《机器学习研究杂志》25 (2024) 1-57 投稿：10/23；修订：2/24；发表：2/24

存在对称性和状态抽象时的策略梯度方法

Prakash Panangaden* PRAKASH@CS.MCGILL.CA
麦吉尔大学计算机科学学院及米拉魁北克人工智能研究所，加拿大魁北克省蒙特利尔

Sahand Rezaei-Shoshtari* SREZAEI@CIM.MCGILL.CA
麦吉尔大学计算机科学学院及 *Mila* – 魁北克人工智能研究所，加拿大魁北克省蒙特利尔

Rosie Zhao* ROSIEZHAO@G.HARVARD.EDU
哈佛大学工程与应用科学学院，美国马萨诸塞州剑桥

David Meger DMEGER@CIM.MCGILL.CA
麦吉尔大学计算机科学学院及 *Mila* – 魁北克人工智能研究所，加拿大魁北克省蒙特利尔

Doina Precup DPRECUP@CS.MCGILL.CA
加拿大魁北克省蒙特利尔，麦吉尔大学计算机科学学院、*Mila*–魁北克人工智能研究所和 *DeepMind*

编辑： Martha White

arXiv:2305.05666v2 [cs.LG] 7 Mar 2024

摘要

高维复杂问题中的强化学习（RL）依赖抽象来提升效率和泛化能力。本文研究连续控制设定中的抽象，并将马尔可夫决策过程（MDP）同态的定义扩展至连续状态与动作空间。我们针对随机策略和确定性策略，推导了抽象MDP上的策略梯度定理。我们的策略梯度结果能够利用环境的近似对称性进行策略优化。基于这些定理，我们提出了一系列行动者-评论家算法，能够利用松弛双模拟度量，同时学习策略和MDP同态映射。最后，我们引入了一系列具有连续对称性的环境，以进一步展示我们的算法在存在此类对称性时进行动作抽象的能力。我们在这些环境以及来自 DeepMind 控制套件、颇具挑战性的视觉控制任务上验证了该方法的有效性。我们的方法能够利用MDP同态进行表征学习，从而带来

*。贡献相同；按字母顺序排列。

©2024 Prakash Panangaden、Sahand Rezaei-Shoshtari、Rosie Zhao、David Meger、Doina Precup。许可协议：CC-BY 4.0，详见 https://creativecommons.org/licenses/by/4.0/。署名要求见 http://jmlr.org/papers/v25/23-1415.html。

性能提升，而潜在空间的可视化清晰地展现了所学习抽象的结构。

关键词：强化学习、策略优化、抽象、对称性、表征学习

1 引言

在高维观测下进行强化学习，依赖表征学习和抽象来学习更简单、能够高效求解的问题（Li 等人，2006年；Abel 等人，2016）。然而，一个主要障碍是状态、动作与奖励之间的耦合，在复杂的连续控制问题中尤为如此。人们已提出多种策略，通过捕捉各个状态之间的行为等价性，寻找缩减状态空间的方法。针对马尔可夫决策过程的一种形式化方法是双模拟（Givan 等人，2003），它最初于20世纪80年代初被引入标记转移系统（Milner，1989b）。双模拟在状态空间上定义了一种等价关系，使我们能够根据该关系下的等价类对状态空间进行商化。双模拟及其相关的双模拟度量（Ferns 等人，2004）——后者用于近似这种等价关系——此前已用于抽象和模型最小化。

或者，也可以利用商状态空间定义一个新的环境，其转移动力学和奖励保留原始状态空间的结构，并定义原始马尔可夫决策过程与新马尔可夫决策过程之间的函数。因此，与双模拟密切相关的是马尔可夫决策过程同态（Ravindran，2004；Ravindran 和 Barto，2001、2004），它们通过在具有特定保持性质的马尔可夫决策过程之间建立映射来刻画行为等价。与双模拟类似，可以利用马尔可夫决策过程同态来利用马尔可夫决策过程的（近似）对称性，实现联合状态-动作抽象。

马尔可夫决策过程同态是在离散状态和动作空间的背景下提出的，是马尔可夫决策过程之间保持结构并保留价值函数的映射。通常，它们用于将一个马尔可夫决策过程映射到一个抽象MDP，同时确保不会丢失相关信息。Ravindran 和 Barto（2001）证明，可以将策略从抽象MDP映射回原始 MDP，即提升策略，同时保持最优性。以这种方式映射回策略是一项复杂的构造，而且明确依赖于状态与动作空间的有限性。从实践角度来看，近期研究表明，利用马尔可夫决策过程同态能够有效地指导离散问题中的学习（van der Pol 等人，2020a、b；Biza 和 Platt，2019年）。图1展示了马尔可夫决策过程同态的示意图及其关键性质，我们将在第3节中正式定义它们。

我们的首项贡献是将 MDP 同态扩展到连续情形。如果要将这些思想用于物理空间中的动力系统控制（例如机器人学），这一步至关重要。相关数学理论比有限情形深奥得多，而且在某些情况下，有限情形无法为如何推进提供任何指导。我们证明，与有限情形一样，随机策略和确定性策略的价值函数以及最优价值函数都得以保留。将策略从抽象空间提升到原始 MDP，是一个关键例子；在这一过程中，我们必须采取与 Ravindran 和 Barto（2001）截然不同的方法，即借助泛函分析中的经典工具。

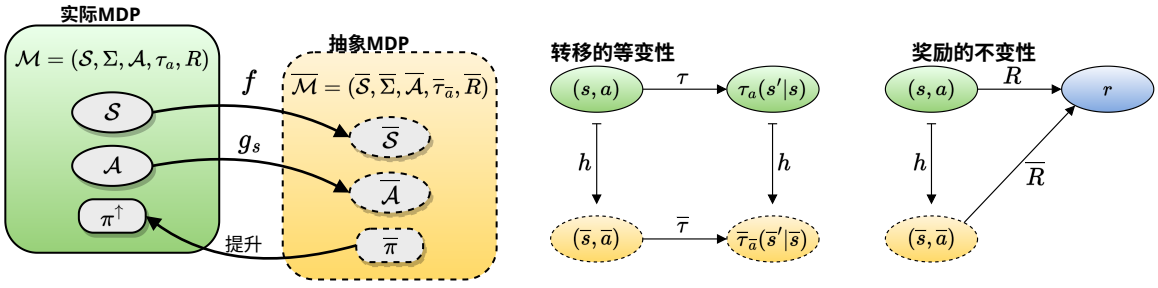
下一个重要贡献是，我们推导出了策略梯度定理（Sutton 等人，2000年；Silver 等人，2014 年）的一个版本，将 MDP同态紧密地整合到策略优化过程中。换言之，可以使用从抽象MDP中获得的策略梯度（称为同态策略梯度 （HPG））来优化定义在原始马尔可夫决策过程上的性能度量。我们严格证明了这一结果对确定性策略和随机策略均成立，并表明，HPG 可以作为额外的梯度估计器，利用近似对称性来提高样本效率。由于策略梯度方法仍然是强化学习算法中的重要一类，尤其适用于连续控制问题（Kiran 等人，2021年；Arulkumaran 等人，2017年），因此，在存在状态抽象的情况下，我们推导的同态策略梯度可能会对策略梯度算法产生重要影响。

我们的第三项贡献是提出了一种深度演员-评论家算法，称为深度同态策略梯度 （DHPG）算法，该算法基于我们新颖的理论结果。DHPG 能够同时优化策略、学习同态映射，并利用 MDP 同态的抽象进行策略优化。我们通过实验证明，通过 MDP 同态学习到的状态-动作抽象为具有挑战性的视觉控制问题中的表征学习提供了自然的归纳偏置，从而在性能和样本效率方面优于强大的基线方法。

最后，我们展示了当对称群同时具有连续性时，如何对一个 MDP 进行折叠。例如，如果一个系统具有球对称性，那么该系统在旋转群 $SO(3)$ 的动作下保持不变，而这显然不是一个有限群。离散对称性确实可能出现在连续系统中，但通常情况下，我们处理的是连续对称性。此外，为了展示 DHPG 学习连续对称性的能力，我们开发了一系列具有连续对称性的环境。总而言之，我们的贡献可以概括如下：

1. 定义连续状态与动作空间上的连续 MDP 同态，并证明在随机策略的一般情形下提升策略的存在性。2. 证明 MDP 同态的价值等价性和最优价值等价性，适用于

随机策略的一般情形。



(a) MDP 同态的组成部分。(b) MDP 同态的交换图。

图1：MDP 同态概述 $h = (f, g_s)$。(a) MDP 同态映射的组成部分，以及实际MDP与抽象MDP之间的关系。(b) 用于展示转移的等变性和奖励不变性的MDP同态交换图。该图改编自 Ravindran 和 Barto（2001）。

3. 推导适用于随机策略和确定性策略的同态策略梯度定理。4. 开发一系列深度演员-评论家算法，使其能够在学习最优策略的同时学习马尔可夫决策过程同态映射。借助一种新颖且计算高效的策略提升过程，我们的算法适用于随机策略和确定性策略。5. 开发一系列具有连续对称性的新型强化学习环境，以便基准测试智能体学习并利用环境连续对称性的能力。

值得注意的是，与 Rezaei-Shoshtari 等人（2022）的先前工作相比，我们的理论和实证贡献并不局限于确定性策略和双射动作编码器。相反，我们证明了在随机策略和满射动作编码器的一般情况下，价值等价性质和同态策略梯度定理均成立，并提出了一种计算效率较高的方法，用于提升一般随机策略。在实证方面，我们表明，在具有连续对称性的环境中，随机 DHPG 优于确定性DHPG，因为它能够实现更强大的动作抽象。我们公开提供了 DHPG 的代码以及具有连续对称性的新型环境[1]。

本文结构如下：第 2 节和第 3 节概述相关工作并介绍相关背景，包括有限马尔可夫决策过程同态、双模拟和策略梯度方法。第 4 节正式介绍连续MDP和连续 MDP 同态，并证明关键的等价性质。第 5 节证明随机和确定性同态策略梯度定理，随后在第 6 节介绍 DHPG 算法。最后，第 7 节给出 DHPG 在连续控制任务上的实验结果。

2 相关工作

状态抽象。 抽象可以定义为一种将原始表征映射为更紧致且更易于处理的抽象表征的过程（Li 等人，2006）。概率双模拟（下文简称“双模拟”）（Larsen 和 Skou，1991b）是系统之间行为等价性的一种概念。Blute 等人（1997b）和 Desharnais 等人（2002）将其扩展到连续状态空间，而 Givan 等人（2003）则将其扩展到马尔可夫决策过程。双模拟度量（Desharnais 等人，1999；Ferns 等人，2005b、2006、2011）定义了一种伪度量，用于量化行为相似性的程度。近期，Zhang 等人（2020）定义了一种损失函数，通过潜在状态的双模拟等价性来学习表征，而 Kemertas 和 Aumentado-Armstrong（2021年）进一步提高了其鲁棒性。Castro（2020）提出了一种方法，用于近似具有连续状态和离散动作的确定性马尔可夫决策过程的双模拟度量。van der Pol 等人（2020a）基于马尔可夫决策过程同态定义了一种对比损失，用于学习适合规划的抽象MDP，但其方法仅适用于有限马尔可夫决策过程。另一种方法是直接将马尔可夫决策过程同态关系嵌入网络架构中（van der Pol 等人，2020b、2021）。近期提出的其他度量（Le Lan 等人，2021）旨在学习能够保持价值（Grimm 等人，2020、2021）或策略（Agarwal 等人，2020）的表征，或者通过基于采样的相似度度量（Castro 等人，】【。

