpha}^{-1}Q(s',a'))\Big]da'\Big) \\ &\leq R(s,a) + \gamma\tilde{\alpha}\log\Big(\int_{\mathcal{A}}\exp(\tilde{\alpha}^{-1}\mu(s',a') + \frac{1}{2}\tilde{\alpha}^{-2}\sigma^2(s',a'))da'\Big)\end{aligned}$$

第一个不等式根据詹森不等式得出（利用对数函数的凹性），随后等式则由托内利定理得出。第二个不等式源自次高斯假设 3.1。 □

推论 *3.1.1* 的证明。 根据定理 3.1，我们可以证明

$$\begin{aligned}\mathcal{T}^*_\beta Q(s,a) &= R(s,a) + \gamma\mathbb{E}_{Q\sim\mathcal{Q}}\Big[\tilde{\alpha}\log\Big(\int_{\mathcal{A}}\exp(\tilde{\alpha}^{-1}Q(s',a') - \beta\sigma(s',a'))da'\Big)\Big] \\ &\leq R(s,a) + \gamma\tilde{\alpha}\log\Big(\int_{\mathcal{A}}\exp(\tilde{\alpha}^{-1}(\mu(s',a') - \beta\sigma(s',a') + \frac{1}{2}\tilde{\alpha}^{-1}\sigma^2(s',a')))da'\Big) \\ &= R(s,a) + \gamma\tilde{\alpha}\log\Big(\int_{\mathcal{A}}\exp(\tilde{\alpha}^{-1}\mu^\dagger(s',a'))da'\Big) = \mathcal{T}^*\mu^\dagger(s,a).\end{aligned}$$

其中，我们定义了 $\mu^\dagger(s',\ a') = \mu(s',\ a') - \beta\sigma(s',\ a') + \frac{1}{2}\tilde{\alpha}^{-1}\sigma^2(s',\ a')$。如果 $\beta \geq \max_{(s',a')}\frac{1}{2}\tilde{\alpha}^{-1}\sigma(s',\ a')$，那么$\mu^\dagger(s',\ a') < \mu(s',\ a')$。因此，我们可以证明

$$\mathcal{T}^*_\beta Q(s,a) \leq \mathcal{T}^*\mu^\dagger(s,a) \leq \mathcal{T}^*\mu(s,a). \tag{16}$$

□

定理 *3.2* 的证明。 根据定理 3.1，我们可以证明

$$\begin{aligned}\mathbb{E}_{s'\sim\tau}[\mathcal{T}^*Q(s,a)] &\leq R(s,a) + \gamma\mathbb{E}_{s'\sim\tau}\Big[\tilde{\alpha}\log\Big(\int_{\mathcal{A}}\exp(\tilde{\alpha}^{-1}\mu(s',a') + \frac{1}{2}\tilde{\alpha}^{-2}\sigma^2(s',a'))da'\Big)\Big] \\ &\leq R(s,a) + \gamma\mathbb{E}_{s'\sim\tau}\Big[\tilde{\alpha}\log\Big(\big(\int_{\mathcal{A}}\exp(\tilde{\alpha}^{-1}\mu(s',a'))da'\big)\cdot\big(\max_{a'}\exp(\frac{1}{2}\tilde{\alpha}^{-2}\sigma^2(s',a'))\big)\Big)\Big] \\ &= R(s,a) + \gamma\mathbb{E}_{s'\sim\tau}\Big[\tilde{\alpha}\log\Big(\int_{\mathcal{A}}\exp(\tilde{\alpha}^{-1}\mu(s',a'))da'\Big)\Big] + \frac{\gamma}{2\tilde{\alpha}}\mathbb{E}_{s'\sim\tau}\Big[\max_{a'}\sigma^2(s',a')\Big].\end{aligned}$$

第二个不等式是由积分中值定理得出的。在最后一个等式中，前两项等于 $\mathcal{T}^*\mu(s,a)$。因此,

$$\epsilon(s,a) = \mathbb{E}_{s'\sim\tau}[\mathcal{T}^*Q(s,a)] - \mathcal{T}^*\mu(s,a) \leq \frac{\gamma}{2\tilde{\alpha}}\mathbb{E}_{s'\sim\tau}\Big[\max_{a'}\sigma^2(s',a')\Big].$$