1. https://github.com/sahandrez/同态策略梯度

2021年）。近期，Kemertas 和 Jepson（2022年）将双模拟关系引入了近似策略迭代中。最后，状态抽象原则上有助于改善策略的迁移（Abel 等人，2019年；Castro 和 Precup，2010年；Soni 和 Singh，2006年；Sorg 和 Singh，2009年；Rajendran 和 Huber，2009年），或学习时间扩展动作（Castro 和 Precup，2011年；Wolfe 和 Barto，2006a,b；Sutton 等人，1999年）。

动作抽象。 动作表示常在大型离散动作空间的背景下进行研究（Sallans 和 Hinton，2004），作为一种预先已知的查找嵌入（Dulac-Arnold 等人，2015）、因子化表示（Sharma 等人，2017）或策略分解（Chandak 等人，2019）。动作表示也可以通过专家演示学习（Tennenholtz 和 Mannor，2019）。与我们工作更相关的是动力学感知嵌入（Whitney 等人，2019），其中学习的是用于连续控制的状态-动作联合嵌入。相比之下，我们采用同态的概念来学习依赖于状态的动作表示，同时保留价值。最后，动作表示可以与时间抽象（Sutton 等人，1999）结合，以发现时间扩展动作（Ravindran 和 Barto，2003；Abel 等人，2020；Castro 和 Precup，2010，2011）。

状态表示学习。现有从原始观测中学习底层状态空间的方法，通常会使用潜在模型（Gelada 等人，2019；Hafner 等人，2019a,b；Ha 和 Schmidhuber，2018；Biza 等人，2021）、辅助预测任务（Jader- berg 等人，2016；Liu 等人，2019；Lyle 等人，2021）、受物理学启发的归纳偏置（Jon- schkowski 和 Brock，2015；Cranmer 等人，2020；Greydanus 等人，2019）、无监督学习（Hjelm 等人，2018；Liu 和 Abbeel，2021）或自监督学习（Anand 等人，2019；Sinha 等人，2021；Hansen 等人，2020；Hansen 和 Wang，2021；Fan 等人，2021）。从另一个角度来看，表征学习可以有效地与强化学习问题解耦（Eslami 等人，2018；Stooke 等人，2021）。环境的对称性也可用于表征学习（Mondal 等人，2022；Mahajan 和 Tulaband- hula，2017；Park 等人，2022；Wang 等人，2021；Higgins 等人，2018，2021；Quessard 等人，2020；Caselles-Dupr´e 等人，2019）。事实上，马尔可夫决策过程同态是此类方法在强化学习中的特化形式。马尔可夫决策过程同态的一个关键区别在于，它们能够像 Thomas 等人（2017）那样，在相同前提下将动作纳入表征学习。近期，简单的图像增强方法已显著提升了强化学习性能（Yarats 等人，2020；Lee 等人，2019）。由于这些方法通常与状态抽象正交，因此可以结合使用。

等变表示学习。 利用等变性来发挥数据中的对称性，是机器学习研究中一条富有成果的路线：在模型架构中施加等变性属性，已使模型在多种数据模态和应用中取得了最先进的性能。这些领域包括计算机视觉中的分割与分类任务（Cohen 和 Welling，2016）、医学影像（Winkels 和 Cohen，2019；Veeling 等人，2018年）、三维模型分类（Thomas 等人，2018；Chen 等人，2021年）、量子化学（Qiao 等人，2021年；Satorras 等人，2021年；Batzner 等人，2022年），以及蛋白质结构分类（Eismann 等人，2021年；Ganea 等人，2021年；Jumper 等人，2021年）。自从平移等变性在传统卷积神经网络中的效用得到证明以来（LeCun 等人，1989，1995），近年来，这些卷积层已被推广为对离散群——如有限旋转、平移和反射（Cohen

和 Welling，2016）——以及连续群——如旋转群 SO(3)、欧几里得群 E(3) 和特殊欧几里得群 SE(3)（Cohen 等人，2018；Kondor 等人，2018；Weiler 等人，2018；Cohen 等人，2019）等变。施加于这些架构的等变性约束非常严格，先前研究表明，实现真正的等变性并不容易（Azulay 和 Weiss，2018；Engstrom 等人，2017）。此外，这些等变网络所对应的群通常是预先固定且已知的；不过，也有研究为任意矩阵李群构建适当的等变网络（Finzi 等人，2021年），并提出了自动发现与李群相关对称性的算法（Dehmamy 等人，2021年）。

3 背景

3.1 马尔可夫决策过程

强化学习基于智能体与环境交互并在此过程中获得奖励。它旨在最大化期望奖励，并通过与环境的交互学习如何实现这一目标。马尔可夫决策过程是形式化描述智能体与环境之间交互的基本模型。

定义 1（马尔可夫决策过程）一个马尔可夫决策过程（马尔可夫决策过程）是一个元组$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \tau_a, R, \gamma)$，其中$\mathcal{S}$ 是状态集合，$\mathcal{A}$ 是动作集合，对于每个 $a \in \mathcal{A}$，我们有 $\tau_a : \mathcal{S} \to \Delta(\mathcal{S})$，其中 $\Delta(\mathcal{S})$ 表示定义在$\mathcal{S}$上的概率分布集合，$R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 是奖励函数，而 $\gamma \in [0, 1)$ 是折扣因子。

最初，我们假设 S 和 A 都是有限的；在第 4 节中，我们将定义具有更一般状态与动作空间的马尔可夫决策过程。从状态 $s \in S$ 出发，遵循策略 $\pi : S \to \Delta(A)$ 行动的智能体会选择动作 $a \sim \pi(\cdot|s)$，并转移到 $s' \sim \tau_a(\cdot|s)$，从而获得奖励 $r = R(s, a)$。其目标是通过学习最优策略来最大化期望回报：

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\tau}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)].$$

这里需要注意，我们假设 $\gamma < 1$ 以确保回报收敛（尽管对于回合制任务，允许使用 $\gamma = 1$）。

价值函数 $V^{\pi}(s)$ 表示从状态 s 开始并遵循策略 π 时的期望回报。动作价值函数 $Q^{\pi}(s, a)$ 表示从状态s开始，采取动作 a 并在此后遵循 π 时的期望回报。价值函数是贝尔曼算子 $T^{\pi} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ 的不动点，其定义如下：

$$T^{\pi}V(s) := \mathbb{E}_{\substack{a\sim\pi(\cdot|s) \\ s'\sim\tau_a(\cdot|s)}} [r + \gamma V(s')].$$

同样，最优价值函数 V^* 是不动点贝尔曼最优算子$T^* : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$的：

$$T^*V(s) := \max_a \left[\mathbb{E}_{s'\sim\tau_a(\cdot|s)}[r + \gamma V(s')]\right].$$

类似地，也可为 Q^{π} 和 Q^* 定义贝尔曼方程（Sutton 和 Barto，2018年）。

3.2 策略梯度定理

强化学习算法大致可分为基于价值的方法和策略梯度（PG）方法。基于价值的方法通过贪心最大化步骤，根据学习到的动作价值选择动作；而策略梯度方法则直接根据性能度量 $J(\theta)$的梯度，优化参数化策略 π_θ ，其定义如下：

$$J(\theta) = \mathbb{E}_\pi[V^\pi(s)], \tag{1}$$

其中，期望是相对于实际MDP的策略、状态转移和初始状态分布计算的。与基于价值的方法不同，策略梯度算法继承了梯度下降所具有的强大（尽管是局部的）收敛保证，并且能够自然地扩展到连续动作。策略梯度方法所依据的基本定理是策略梯度定理（Sutton 等人，2000年）：

定理 2（Sutton 等人（2000年））设 $\pi_\theta : \mathcal{S} \to \Delta(\mathcal{A})$ 是定义在$\mathcal{M}$上的随机策略。那么，性能度量 $J(\theta)$ 关于 θ 的梯度为：

$$\nabla_\theta J(\pi_\theta) = \int_{s\in\mathcal{S}} \rho^{\pi_\theta}(s) \int_{a\in\mathcal{A}} \nabla_\theta \pi_\theta(da|s) Q^{\pi_\theta}(s,a) ds,$$

其中， $\rho^{\pi_\theta}(s) = \lim_{t\to\infty} \gamma^t P(s_t = s|s_0, a_{0:t} \sim \pi_\theta)$ 是 π_θ下状态的折扣平稳分布。

在定理 2 中，假设 $\rho^{\pi_\theta}(s)$ 存在，且与初始状态分布无关（遍历性假设）。策略梯度定理的重要意义在于，策略变化对状态分布的影响不会出现在其表达式中，从而能够通过基于样本的估计来计算梯度（Williams，1992）。Silver 等人（2014）针对确定性策略推导出了确定性策略梯度（DPG）：

定理 3（Silver 等人（2014 年））设 $\pi_\theta : \mathcal{S} \to \mathcal{A}$ 是定义在$\mathcal{M}$上的确定性策略。那么，性能度量 $J(\theta)$ 关于 θ 的梯度为：

$$\nabla_\theta J(\pi_\theta) = \int_{s\in\mathcal{S}} \rho^{\pi_\theta}(s) \nabla_\theta \pi_\theta(s) \nabla_a Q^{\pi_\theta}(s,a)\big|_{a=\pi_\theta(s)} ds,$$

其中， $\rho^{\pi_\theta}(s) = \lim_{t\to\infty} \gamma^t P(s_t = s|s_0, a_{0:t} \sim \pi_\theta)$ 是策略 π_θ下状态的折扣平稳分布。

由于 DPG 无需对动作空间进行积分，因此通常比随机策略梯度（Silver 等人，2014 年）更具样本效率。然而，探索过程中需要手动注入噪声，因为确定性策略本身不具备探索能力。最后，值得注意的是，由于价值函数需要对 a求微分，DPG 仅适用于连续动作。

3.3 双模拟与双模拟度量

双模拟是状态空间上的一种基本等价关系，体现了行为相似性的概念。它由 Milner（1980、1989a）和 Park（1981）于20世纪70年代末和20世纪80年代初在非概率语境下提出，随后被扩展至

Larsen 和 Skou（1991a）研究了概率系统。Blute 等人（1997a）和 Desharnais 等人（2002）将其扩展到了连续状态空间。这些模型并未涉及奖励，但只需进行少量修改即可加入奖励，正如 Givan 等人（2003）所做的那样。MDP 上的双模拟关系正式定义如下：

定义 4（双模拟）在 *MDP* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \tau_a, R, \gamma)$上的双模拟关系是$\mathcal{S}$ 上的一个等价关系 B，满足：若 sBt 成立，则对于任意动作 a 以及 B 的任意等价类 C，都有：

- $R(s, a) = R(t, a)$ 以及
- $\tau_a(C|s) = \tau_a(C|t)$。

如果两个状态 s 和 t 之间存在这样的关系，我们就称 s 和 t 双模拟等价，并记作 $s \sim t$。

可以将双模拟定义为 S 上等价关系所构成的完备格中某个适当算子的不动点的最大值（Milner，1989b）。双模拟无法应对系统参数的微小扰动。在马尔可夫决策过程这类定量场景中，我们需要使用度量来对相似性进行定量刻画。

为了定义一种可视为双模拟定量版本的“度量”，自然会想到一种伪度量，其核为双模拟等价关系。我们通过不动点构造来定义这种伪度量。我们在 $\mathfrak{M}$（S 上 1-有界伪度量的空间）上赋予如下度量：

$$\Delta(m_1, m_2) := \sup_{x,y \in \mathcal{S}} |m_1(x, y) - m_2(x, y)|.$$

这里，m_1，m_2 是 $\mathfrak{M}$ 的元素，即 1-有界伪度量。然后，我们定义一个称为 $T_K : \mathfrak{M} \to \mathfrak{M}$ 的算子，如下所示：

$$T_K(m)(x, y) = \max_{a \in \mathcal{A}} [|R(x, a) - R(y, a)| + \gamma W_1(m)(\tau_a(x), \tau_a(y))].$$

这里，$\tau_a(x)$表示系统从 x 出发执行 a-转移时在状态空间上的概率分布，$\tau_a(y)$也类似。概率分布上的度量 W_1 是众所周知的Kantorovich度量[2]，它取决于 m。不难证明，空间 $\mathfrak{M}$ 配备 Δ 后是一个完备度量空间，并且函数或算子 T_K 关于度量 Δ是压缩的。因此，根据巴拿赫不动点定理，它有唯一的不动点。这就是双模拟度量[3]。

3.4 有限马尔可夫决策过程同态

与马尔可夫决策过程中状态的行为等价这一概念密切相关的是模型最小化方法。这些方法旨在识别原始马尔可夫决策过程中的可约化部分，从而得到一个等价且规模更小的马尔可夫决策过程。这催生了 MDP同态这一概念，该概念最初由 Ravindran 和 Barto（2001）提出。我们将在状态与动作空间有限的假设下，介绍 MDP同态的定义及相关结果。

2. 更常被称为“Wasserstein”度量，但这种称呼在历史上并无依据。3. 在 Ferns 等人（2005a）中，使用了一种基于格理论的不同不动点定理。

定义 5（MDP 同态）在马尔可夫决策过程 $M=(S, A\ \tau_a, R, \gamma)$ 与 $\overline{M}=(\overline{S}, \overline{A}\ \overline{\tau}_a, \overline{R}, \overline{\gamma})$ 之间的MDP同态 h 是由满射 $h=(f, g_s)$ 构成的元组，其中 $f: S \to \overline{S}$，且对于每个 $s \in S$， $g_s: A \to \overline{A}$，满足：

1. $R(s,a) = \overline{R}(f(s)\ g_s(a))$ 对每个 $s \in S$ $a \in A$均成立；2. 对每个 $s, s' \in S$ 及 $a \in A$，

$$\overline{\tau}_{g_s(a)}(f(s')|f(s)) = \sum_{s'' \in [s']_{B_h|S}} \tau_a(s''|s), \tag{2}$$

其中 B_h 是由 h, $B_h \mid S$ 的等价关系所诱导的 $S \times A$ 的划分， B_h 到 S的投影为 B_h，而 $[s']_{B_h|S}$ 是包含 s'的 $B_h|S$ 的划分。

换言之，在像 MDP $\overline{M}$ 中，采取动作 $g_s(a)$ 时，从 $f(s)$ 转移到 $f(s')$ 的概率，等于在原始马尔可夫决策过程 M 中采取动作 a 时，从 s 转移到子集 $[s']_{B_h|S}$ 的概率。图1b展示了马尔可夫决策过程同态的交换图。MDP同态的一个关键性质是最优价值等价性，这表明最优价值函数在此映射下保持不变。

定理6（Ravindran 和 Barto（2001））设 h 是从$M=(S, A\ \tau_a, R, \gamma)$ 到$\overline{M}=(\overline{S}, \overline{A}\ \overline{\tau}_a, \overline{R}, \overline{\gamma})$的 *MDP*同态。那么，对于任意$(s, a) \in S \times A$，我们有：

$$Q^*(s,a) = \overline{Q}^*(f(s), g_s(a)).$$

马尔可夫决策过程及其在 MDP同态下的像 MDP 的最优策略也密切相关。给定像 MDP 上的一个策略，我们可以在原始马尔可夫决策过程上定义一个新的提升策略，使其具有“等价行为”。

定义*7*（提升策略）设 h 是从 $M=(S, A, \tau_a, R, \gamma)$ 到 $\overline{M}=(\overline{S}, \overline{A}\ \overline{\tau}_a, \overline{R}, \overline{\gamma})$ 的 *MDP*同态，并设 $\overline{\pi}: \overline{S} \to \Delta(\overline{A})$ 是 $\overline{M}$ 上的一个策略。那么，将 $\overline{\pi}$ 提升到 M 后得到的是一个策略 $\pi^{\uparrow}: S \to \Delta(A)$，满足对于任意 $(s, a) \in S \times A$，我们有：

$$\pi^{\uparrow}(a|s) = \frac{\overline{\pi}(g_s(a)|f(s))}{|g_s^{-1}(\{g_s(a)\})|}.$$

请注意，要使这些结果成立，只需提升策略满足：

$$\sum_{a \in g_s^{-1}(\{g_s(a)\})} \pi^{\uparrow}(a|s) = \overline{\pi}(g_s(a)|f(s)) \quad \forall\ s \in S \tag{3}$$

但为了使提升策略唯一，Ravindran 和 Barto（2001）选择将从 $f(s)$ 执行 $g_s(a)$ 的概率，均匀分配到所有满足 $g_s(a) = g_s(a')$ 的动作 a' 上。我们有以下结果：M 的最优策略所对应的提升策略是 M 的最优策略：

定理 8（Ravindran 和 Barto（2001））设 h 是从$M=(S、A\ \tau_a、R、\gamma)$ 到$\overline{M}=(\overline{S}、\overline{A}\ \overline{\tau}_a、\overline{R}、\overline{\gamma})$的 *MDP*同态映射，并设 $\overline{\pi}^*: \overline{S} \to \overline{\Delta(A)}$ 是 $\overline{M}$上的最优策略。那么，提升策略 $\pi^*: S \to \Delta(A)$ 是 M的最优策略。

此外，Rezaei-Shoshtari 等人（2022年）表明，根据提升策略的这一定义，我们有一个价值等价 结果，说明所有价值函数——而不仅仅是最优价值函数——都在 MDP 同态映射下得以保留。

定理9（Rezaei-Shoshtari 等人（2022）） 设 h 是从$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \tau_a, R, \gamma)$到$\overline{\mathcal{M}} = (\overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\tau}_a, \overline{R}, \gamma)$的一个*MDP*同态。对于任意$(s, a) \in \mathcal{S} \times \mathcal{A}$、抽象策略$\pi : \mathcal{S} \to \Delta(\mathcal{A})$及其提升策略 $\pi^\uparrow : \mathcal{S} \to \Delta(\mathcal{A})$，均有：

$$Q^{\pi^\uparrow}(s, a) = \overline{Q}^{\overline{\pi}}(f(s), g_s(a)).$$

4 连续 MDP 同态

上一节中，我们介绍的 MDP 同态仅适用于状态与动作空间有限的情形。本节将对一般连续域中的 MDP 同态进行形式化定义。首先，我们定义连续 MDP，并陈述基本假设；这些假设要求我们谨慎处理空间的可测性与可微性。

定义 10（连续马尔可夫决策过程） 一个连续马尔可夫决策过程（马尔可夫决策过程）是一个6元组：

$$\mathcal{M} = (\mathcal{S}, \Sigma, \mathcal{A}, \forall a \in \mathcal{A}\ \tau_a : \mathcal{S} \times \Sigma \to [0, 1], R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}, \gamma)$$

其中，$\mathcal{S}$ 是状态空间，假设它是一个适当的拓扑空间；Σ 是$\mathcal{S}$上的一个 σ-代数；[4]，$\mathcal{A}$是动作空间，它是一个局部紧度量空间，通常取为 $\mathbb{R}^n$ 的子集；τ_a 是针对每个可能动作的转移概率核 a；对于每个固定的s，$\tau_a(\cdot|s)$ 是 Σ 上的一个概率分布；R 是奖励函数，γ 是折扣因子。此外，对于所有 $s \in \mathcal{S}$ 和 $B \in \Sigma$，映射 $a \mapsto \tau_a(B|s)$ 是光滑的。

最后一个假设是保证关于动作 a 可微所必需的，这也是第 5 节推导同态策略梯度定理所需的。

当假设底层空间是波兰空间时，连续空间上的概率论具有良好的性质（定义见附录 B），但我们的结果并不需要波兰空间的性质。关于动作空间的假设是证明策略可以被提升所必需的；也许可以通过其他假设证明这一点，但局部紧度量空间足够一般，能够涵盖我们见过的所有例子。

接下来，我们将定义连续 MDP 同态，并分别针对最优价值等价和价值等价建立相关结论。

定义 11（连续 MDP 同态） *A* 连续 MDP 同态是一个映射 $h = (f, g_s)$ ：马尔可夫决策过程 $\to$ 马尔可夫决策过程，其中 $f : \mathcal{S} \to \overline{\mathcal{S}}$，并且对于每个 $s \in \mathcal{S}$，$g_s : \mathcal{A} \to \mathcal{A}$ 都是可测的满射映射，且满足以下条件：

$$\textit{Invariance of reward: } \overline{R}(f(s), g_s(a)) = R(s, a) \qquad \forall s \in \mathcal{S}, a \in \mathcal{A} \tag{4}$$

$$\textit{Equivariance of transitions: } \overline{\tau}_{g_s(a)}(\overline{B}|f(s)) = \tau_a(f^{-1}(\overline{B})|s) \qquad \forall\ s \in \mathcal{S}, a \in \mathcal{A}, \overline{B} \in \overline{\Sigma} \tag{5}$$

4. 通常是博雷尔代数。

注意，如果 g_s 是恒等映射，则第二个条件简化为 $\tau_a(B|f(s)) = \tau_a(f^{-1}(B)|s)$，这仅仅是双模拟中所使用的保持转移概率的条件（Desharnais 等人，2002）。

奖励上的条件可直接从有限情形平移而来。转移的等变性使用像 MDP 上定义的 σ-代数来定义；它表明，测度 $\tau_{g_s(a)}(\cdot|f(s))$ 是测度 $\tau_a(\cdot|s)$ 在状态映射 f 下的推前测度。在接下来的结果中，我们将看到要求这一条件的原因。

4.1 最优价值等价性

在这种连续情形下，我们将证明最优价值等价性仍然成立。证明与定理6类似，不过，我们利用变量替换公式（见附录 B 中的定理 25）来改变连续 Bellman 方程中的积分域，而不是对求和重新编排索引。

定理 12（最优价值等价性） 设 $\mathcal{M} = (\mathcal{S}, \Sigma, \mathcal{A}, \tau_a, R, \gamma)$ 是连续 MDP 同态 $h = (f, g_s)$ 的像，该同态从 $\mathcal{M} = (\mathcal{S}, \Sigma, \mathcal{A}, \tau_a, R, \gamma)$ 出发。那么对于任意 $(s,a) \in \mathcal{S} \times \mathcal{A}$，我们有：

$$Q^*(s,a) = \overline{Q}^*(f(s), g_s(a)),$$

其中，Q^* 与 $\overline{Q}^*$ 分别是 $\mathcal{M}$ 和 $\overline{\mathcal{M}}$ 的最优动作价值函数。 —