□

附录 B 消融研究结果

B.1 悲观性消融

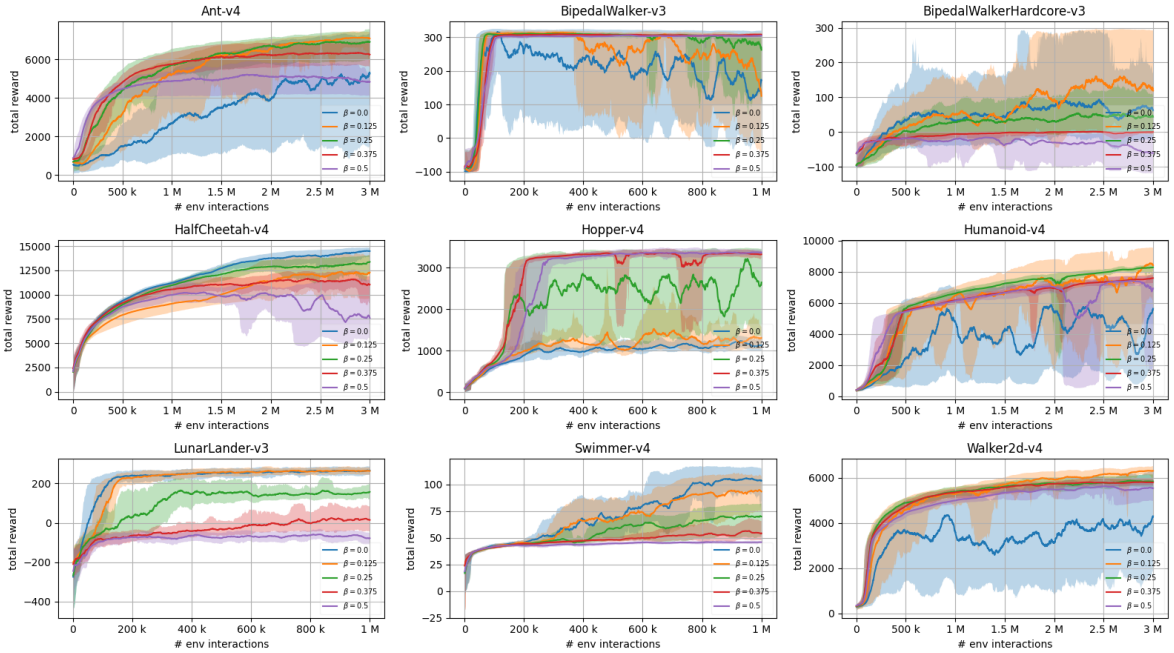


图 5：STAC 的周期得分曲线，其中悲观主义（β）参数取不同值，演员和评论家丢弃法均设为 0。01。

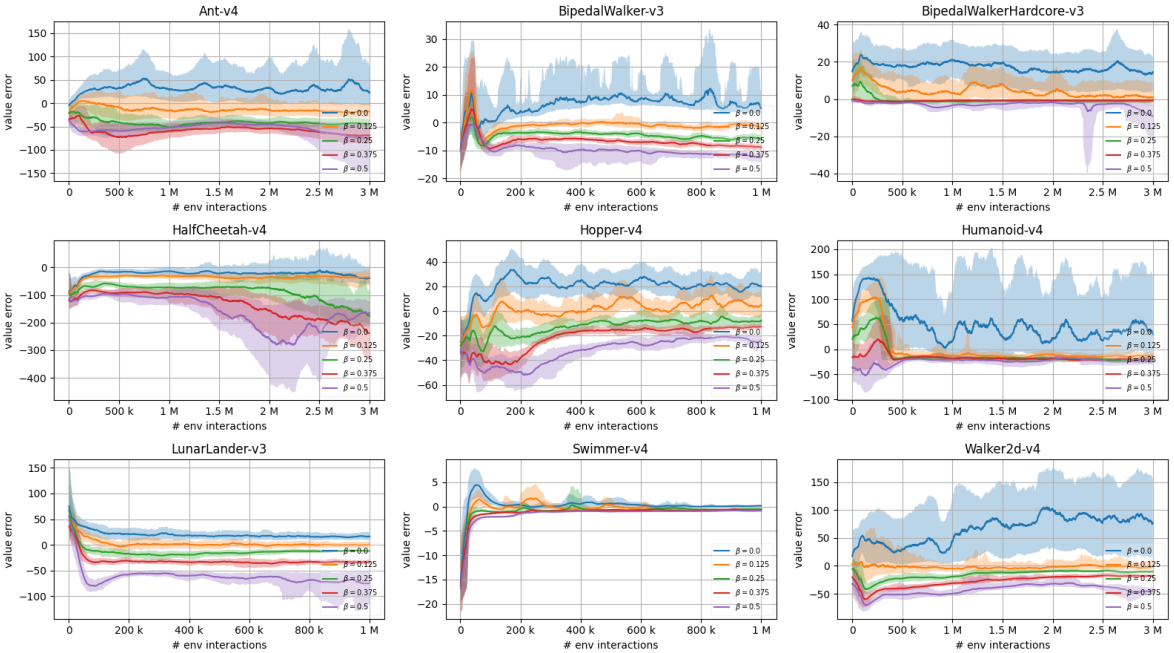


图 6：STAC 在不同悲观主义（β）参数下的平均回合值估计误差曲线，演员和评论家的丢弃率均为 0。01。

B.2 丢弃配置消融

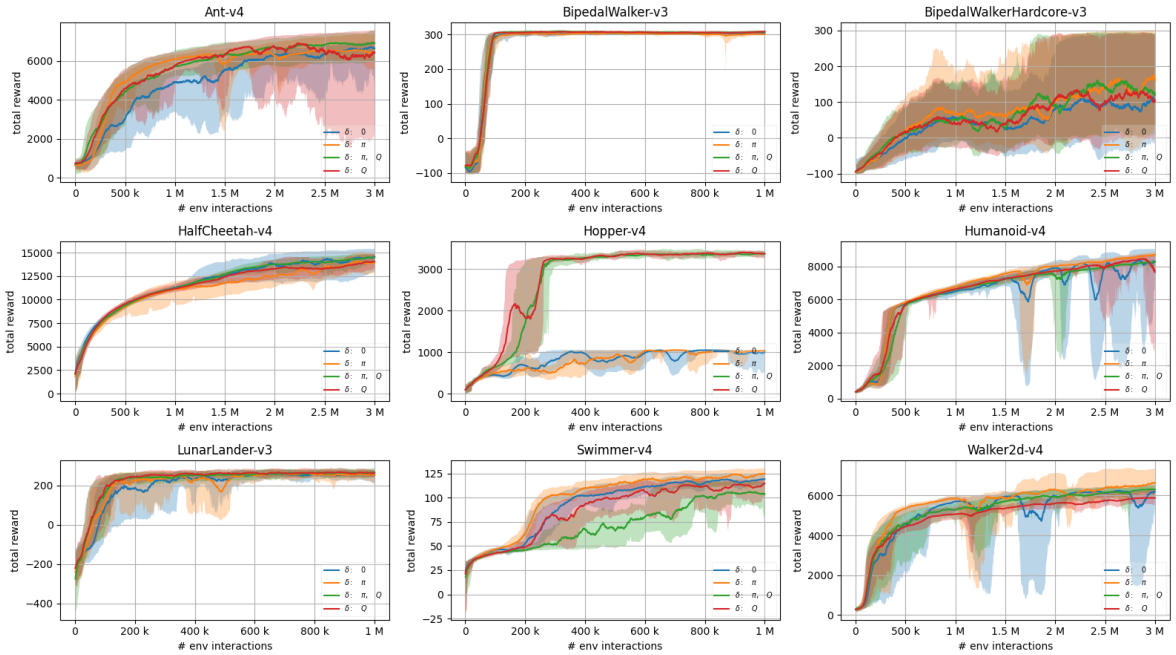


图 7：在固定的悲观程度下，采用 4 种不同丢弃配置的 STAC 的回合得分曲线，该悲观程度可实现最佳性能（见表 4）。

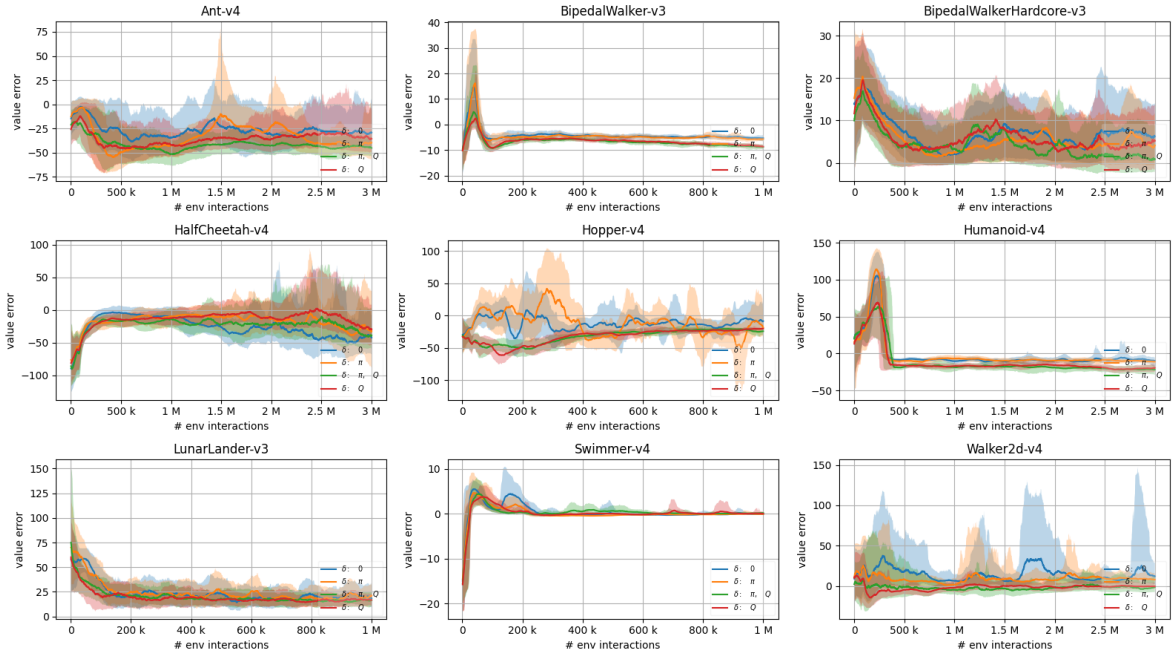


图 8：在固定的悲观程度下，采用 4 种不同丢弃配置的 STAC 的平均回合价值估计误差曲线，该悲观程度可实现最佳性能（见表 4）。

附录 C 超参数与实验细节