证明 我们先证明以下断言：

断言 13 对于 $m \geq 1$，定义序列 $Q_m : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ 如下：

$$Q_m(s,a) = R(s,a) + \gamma \int_{s' \in \mathcal{S}} \tau_a(ds'|s) \sup_{a' \in \mathcal{A}} Q_{m-1}(s', a')$$

以及 $Q_0(s,a) = R(s,a)$。类似地，定义序列 $\overline{Q}_m : \overline{\mathcal{S}} \times \overline{\mathcal{A}} \to \mathbb{R}$。那么，对于任意 $(s,a) \in \mathcal{S} \times \mathcal{A}$，我们断言：

$$Q_m(s,a) = \overline{Q}_m(f(s), g_s(a)).$$

我们将通过对 m 进行归纳法来证明该断言。基例 $m = 0$ 由连续 MDP 同态的奖励不变性推出：

$$Q_0(s,a) = R(s,a) = \overline{R}(f(s), g_s(a)) = \overline{Q}_0(f(s), g_s(a)).$$

对于归纳情形，请注意：

$$\begin{aligned}
Q_m(s,a) &= R(s,a) + \gamma \int_{s' \in \mathcal{S}} \tau_a(ds'|s) \sup_{a' \in \mathcal{A}} Q_{m-1}(s', a') && (6)\\
&= \overline{R}(f(s), g_s(a)) + \gamma \int_{s' \in \mathcal{S}} \tau_a(ds'|s) \sup_{a' \in \mathcal{A}} \overline{Q}_{m-1}(f(s'), g_{s'}(a')) && (7)\\
&= \overline{R}(f(s), g_s(a)) + \gamma \int_{s' \in \mathcal{S}} \tau_a(ds'|s) \sup_{a' \in \overline{\mathcal{A}}} \overline{Q}_{m-1}(f(s'), a') && (8)\\
&= \overline{R}(f(s), g_s(a)) + \gamma \int_{s' \in \overline{\mathcal{S}}} \overline{\tau}_{g_s(a)}(ds'|f(s)) \sup_{a' \in \overline{\mathcal{A}}} \overline{Q}_{m-1}(s', a') && (9)\\
&= \overline{Q_{m-1}}(f(s), g_s(a)), && (10)
\end{aligned}$$

其中，公式 7 根据归纳假设得出，公式 8 根据 g_s 是满射得出，公式 9 根据变量替换公式得出；事实上，根据定义 11， f 下 $\tau_a(\cdot|s)$ 的推前测度等于 $\tau_{g_s(a)}(\cdot|f(s))$，而这里我们对一个定义在 S $\to \mathbb{R}$ 上、形式为 $s' \mapsto \sup_{a' \in \mathcal{A}} Q_{m-1}(s', a')$ 的函数进行积分。这就完成了归纳证明。由于 $\lim_{m\to\infty} Q_m(s,a) = Q^*(s,a)$，因此可得$Q^*(s,a) = Q^*(f(s)\ g_s(a))$。

■

4.2 策略提升与价值等价

回顾一下，在有限情形中，我们有一个通过 MDP同态定义提升策略的精确方程。在连续情形中，找到一个普遍存在且同时能够给出价值等价结果的提升策略并非易事。我们将使用以下条件来定义连续 MDP 同态的提升策略。

定义 *14*（策略提升）设 $M = (S\ \Sigma, A\ \tau_a, R, \gamma)$ 是从 $\overline{M} = (\overline{S}\ \overline{\Sigma}, \overline{A}\ \overline{\tau_a}, \overline{R}, \gamma)$ 经连续 *MDP* 同态 $h = (f, g_s)$ 映射得到的像。则对于定义在 M 上的任意策略 $\pi : S \to \Delta(A)$，M 上的策略 $\pi^\uparrow : S \to \Delta(A)$ 是 $\overline{\pi}$ 的提升策略，当且仅当：

$$\pi^\uparrow(g_s^{-1}(\beta)|s) = \overline{\pi}(\beta|f(s)) \tag{11}$$

对于每个博雷尔集 $\beta \subseteq \overline{\mathcal{A}}$和 $s \in S$。换句话说， $\pi(f(s)) \cdot$ 是对所有 $s \in S$ 相对于 g_s 的 $\pi^\uparrow(s, \cdot)$ 的推前测度。

请注意，定义 14 并未定义测度，因为我们需要指定对 $\pi^\uparrow(s,\ B)$ 中所有 Borel 集 B in A 所赋的值，而不仅仅是那些作为原像 $g_s^{-1}(\beta)$ 出现的集合。然而，直接定义：

$$\pi^\uparrow(B|s) = \overline{\pi}(g_s(B)|f(s))$$

由于 g_s 不会将博雷尔集映射为博雷尔集，且一般情况下 $B \subsetneq g_s^{-1}(g_s(B))$，因此会立即产生一些问题。换言之，只有当 g_s 是双射，且将可测集映射为可测集时，我们才能使用这一含义。然而，正如下一个结果所示，假设 A 和 A 是局部紧度量空间，则一般确实存在满足定义 14 中条件的测度。该证明运用了泛函分析中的结果，具体而言，即 Hahn-Banach 定理和 Riesz 表示定理。值得注意的是， g_s 的双射假设是 Rezaei-Shoshtari 等人（2022）先前研究的局限之一，而这一局限已在我们的论文中得到消除。

命题15 设 $\mathcal{M} = (\mathcal{S}\ \Sigma, \overline{\mathcal{A}}\ \tau_a, \overline{R}, \gamma)$ 是连续马尔可夫决策过程同态 $h = (f, g_s)$ 将 $\mathcal{M} = (\mathcal{S}\ \Sigma, \mathcal{A}\ \tau_a, R, \gamma)$映射得到的像，其中 $\mathcal{A}$和 $\mathcal{A}$是局部紧度量空间。那么，对于定义在 $\mathcal{M}$上的任意策略 $\pi : \mathcal{S} \to \Delta(\mathcal{A})$，都存在一个符合定义 *14* 意义的提升策略$\pi^\uparrow : S \to \Delta(\mathcal{A})$。

证明 定义泛函 $p : C_0(\mathrm{A}) \to \mathbb{R}$ 如下：

$$p(\psi) = \max_{a \in \mathcal{A}} \psi(a).$$

显然，对于每个 ψ，$\varphi \in C_0(\mathrm{A})$ 和 $0 < \alpha < \infty$，$p(\varphi + \psi) \le p(\varphi) + p(\psi)$，且 $p(\alpha\psi) = \alpha p(\psi)$。事实上，$p$ 是一个半范数。注意，由于 g_s 是满射，我们可以定义子空间 $U := \{\eta \bigcirc g_s : \eta \in C_0(\mathrm{A})\} \subseteq C_0(\mathrm{A})$。令 ρ 为定义在 U 上的线性泛函，其定义为

$$\rho(\eta \circ g_s) = \int_{a' \in \overline{\mathcal{A}}} \eta(a') \overline{\pi}(da' | f(s)).$$

我们有：

$$\rho(\eta \circ g_s) \le \overline{\pi}(f(s), \overline{\mathcal{A}}) \max_{a' \in \overline{\mathcal{A}}} \eta(a') = \max_{a \in \mathcal{A}} (\eta \circ g_s)(a) = p(\eta \circ g_s),$$

因为 $\pi(\cdot|f(s))$ 是一个概率测度，且 g_s 是满射。根据 Hahn-Banach 定理，我们可以将 ρ 延拓为 $C_0(\mathrm{A})$ 上的一个线性泛函 $\hat{}\rho$，使得对于每个 $\psi \in C_0(\mathrm{A})$，都有 $\hat{}\rho(\psi) \le p(\psi)$。由此可知，如果 $\psi \le 0$，那么 $\hat{}\rho(\psi) \le 0$；因此，如果 $\psi \ge 0$，那么 $\hat{}\rho(\psi) = -\hat{}\rho(-\psi) \ge 0$。由于这意味着 $\hat{}\rho$ 是一个正线性泛函，且 A 是一个局部紧度量空间，根据 Riesz 表示定理，存在 A 上唯一的 Radon测度 μ，使得：

$$\hat{\rho}(\psi) = \int_{a \in \mathcal{A}} \psi(a) d\mu(a).$$

由此可得，对于每个 $\eta \in C_0(\mathrm{A})$：

$$\int_{a \in \overline{\mathcal{A}}} \eta(a') \overline{\pi}(da' | f(s)) = \rho(\eta \circ g_s) = \int_{a \in \mathcal{A}} (\eta \circ g_s)(a) d\mu(a) = \int_{a' \in \overline{\mathcal{A}}} \eta(a') dg_{s*}\mu(da'),$$

其中，第一个等号根据 ρ 的定义成立，第二个等号由 μ 扩展ρ 推出，最后一个等号则由变量替换公式得出。因此，$\pi(\cdot|f(s))$ 是 μ 关于 g_s 的推前测度。令 $\pi^{\uparrow}(\cdot|s) = \overline{\mu}$，即可得证。 ■

回想一下，寻找提升策略可归结为以下问题：给定一个满射 $\bar{}$ $\bar{}$可测函数 $g_s : \mathrm{A} \to \mathrm{A}$ 和 A 上的一个概率测度 $\bar{\pi}$，是否存在 A 上的测度 $\pi^{\uparrow}$，使得由此得到的推前测度 $g_{s*}\pi^{\uparrow} = \bar{\pi}$? 这一结论对于波兰空间的解析子集也普遍成立。该结果最初由 Varadarajan（1963）证明（见引理 2.2）。

既然我们已经证明在连续情形下提升策略存在，接下来便证明连续 MDP 同态的价值等价性。证明与最优价值等价性的证明非常相似，实际上只需再对提升策略应用一次变量替换即可。

定理 16（价值等价）设 $\mathcal{M} = (\mathcal{S}\ \Sigma, \mathcal{A}\ \tau_a, R, \gamma)$ 是连续 $\overline{MDP}$ 同态 $\hbar = (f, g_s)$ 作用于 $\mathcal{M} = (\mathcal{S}\ \Sigma, \mathcal{A}\ \tau_a, R, \gamma)$ 所得的像，设 $\pi^{\uparrow}$ 是与 π 对应的提升策略。那么对于任意 $(s, a) \in \mathcal{S} \times \mathcal{A}$，我们有：

$$Q^{\pi^{\uparrow}}(s, a) = \overline{Q}^{\overline{\pi}}(f(s), g_s(a)),$$

其中，$Q^{\pi^{\uparrow}}(s, a)$ 和 $\overline{Q}^{\overline{\pi}}(f(s)\ g_s(a))$ 分别是策略 $\pi^{\uparrow}$ 和 π的动作价值函数。

证明 与定理 12 类似，我们将序列 $Q_m^{\pi^\uparrow}: S \times A \to \mathbb{R}$ 定义为

$$Q_m^{\pi^\uparrow}(s,a) = R(s,a) + \gamma \int_{s' \in \mathcal{S}} \tau_a(ds'|s) \int_{a' \in \mathcal{A}} \pi^\uparrow(da'|s') Q_{m-1}^{\pi^\uparrow}(s',a')$$

对于 $m \geq 1$ 和 $Q_0^{\pi^\uparrow}(s, a) = 0$。类似地，定义 $Q_{m-1}^{\overline{\pi}}: S \rightarrow A \to \mathbb{R}$。对于归纳情形，我们可以通过两次变量替换，分别将积分域从 S 变为 S，并将 A 变为 A:

$$\begin{aligned}
Q_m^{\pi^\uparrow}(s,a) &= R(s,a) + \gamma \int_{s' \in \mathcal{S}} \tau_a(ds'|s) \int_{a' \in \mathcal{A}} \pi^\uparrow(da'|s') Q_{m-1}^{\pi^\uparrow}(s',a') && (12)\\
&= \overline{R}(f(s), g_s(a)) + \gamma \int_{s' \in \mathcal{S}} \tau_a(ds'|s) \int_{a' \in \mathcal{A}} \pi^\uparrow(da'|s') \overline{Q}_{m-1}^{\overline{\pi}}(f(s'), g_s(a')) && (13)\\
&= \overline{R}(f(s), g_s(a)) + \gamma \int_{s' \in \mathcal{S}} \tau_a(ds'|s) \int_{\overline{a} \in \overline{\mathcal{A}}} \overline{\pi}(d\overline{a}|f(s')) \overline{Q}_{m-1}^{\overline{\pi}}(f(s'), \overline{a}) && (14)\\
&= \overline{R}(f(s), g_s(a)) + \gamma \int_{\overline{s} \in \overline{\mathcal{S}}} \overline{\tau}_{g_s(a)}(d\overline{s}|f(s)) \int_{\overline{a} \in \overline{\mathcal{A}}} \overline{\pi}(d\overline{a}|\overline{s}) \overline{Q}_{m-1}^{\overline{\pi}}(\overline{s}, \overline{a}) && (15)\\
&= \overline{Q}_{m-1}^{\overline{\pi}}(f(s),\ g_s(a))。 && (16)
\end{aligned}$$

与定理 12 类似，我们得出 $Q^{\pi^\uparrow}(s,\ a) = Q^{\overline{\pi}}(f(s),\ g_s(a))$. ■

定理 16 指出，约化马尔可夫决策过程上任意策略的价值函数，等于其对应的提升策略在原始马尔可夫决策过程上的价值函数。由于这对任意最优策略都成立，根据定理 12，提升后的最优策略也是原始马尔可夫决策过程的最优策略。因此，我们从有限情形中重新获得了连续 MDP 同态的所有理想性质。

5 同态策略梯度

为了将 MDP同态的概念直接整合到策略梯度中，并将其状态-动作抽象作为策略优化的归纳偏置，我们推导出了同态策略梯度（HPG）定理。值得注意的是，我们的结果针对连续状态与动作进行推导，并且同时适用于随机策略和确定性策略；这与 Rezaei-Shoshtari 等人（2022）的先前工作形成对比，后者对同态策略梯度定理的推导仅限于确定性策略。

在本节中，我们假设策略由可微函数（例如，神经网络）参数化，并且可以通过参数化的同态映射获得 MDP 同态像。重要的是，第 6 节详细介绍了如何学习这种参数化的 MDP 同态映射。最后，遵循有关策略梯度方法的先前研究（Sutton 等人，2000年；Silver 等人，2014 年），我们按照公式（1）中的描述定义实际MDP上的性能度量。

由于随机策略和确定性策略的策略梯度定理推导过程大不相同，所需的步骤和假设也各异，因此在本节余下部分，我们将分别独立推导随机策略和确定性策略的同态策略梯度定理。

5.1 随机 HPG 定理

随机 HPG 定理可在连续 MDP 同态的基本假设（如定义 11 所述）以及附录 A 中描述的正则性条件下推导得出。值得注意的是，对 MDP 同态映射的唯一要求是 $f : S \to S$ 和 $g_s : A \to A$ 为可测的满射，并满足定义 11 中的奖励不变性和转移的等变性。这与确定性 HPG 定理不同，后者对 g_s 另有进一步限制。

定理 17（随机同态策略梯度）设 $\mathcal{M} = (\mathcal{S}, \Sigma, \mathcal{A}, \tau, R, \gamma)$ 是连续 *MDP* 同态 $h = (f, g_s)$ 的像，该同态从 $\underline{\mathcal{M}} = (\underline{\mathcal{S}}, \underline{\Sigma}, \underline{\mathcal{A}}, \underline{\tau}, \underline{R}, \gamma)$ 映射而来。设 $\pi_\theta : \mathcal{S} \to \Delta(\underline{\mathcal{A}})$ 是定义在 $\mathcal{M}$ 上的随机策略。那么，性能度量 $J(\theta)$ 关于 θ 的梯度为：

$$\nabla_\theta J(\theta) = \int_{\bar{s} \in \bar{\mathcal{S}}} \rho^{\bar{\pi}_\theta}(\bar{s}) \int_{\bar{a} \in \bar{\mathcal{A}}} \bar{Q}^{\bar{\pi}_\theta}(\bar{s}, \bar{a}) \nabla_\theta \bar{\pi}_\theta(d\bar{a}|\bar{s}) d\bar{s},$$

其中 $\rho^{\bar{\pi}_\theta}(\bar{s})$ 是 $\mathcal{M}$ 遵循随机策略 $\pi_\theta(a|s)$ 时的折扣状态分布。

证明 证明过程与随机策略梯度定理（Sutton 等人，2000年）类似。首先，我们推导 $\nabla_\theta V^{\pi_\theta^\uparrow}(s)$ 的递归表达式如下：

$$\begin{aligned}
\nabla_\theta V^{\pi_\theta^\uparrow}(s) &= \nabla_\theta \int_{a \in \mathcal{A}} \pi_\theta^\uparrow(da|s) Q^{\pi_\theta^\uparrow}(s, a) \\
&= \int_{a \in \mathcal{A}} \Big[\nabla_\theta \pi_\theta^\uparrow(da|s) Q^{\pi_\theta^\uparrow}(s, a) + \pi_\theta^\uparrow(da|s) \nabla_\theta Q^{\pi_\theta^\uparrow}(s, a) \Big] \\
&= \int_{a \in \mathcal{A}} \Big[\nabla_\theta \pi_\theta^\uparrow(da|s) Q^{\pi_\theta^\uparrow}(s, a) + \pi_\theta^\uparrow(da|s) \nabla_\theta \Big(R(s, a) + \gamma \int_{s' \in \mathcal{S}} \tau_a(ds'|s) V^{\pi_\theta^\uparrow}(s') \Big) \Big] \\
&= \int_{a \in \mathcal{A}} \nabla_\theta \pi_\theta^\uparrow(da|s) Q^{\pi_\theta^\uparrow}(s, a) + \gamma \int_{a \in \mathcal{A}} \pi_\theta^\uparrow(da|s) \int_{s' \in \mathcal{S}} \tau_a(ds'|s) \nabla_\theta V^{\pi_\theta^\uparrow}(s') \\
&= \int_{a \in \mathcal{A}} \nabla_\theta \pi_\theta^\uparrow(da|s) \bar{Q}^{\bar{\pi}_\theta}(f(s), g_s(a)) + \gamma \int_{a \in \mathcal{A}} \pi_\theta^\uparrow(da|s) \int_{s' \in \mathcal{S}} \tau_a(ds'|s) \nabla_\theta V^{\bar{\pi}_\theta}(f(s')) \qquad (17) \\
&= \int_{a \in \mathcal{A}} \nabla_\theta \pi_\theta^\uparrow(da|s) \bar{Q}^{\bar{\pi}_\theta}(f(s), g_s(a)) + \gamma \int_{a \in \mathcal{A}} \pi_\theta^\uparrow(da|s) \int_{\bar{s} \in \bar{\mathcal{S}}} \bar{\tau}_{g_s(a)}(d\bar{s}|f(s)) \nabla_\theta V^{\bar{\pi}_\theta}(\bar{s}) \qquad (18) \\
&= \int_{\bar{a} \in \bar{\mathcal{A}}} \nabla_\theta \bar{\pi}_\theta(d\bar{a}|f(s)) \bar{Q}^{\bar{\pi}_\theta}(f(s), \bar{a}) + \gamma \int_{\bar{a} \in \bar{\mathcal{A}}} \bar{\pi}_\theta(d\bar{a}|f(s)) \int_{\bar{s} \in \bar{\mathcal{S}}} \bar{\tau}_{\bar{a}}(d\bar{s}|f(s)) \nabla_\theta V^{\bar{\pi}_\theta}(\bar{s}).
\end{aligned}$$

这里，我们在公式（17）中应用了价值等价；在公式（18）中，分别通过 $\tau_a(\cdot|s)$ 和 $\bar{\tau}_{g_s(a)}(\cdot|f(s))$ 对 S 到 S 上方的变量进行变量替换；并分别通过提升策略和抽象MDP上的策略对 A 到 A 上方的变量进行变量替换。请注意，要严格验证对 θ 的梯度与对 A 的积分能否交换，可能需要谨慎处理；然而，这是证明连续域上任何类型的策略梯度结果所必需的条件，并非随机 HPG 定理所特有。

与随机策略梯度定理的证明一样，我们可以继续在抽象*MDP* M的空间中展开 $\nabla_\theta V^{\pi_\theta}(s)$ 的定义。令 $\pi^k(\cdot|f(s))$ 表示从状态 $f(s)$ 出发，遵循 π 滚转 k 步后在S上的概率分布，则有：

$$\nabla_\theta V^{\pi_\theta^\uparrow}(s) = \sum_{k=0}^{\infty} \gamma^k \int_{\bar{s} \in \bar{\mathcal{S}}} \bar{\pi}_\theta^k(d\bar{s}|f(s)) \int_{\bar{a} \in \bar{\mathcal{A}}} \nabla_\theta \bar{\pi}_\theta(d\bar{a}|f(s)) \bar{Q}^{\bar{\pi}_\theta}(f(s), \bar{a}).$$

最后，我们得出：

$$\begin{aligned}\nabla_\theta J(\theta) &= \nabla_\theta V^{\pi^\uparrow_\theta}(s)\\ &= \int_{\overline{s}\in\overline{\mathcal{S}}}\sum_{k=0}^{\infty}\gamma^k\overline{\pi}^k_\theta(d\overline{s}|f(s))\int_{\overline{a}\in\overline{\mathcal{A}}}\nabla_\theta\overline{\pi}_\theta(d\overline{a}|f(s))\overline{Q}^{\overline{\pi}_\theta}(f(s),\overline{a})\\ &= \int_{\overline{s}\in\overline{\mathcal{S}}}\rho^{\overline{\pi}_\theta}(d\overline{s})\int_{\overline{a}\in\overline{\mathcal{A}}}\overline{Q}^{\overline{\pi}_\theta}(\overline{s},\overline{a})\nabla_\theta\overline{\pi}_\theta(d\overline{a}|\overline{s}),\end{aligned}$$

如期望的那样，这里，$\rho^{\overline{\pi}_\theta}(\overline{s})$ 是由策略 $\overline{\pi}_\theta$ 诱导的折扣平稳分布。 ■

w

5.2 确定性 HPG 定理

与随机 HPG 中同态映射可以是任意可测满射不同，确定性情形要求动作编码器 $g_s : A \to A$ 是一个局部微分同胚（定义见附录 B）。这一要求的重要含义是，动作编码器 g_s 必须是局部双射，因此抽象动作空间的维数必须与实际动作空间相同。首先，我们展示策略梯度的等价性：