表 3：各算法的实验参数

算法	参数	值
DSAC、ESTAC、SAC、STAC	优化器	Adam（[Kingma 和 Ba, 2014]）
	评论家学习率	3×10^{-4}
	演员学习率	3×10^{-4}
	折扣率（γ）	0.99
	目标平滑系数（ρ）	0.995
	目标熵（ $\bar{H}$）	$-\lvert\mathcal{A}\rvert$
	重放缓冲区大小	1×10^{6}
	小批量大小	256
	学习起始步数	10000
	更新到数据比率（G）	1
DSAC	分位数数量	25
	Huber 损失阈值（κ）	1.0

表 4：各环境下取得最佳结果的目标策略熵（$\bar{H}$）、悲观主义 β 和丢弃率。所有算法使用相同的策略熵。

环境	熵（ $\bar{H}$）	悲观主义（β）	演员丢弃	评论家丢弃
Ant-v4	-4	0.25	0.01	0.01
BipedalWalker-v3	-2	0.375	0	0.01
BipedalWalkerHardcore-v3	-2	0.125	0.01	0
HalfCheetah-v4	-3	0.0	0	0
Hopper-v4	-1	0.5	0.01	0.01
Humanoid-v4	-8	0.25	0.01	0
LunarLander-v3	-2	0.0	0.01	0.01
Swimmer-v4	-1	0.0	0.01	0
Walker2d-v4	-3	0.125	0.01	0

附录 D STAC 的网络架构

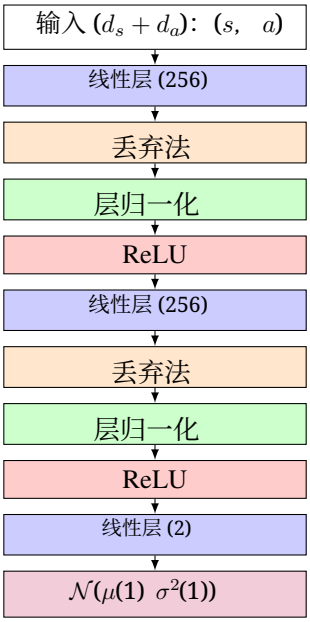


图 9：评论家网络架构

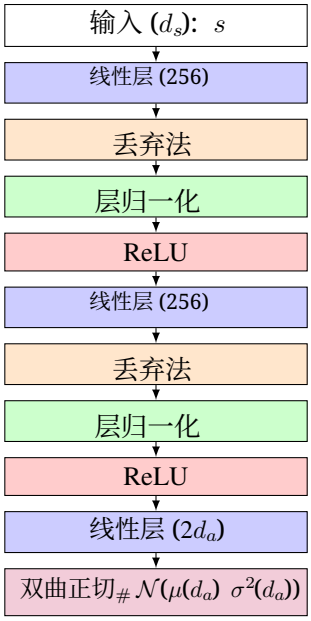


图 10：演员网络架构

附录 E 源代码

我们的结果可在 https://github.com/ugurcanozalp/stochastic-actor-critic 公开获取。这段代码将我们内部开发的强化学习框架作为子仓库使用，该框架可通过 onhttps://github.com/ ugurcanozalp/rl-warehouse 获取。