定理 18（确定性策略梯度的等价性） 设 $\overline{M} = (\overline{S}, \overline{\Sigma}, \overline{A}, \overline{\tau}, R, \overline{\gamma})$ 是连续 *MDP* 同态 $h = (f, g_s)$ 作用于 $M = (S, \Sigma, A, \tau, R, \gamma)$ 后得到的像，设 $\pi^\uparrow_\theta : S \to A$ 是与抽象确定性策略 $\pi_\theta : \overline{S} \to \overline{A}$ 相对应的提升的确定性策略。那么，对于任意 $(s,a) \in S \times A$，我们有：

$$\nabla_a Q^{\pi^\uparrow_\theta}(s,a)\Big|_{a=\pi^\uparrow_\theta(s)}\nabla_\theta\pi^\uparrow_\theta(s) = \nabla_{\overline{a}}\overline{Q}^{\overline{\pi}_\theta}(\overline{s},\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(\overline{s})}\nabla_\theta\overline{\pi}_\theta(\overline{s}).$$

证明 在假设附录 A 中所述条件成立的情况下，我们首先使用链式法则，对确定性策略提升关系关于策略参数 θ 求导：

$$\begin{aligned}(g_s\circ\pi^\uparrow_\theta)(s) &= (\overline{\pi}_\theta\circ f)(s)\\ d(g_s\circ\pi^\uparrow_\theta)_\theta(s) &= d(\overline{\pi}_\theta\circ f)_\theta(s)\\ d(g_s)_{\pi^\uparrow_\theta(s)}\circ d(\pi^\uparrow_\theta)_\theta(s) &= d(\overline{\pi}_\theta\circ f)_\theta(s)\\ \underbrace{\nabla_a g_s(a)\big|_{a=\pi^\uparrow_\theta(s)}}_{P}\nabla_\theta\pi^\uparrow_\theta(s) &= \nabla_\theta\overline{\pi}_\theta(f(s)),\end{aligned}\tag{19}$$

其中 $\bigcirc$ 是复合算子，矩阵的维度分别为 $P \in \mathbb{R}^{|\mathcal{A}|\times|A|}$，$\nabla_\theta\pi^\uparrow_\theta(s) \in \mathbb{R}^{|A|\times|\theta|}$ 和 $\nabla_\theta\overline{\pi}_\theta(s) \in \mathbb{R}^{|\mathcal{A}|\times|\theta|}$。其次，我们对价值等价定理关于动作 a 使用链式法则求导：

$$\begin{aligned}Q^{\pi^\uparrow_\theta}(s,a) &= \overline{Q}^{\overline{\pi}_\theta}(f(s),g_s(a))\\ dQ^{\pi^\uparrow_\theta}(s,a)_a &= d\overline{Q}^{\overline{\pi}_\theta}(f(s),g_s(a))_a\\ \nabla_a Q^{\pi^\uparrow_\theta}(s,a)\big|_{a=\pi^\uparrow_\theta(s)} &= \nabla_{\overline{a}}\overline{Q}^{\overline{\pi}_\theta}(f(s),\overline{a})\big|_{\overline{a}=\overline{\pi}_\theta(f(s))}\underbrace{\nabla_a g_s(a)\big|_{a=g_s^{-1}(\overline{\pi}_\theta(f(s)))}}_{P},\end{aligned}\tag{20}$$

其中，矩阵的维度为 $\nabla_a Q^{\pi_\theta^\uparrow}(s,\ a) \in \mathbb{R}^{|A|}$、$\nabla_a Q^{\pi_\theta}(s,\ a) \in \mathbb{R}^{|\mathcal{A}|}$，与之前类似，$P \in \mathbb{R}^{|\mathcal{A}|\times|A|}$。由于我们假设 g_s 是局部微分同胚，反函数定理指出矩阵 P 可逆，因此我们将方程 (20) 两边右乘 P^{-1}，再将所得方程左乘方程 (19)，从而得到所需结果：

$$\nabla_a Q^{\pi_\theta^\uparrow}(s,a)\big|_{a=\pi_\theta^\uparrow(s)} P^{-1} P \nabla_\theta \pi_\theta^\uparrow(s) = \nabla_{\bar{a}} \overline{Q}^{\bar{\pi}_\theta}(f(s),\bar{a})\big|_{\bar{a}=\bar{\pi}(f(s))} \nabla_\theta \bar{\pi}_\theta(f(s))$$
$$\nabla_a Q^{\pi_\theta^\uparrow}(s,a)\big|_{a=\pi_\theta^\uparrow(s)} \nabla_\theta \pi_\theta^\uparrow(s) = \nabla_{\bar{a}} \overline{Q}^{\bar{\pi}_\theta}(f(s),\bar{a})\big|_{\bar{a}=\bar{\pi}_\theta(f(s))} \nabla_\theta \bar{\pi}_\theta(f(s)).$$

■

定理 18 指出，尽管底层空间经过了抽象化，抽象MDP的梯度仍与原始 MDP的梯度等价。这意味着，在抽象MDP上执行 HPG 等同于在实际MDP上执行 DPG，因此我们可以如图 2 所示，协同使用二者来更新同一组参数 θ。

虽然可以直接套用定理 18 来替换标准 DPG 的梯度，但从理论上讲，这样做并不能得到任何有用的结果，因为期望仍然是相对于参数为 $\pi_\theta^\uparrow(s)$ 时实际MDP M 的平稳状态分布来估计的。然而，利用连续 MDP 同态的性质，我们可以将积分空间从 S 改为 S，从而根据参数为 $\pi_\theta(s)$ 时抽象MDP M 的平稳状态分布来估计策略梯度：

定理 19（确定性同态策略梯度）设 $\overline{\mathcal{M}} = (\overline{\mathcal{S}}, \overline{\Sigma}, \overline{\mathcal{A}}, \overline{\tau}, \overline{R}, \overline{\gamma})$ 是连续 *MDP* 同态 $h = (f, g_s)$ 的像，该同态从 $M = (\mathcal{S}, \Sigma, \mathcal{A}, \tau, R, \gamma)$ 出发。设 $\pi_\theta : \mathcal{S} \to \mathcal{A}$ 是定义在 $\mathcal{M}$ 上的确定性抽象策略。那么，定义在实际 *MDP* $\mathcal{M}$ 上的性能度量 $J(\theta)$ 关于 θ 的梯度为：

$$\nabla_\theta J(\theta) = \int_{\bar{s}\in\overline{\mathcal{S}}} \rho^{\bar{\pi}_\theta}(\bar{s}) \nabla_{\bar{a}} \overline{Q}^{\bar{\pi}_\theta}(\bar{s},\bar{a})\Big|_{\bar{a}=\bar{\pi}_\theta(\bar{s})} \nabla_\theta \bar{\pi}_\theta(\bar{s}) d\bar{s},$$

其中，$\rho^{\pi_\theta}(s)$ 是遵循确定性策略 $\pi_\theta(s)$ 的 $\mathcal{M}$ 的折扣状态分布。

证明 证明过程与确定性策略梯度定理（Silver 等人，2014 年）的推导思路相同，但需要额外步骤将积分空间从 S 变为 S。首先，我们推导出 $\nabla_\theta V^{\pi_\theta^\uparrow}(s)$ 的递归表达式如下：

$$\begin{aligned}
\nabla_\theta V^{\pi_\theta^\uparrow}(s) &= \nabla_\theta Q^{\pi_\theta^\uparrow}\left(s, \pi_\theta^\uparrow(s)\right) \\
&= \nabla_\theta \left[R(s, \pi_\theta^\uparrow(s)) + \gamma \int_{s'\in\mathcal{S}} \tau_{\pi_\theta^\uparrow(s)}(ds'|s) V^{\pi_\theta^\uparrow}(s') \right] \\
&= \nabla_\theta \pi_\theta^\uparrow(s) \nabla_a R(s,a)\Big|_{a=\pi_\theta^\uparrow(s)} \\
&\quad + \gamma \int_{s'\in\mathcal{S}} \left[\tau_{\pi_\theta^\uparrow(s)}(ds'|s) \nabla_\theta V^{\pi_\theta^\uparrow}(s') + \nabla_\theta \pi_\theta^\uparrow(s) \nabla_a \tau_a(ds'|s)\Big|_{a=\pi_\theta^\uparrow(s)} V^{\pi_\theta^\uparrow}(s') \right] \quad (21) \\
&= \nabla_\theta \pi_\theta^\uparrow(s) \nabla_a \left[R(s,a) + \gamma \int_{s'\in\mathcal{S}} \tau_a(ds'|s) V^{\pi_\theta^\uparrow}(s') \right]\Big|_{a=\pi_\theta^\uparrow(s)} + \gamma \int_{s'\in\mathcal{S}} \tau_{\pi_\theta^\uparrow(s)}(ds'|s) \nabla_\theta V^{\pi_\theta^\uparrow}(s')
\end{aligned}$$

$$= \nabla_\theta \pi_\theta^\uparrow(s) \nabla_a Q^{\pi_\theta^\uparrow}(s,a)\Big|_{a=\pi_\theta^\uparrow(s)} + \gamma \int_{s' \in \mathcal{S}} \tau_{\pi_\theta^\uparrow(s)}(ds'|s) \nabla_\theta V^{\overline{\pi}_\theta}(f(s')) \tag{22}$$

$$= \nabla_\theta \pi_\theta^\uparrow(s) \nabla_a Q^{\pi_\theta^\uparrow}(s,a)\Big|_{a=\pi_\theta^\uparrow(s)} + \gamma \int_{\overline{s}' \in \overline{\mathcal{S}}} \overline{\tau}_{g_s(\pi_\theta^\uparrow(s))}(d\overline{s}'|f(s)) \nabla_\theta V^{\overline{\pi}_\theta}(f(s')) \tag{23}$$

$$= \nabla_\theta \overline{\pi}_\theta(f(s)) \nabla_{\overline{a}} \overline{Q}^{\overline{\pi}_\theta}(f(s),\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(f(s))} + \gamma \int_{\overline{s}' \in \overline{\mathcal{S}}} \overline{\tau}_{\overline{\pi}_\theta(\overline{s})}(d\overline{s}'|\overline{s}) \nabla_\theta V^{\overline{\pi}_\theta}(\overline{s}') \tag{24}$$

$$= \nabla_\theta \overline{\pi}_\theta(f(s)) \nabla_{\overline{a}} \overline{Q}^{\overline{\pi}_\theta}(f(s),\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(f(s))} + \gamma \int_{\overline{s}' \in \overline{\mathcal{S}}} p(\overline{s} \to \overline{s}', 1, \overline{\pi}_\theta) \nabla_\theta V^{\overline{\pi}_\theta}(\overline{s}') d\overline{s}'.$$

其中，$p(\overline{s} \to \overline{s}', t, \pi_\theta)$ 表示在策略 $\pi_\theta(s)$ 下，经过 t 个时间步从 s 到 s' 的概率。在公式 (21) 中，由于函数连续性方面的正则性条件，我们得以应用莱布尼茨积分法则，交换求导与积分的顺序。在公式 (22) 中，我们使用了价值等价性质；在公式 (23) 中，我们使用了基于 $\tau_a(\cdot|s)$ 关于 f 的推前测度的变量替换公式。最后，在公式 (24) 中，我们使用了定理 18 中策略梯度的等价性。通过递归地滚转上述公式，我们得到：

$$\begin{aligned}
\nabla_\theta V^{\pi_\theta^\uparrow}(s) &= \nabla_\theta \overline{\pi}_\theta(f(s)) \nabla_{\overline{a}} \overline{Q}^{\overline{\pi}_\theta}(f(s),\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(f(s))} \\
&\quad + \gamma \int_{\overline{s}' \in \overline{\mathcal{S}}} p(\overline{s} \to \overline{s}', 1, \overline{\pi}_\theta) \nabla_\theta \overline{\pi}_\theta(f(s')) \nabla_{\overline{a}} \overline{Q}^{\overline{\pi}_\theta}(f(s'),\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(f(s'))} d\overline{s}' \\
&\quad + \gamma^2 \int_{\overline{s}' \in \overline{\mathcal{S}}} p(\overline{s} \to \overline{s}', 1, \overline{\pi}_\theta) \int_{\overline{s}'' \in \overline{\mathcal{S}}} p(\overline{s}' \to \overline{s}'', 1, \overline{\pi}_\theta) \nabla_\theta V^{\pi_\theta^\uparrow}(f(s'')) d\overline{s}'' d\overline{s}' \\
&= \nabla_\theta \overline{\pi}_\theta(f(s)) \nabla_{\overline{a}} \overline{Q}^{\overline{\pi}_\theta}(f(s),\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(f(s))} \\
&\quad + \gamma \int_{\overline{s}' \in \overline{\mathcal{S}}} p(\overline{s} \to \overline{s}', 1, \overline{\pi}_\theta) \nabla_\theta \overline{\pi}_\theta(f(s')) \nabla_{\overline{a}} \overline{Q}^{\overline{\pi}_\theta}(f(s'),\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(f(s'))} d\overline{s}' \\
&\quad + \gamma^2 \int_{\overline{s}'' \in \overline{\mathcal{S}}} p(\overline{s} \to \overline{s}'', 2, \overline{\pi}_\theta) \nabla_\theta V^{\overline{\pi}_\theta}(f(s'')) d\overline{s}'' \qquad (25) \\
&\vdots \\
&= \int_{\overline{s}' \in \overline{\mathcal{S}}} \sum_{t=0}^{\infty} \gamma^t p(\overline{s} \to \overline{s}', t, \overline{\pi}_\theta) \nabla_\theta \overline{\pi}_\theta(f(s)) \nabla_{\overline{a}} \overline{Q}^{\overline{\pi}_\theta}(f(s),\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(f(s))} d\overline{s}'. \qquad (26)
\end{aligned}$$

在公式 (25) 中，我们使用富比尼定理交换了积分顺序；该定理要求 $\|\nabla_\theta V^{\pi_\theta}(s)\|$ 有界，正如正则性条件所述。最后，我们对初始状态分布取 $\nabla_\theta V^{\pi_\theta^\uparrow}(s)$ 的期望：

$$\begin{aligned}
\nabla_\theta J(\theta) &= \nabla_\theta \int_{s \in \mathcal{S}} p_1(s) V^{\pi_\theta^\uparrow}(s) ds \\
&= \int_{s \in \mathcal{S}} p_1(s) \nabla_\theta V^{\pi_\theta^\uparrow}(s) ds \\
&= \int_{s \in \mathcal{S}} p_1(s) \int_{\overline{s}' \in \overline{\mathcal{S}}} \sum_{t=0}^{\infty} \gamma^t p(\overline{s} \to \overline{s}', t, \overline{\pi}_\theta) \nabla_\theta \overline{\pi}_\theta(f(s)) \nabla_{\overline{a}} \overline{Q}^{\overline{\pi}_\theta}(f(s),\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(f(s))} d\overline{s}' ds \\
&= \int_{\overline{s} \in \overline{\mathcal{S}}} \overline{p}_1(\overline{s}) \int_{\overline{s}' \in \overline{\mathcal{S}}} \sum_{t=0}^{\infty} \gamma^t p(\overline{s} \to \overline{s}', t, \overline{\pi}_\theta) \nabla_\theta \overline{\pi}_\theta(f(s)) \nabla_{\overline{a}} \overline{Q}^{\overline{\pi}_\theta}(f(s),\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(f(s))} d\overline{s}' d\overline{s} \qquad (27) \\
&= \int_{\overline{s} \in \overline{\mathcal{S}}} \rho^{\overline{\pi}_\theta}(\overline{s}) \nabla_\theta \overline{\pi}_\theta(\overline{s}) \nabla_{\overline{a}} \overline{Q}^{\overline{\pi}_\theta}(\overline{s},\overline{a})\Big|_{\overline{a}=\overline{\pi}_\theta(\overline{s})} d\overline{s}. \qquad (28)
\end{aligned}$$

其中， $\bar{\rho}^{\pi_\theta}(\bar{s})$ 是由策略 π_θ 所诱导的折扣平稳分布。在公式 (27) 中，我们使用了变量替换公式。与前面的步骤类似，我们使用莱布尼茨积分法则交换积分与求导的顺序，并使用富比尼定理交换积分的顺序。

■

5.3 比较随机与确定性 HPG 定理

同态策略梯度（定理 17 和 19）构成了我们提出的同态演员-评论家算法的基础，其意义有两方面。首先，除了标准策略梯度估计器之外，我们还可以基于近似 MDP 同态像，获得另一种策略梯度估计。尽管这两种策略梯度估计并非统计独立，因为它们通过同态映射联系在一起，但 HPG 有望以引入一定偏差为代价，降低方差；这种偏差源于 MDP 同态的近似。

其次，由于 MDP 的最小像就是 MDP 同态像（Ravindran 和 Barto，2001年），抽象评论家 Q^{π_θ} 是在一个简化的问题上训练的。换言之，用于训练 Q^{π_θ} 的每个抽象状态-动作对（$s,\ a$）都代表了在 MDP 同态关系下等价的所有（$\bar{s},\ a$）对，从而提升样本效率。然而，复杂度降低的幅度取决于环境中的近似对称性，我们的实证结果也支持这一点。

图 2 展示了同态策略梯度定理及其与标准策略梯度定理并用时的示意图。作为本节的总结，我们对随机和确定性 HPG 变体进行概念性比较，并将在第 7 节进一步进行实证比较。

动作空间中的维度缩减。 马尔可夫决策过程同态的一个关键方面是“折叠”的概念：状态映射 f 和依赖状态的动作映射 g_s都是满射。例如，物理系统相对于某个动作的连续对称性对应于某个量的不变性，并能有效降低动作空间的维度（Noether，1971；Bluman 和 Kumei，2013）。在强化学习智能体的背景下，识别并利用环境的连续对称性的能力可以降低动作空间的维度，从而显著简化学习问题。然而，这类动作缩减不满足确定性 HPG 定理的条件，因为该定理要求动作映射 g_s是局部微分同胚。因此，其基本假设不涵盖严格折叠。相比之下，随机 HPG 不对 g_s施加任何额外结构，因此能够有效缩减动作空间的维度，同时不会损害策略的最优性。

最大熵强化学习。 随机策略的相关结果为在更多种类的算法中集成马尔可夫决策过程同态提供了理论保证。例如，最大熵强化学习框架将期望回报形式推广为同时考虑策略的熵，从而提升鲁棒性和探索能力（Ziebart et al., 2008; Ziebart, 2010; Haarnoja 等人, 2018）。当然，这类方法仅适用于随机策略。因此，与确定性同态策略梯度相比，

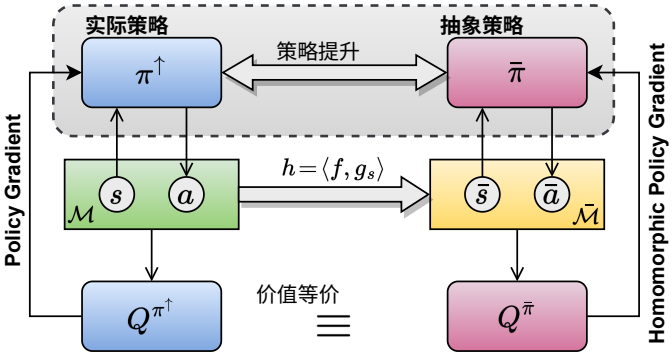


图 2：HPG 示意图。实际MDP M 用于训练 $Q^{\pi^{\uparrow}}$，并使用标准策略梯度定理更新 $\pi^{\uparrow}$；抽象MDP M 用于训练 $Q^{\bar{\pi}}$，并使用同态策略梯度定理更新 $\bar{\pi}$。M 是通过学习同态映射 $h=(f, g_s)$ 获得的 MDP M 的 MDP 同态像。策略 $\pi^{\uparrow}$ 和 $\bar{\pi}$ 通过提升过程相互耦合。

随机 HPG 能够受益于加入策略的熵。重要的是，推前测度的熵至多等于原始测度的熵（这是条件詹森不等式的推论——参见 Smorodinsky（2006年））；因此，在抽象MDP中寻求最大熵策略，也会对应于原始空间中的高熵策略。

样本效率。确定性和随机策略梯度方法的样本效率会因算法选择、网络架构、探索策略和实现方式而有很大差异（Henderson 等人，2018年）；不过，通常观察到确定性策略梯度方法的样本效率更高（Lillicrap 等人，2015年；Fujimoto 等人，2018年；Barth-Maron 等人，2018年）。一种假设是，与同时对状态空间和动作空间积分的随机策略梯度不同，确定性策略梯度只对状态空间积分，因此策略梯度估计的样本效率更高，尤其是在高维动作空间中（Silver 等人，2014 年）。同样的推理也适用于 HPG 的各种变体。我们已在第 7 节中针对多种环境开展了全面的实证研究，以进一步探究这两种 HPG 变体的特性。

6 同态演员-评论家算法

本节将概述一种基于随机与确定性 HPG 定理的实用深度强化学习算法，称为深度同态策略梯度 （DHPG）算法。尽管在随机策略和确定性策略这两种情况下，算法的整体结构以及学习 MDP 同态映射的方式相似，但随机情况下的策略提升过程还需要额外的复杂步骤。算法 1 给出了 DHPG 算法的伪代码。

将像素观测记为 o_t，底层状态记为 s_t，抽象状态记为 s_t，DHPG 算法的主要组成部分包括：MDP 同态映射 $h_{\phi,\eta}=(f_\phi(s_t)\, g_\eta(s_t, a_t))$、像素编码器 $E_\mu(o_t)$、实际评论家 $Q_\psi(s_t, a_t)$ 和策略 $\pi_\theta^{\uparrow}(a_t|s_t)$，

抽象评论家 $Q_{\psi}(\bar{s}_t, \bar{a}_t)$ 和策略 $\pi_\theta(a_t|\bar{s}_t)$、奖励预测器 $R_\rho(s_t)$，以及概率转移动力学 $\tau_\nu(s_{t+1}|s_t, a_t)$，其输出为高斯分布。最后，我们利用目标评论家网络 $Q_{\psi'}$ 和 $Q_{\psi'}$，以实现更稳定的训练，并使用标准回放缓冲区（Mnih 等人，2013年；Lillicrap 等人，2015年）。

策略提升过程。一般而言，对于每个博雷尔集 $\beta \subseteq \mathcal{A}$ 和 $s \in \mathcal{S}$，提升策略需要满足关系 $\pi^\uparrow(g_s^{-1}(\beta)|s) = \pi(\beta|f(s))$。如第4.2节所述，命题15证明了提升策略 $\pi^\uparrow$ 可由抽象策略 π 得到，但没有提供构造提升策略的明确方法。

如果抽象策略是确定性的，则可以简单地通过为原像 $g_s^{-1}(\pi(f(s)))$ 选取一个代表元来得到提升策略。如果我们像第5.2节中假设的那样，选择 g_s 为双射，则可将提升策略唯一地定义为 $\pi^\uparrow(s) = g_s^{-1}(\pi(f(s)))$。这样就可以使用同一个网络对这两种策略进行参数化。在实践中，我们对实际策略 $\pi_\theta^\uparrow$ 进行参数化，并将抽象策略表示为 $\pi_\theta(f(s)) = g_s(\pi_\theta^\uparrow(s))$。

对于随机抽象策略，求解过程并没有那么直接；尽管已有用于构造随机反问题解的贝叶斯方法（Butler 等人，2018年），我们选择基于采样的方法，推导出一个损失函数，作为策略提升过程的近似（Kaipio 和 Somersalo，2006年）。利用推前测度的变量替换公式，我们可以证明，两种策略下抽象动作的条件期望相等：

$$\mathbb{E}_{\pi^\uparrow}[g_s(a)|s] = \int_{a\in\mathcal{A}} g_s(a)\pi^\uparrow(da|s) = \int_{\bar{a}\in\bar{\mathcal{A}}} \bar{a}\,\bar{\pi}(d\bar{a}|\bar{s}) = \mathbb{E}_{\bar{\pi}}[\bar{a}|f(s)].$$

所有有限阶矩也有类似的结果；特别是，两种策略下抽象动作的条件方差相等——即：

$$\mathrm{Var}_{\pi^\uparrow}[g_s(a)|s] = \mathrm{Var}_{\bar{\pi}}[\bar{a}|f(s)].$$

因此，我们可以推导出策略提升损失，用以衡量两种策略相对于 MDP 同态映射和提升过程的一致性。假设策略 $\pi_\theta^\uparrow$ 和 π_θ 由相互独立的神经网络参数化，则通过匹配在回放缓冲区中采样的观测条件下抽象动作的条件期望和标准差（SD），即可得到损失函数：

$$\mathcal{L}_{\text{lift.}}(\theta,\bar{\theta}) = \left(\mathbb{E}_{\pi_\theta^\uparrow}[g_s(a)|s] - \mathbb{E}_{\bar{\pi}_{\bar{\theta}}}[\bar{a}|f(s)]\right)^2 + \left(\mathrm{SD}_{\pi_\theta^\uparrow}[g_s(a)|s] - \mathrm{SD}_{\bar{\pi}_{\bar{\theta}}}[\bar{a}|f(s)]\right)^2. \tag{29}$$

如前所述，对于确定性DHPG，公式（29）中的策略提升损失并非必需。

评论家训练。为更快地传播奖励，实际评论家和抽象评论家均使用 n 步 TD 误差进行训练（Barth-Maron 等人，2018年）。因此，每个评论家的损失函数定义为对从回放缓冲区 B 中抽取的状态转移样本所估计的 n 步贝尔曼误差的期望：

$$\mathcal{L}_{\text{actual critic}}(\psi) = \mathbb{E}_{(s,a,s',r)\sim\mathcal{B}}\Big[\Big(R_t^{(n)} + \gamma^n Q_{\psi'}(s_{t+n}, a_{t+n}) - Q_\psi(s_t,a_t)\Big)^2\Big] \tag{30}$$

$$\mathcal{L}_{\text{abstract critic}}(\bar{\psi},\phi,\eta) = \mathbb{E}_{(s,a,s',r)\sim\mathcal{B}}\Big[\Big(R_t^{(n)} + \gamma^n \bar{Q}_{\bar{\psi}'}(\bar{s}_{t+n}, \bar{a}_{t+n}) - \bar{Q}_{\bar{\psi}}(\bar{s}_t,\bar{a}_t)\Big)^2\Big], \tag{31}$$

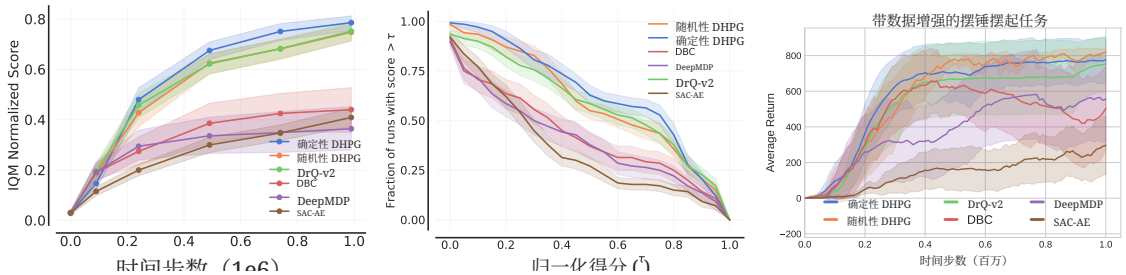
其中，$\bar{s}_t = f_\phi(s_t)$ 和 $\bar{a}_t = g_\eta(s_t, a_t)$ 使用学习得到的MDP同态计算，ψ' 和 $\bar{\psi}'$ 表示目标网络的参数，而 $R_t^{(n)} = \sum_{i=0}^{n-1}\gamma^i r_{t+i}$ 是 n 步回报。

训练演员。 对于随机策略，我们使用标准策略梯度定理（Sutton 等人，2000年）训练实际策略，并通过随机 HPG（定理 17）训练抽象策略。虽然我们可以采用任何随机演员-评论家算法来训练实际策略，但我们使用 SAC（Haarnoja 等人，2018年），以进一步展示我们的方法在最大熵强化学习框架中的适用性。值得注意的是，正如第5.3节所述，只有在更新实际策略时才需要考虑熵正则项。

对于确定性策略，使用确定性策略梯度（Silver 等人，2014 年）训练实际策略，并使用确定性同态策略梯度（定理 19）更新抽象策略。值得注意的是，由于在此情况下实际策略和抽象策略共享同一组参数（$\theta = \theta$），因此两种策略更新都应用于同一组策略参数。

学习连续 MDP 同态映射。 我们使用松弛双模拟度量（Taylor 等人，2008）学习连续 MDP 同态映射，这与此前关于连续 MDP 同态的研究（Rezaei-Shoshtari，2022 年）类似。我们使用松弛双模拟度量（Taylor 等人，2008）及公式（32）提出一个损失

算法 1 深度同态策略梯度（DHPG） **1: 初始化目标网络** $\psi' \leftarrow \psi$, $\psi' \leftarrow \psi$. **2: 对于** $t = 0$ **到** T **执行 3: 编码观测：** $s_t = E_\mu(o_t)$ **4: 选择并执行动作：5: 如果是随机策略则 6:** $a_t \sim \pi^{\uparrow}_\theta(\cdot|s_t)$ **7: 否则 8:** $a_t = \pi^{\uparrow}_\theta(s_t) + \epsilon$**，其中** $\epsilon \sim N(0, \sigma)$ **9: 结束如果 10: 将转移存储到回放缓冲区：** $B \leftarrow B \cup (s_t, a_t, s_{t+1}, r_t)$ **11: 采样一个小批量：** $B_i \sim B$ **12: 对小批量进行置换：** $B_j = \text{permute}(B_i)$ **13: 训练** $h_\phi, \eta, E_\mu, \tau_\nu, R_\rho$**：** $\mathbf{L}_{\text{lax}} + \mathbf{L}_{\text{h}}$ ▷ **公式 33-34 14: 训练评论家** Q_ψ, Q_ψ**：** $\mathbf{L}_{\text{actual}} + \mathbf{L}_{\text{abstract}}$ ▷ **公式 30-31 15: 如果是随机策略则 16: # 提升策略和抽象策略分别由** θ **和** θ**参数化 17: 使用 PG** $+$ **最大熵 SAC（Haarnoja 等人，2018）训练** $\pi^{\uparrow}_\theta$ **18: 使用随机 HPG** ▷ **定理 17 训练**π_θ **19: 使用策略提升损失更新策略** $\pi^{\uparrow}_\theta$ **和** π_θ **：** $\mathbf{L}_{\text{lift.}}$ ▷ **公式 29 20: 否则 21: # 提升策略和抽象策略共享相同参数** θ**22: 使用 DPG** ▷ **TD3（Fujimoto 等人，2018）训练** $\pi^{\uparrow}_\theta$ **23: 使用确定性同态策略梯度** ▷ **定理 19 训练** π_θ **24: 结束如果 25: 更新目标网络：** $\psi' \leftarrow \alpha\psi + (1-\alpha)\psi'$, $\psi' \leftarrow \alpha\psi + (1-\alpha)\psi'$**26: 结束循环**



(a) 样本效率。 (b) 性能剖面。 (c) 学习曲线。

图3：使用 10 个随机种子获得的 DM Control 像素观测任务结果。RLiable 指标汇总了 14 项任务的结果。所有方法均**使用**图像增强。(a)RLiable 四分位均值分数随步数的变化，用于比较样本效率；(b)在 500k 步时的 RLiable 性能剖面；(c) 摆锤摆起任务的学习曲线。完整结果见附录 C.1。阴影区域表示 95% 置信区间。

将松弛双模拟等价状态映射到抽象空间中更近位置的函数。宽松双模拟度量适用于连续动作，并且作为一种（伪）度量，它可以自然地表示近似MDP同态。两个状态-动作对 (s_i, a_i) 和 (s_j, a_j) 之间的宽松双模拟度量定义为：

$$d_{\text{lax}}\big((s_i,a_i),(s_j,a_j)\big) = c_r|R(s_i,a_i) - R(s_j,a_j)| + c_t W_1\big(\tau(\cdot|s_i,a_i),\tau(\cdot|s_j,a_j)\big), \tag{32}$$

其中，第一项用于衡量奖励项之间的距离，而 W_1 是 Kantorovich 度量，用于衡量转移概率之间的距离。遵循先前利用双模拟进行表征学习的研究思路（Zhang 等人，2020年），我们针对从回放缓冲区中采样得到的转移元组对，定义了所提出的松弛双模拟损失。我们对样本进行置换，以计算它们在抽象空间中的成对距离以及成对松弛双相似距离。因此，我们最小化这两个项之间的距离：

$$\mathrm{L}_{\text{lax}}(\phi,\eta) = \mathbb{E}_{\mathcal{B}}\Big[\|f_\phi(s_i) - f_\phi(s_j)\|_1 - \|r_i - r_j\|_1 - \alpha W_2\big(\overline{\tau}_\nu(\cdot|f_\phi(s_i),g_\eta(s_i,a_i)),\overline{\tau}_\nu(\cdot|f_\phi(s_j),g_\eta(s_j,a_j))\big)\Big]^2 \tag{33}$$

这里的期望是对从回放缓冲区中采样的两个样本 $(s_i,a_i,s_i',r_i),(s_j,a_j,s_j',r_j) \sim \mathrm{B}$ 计算的。

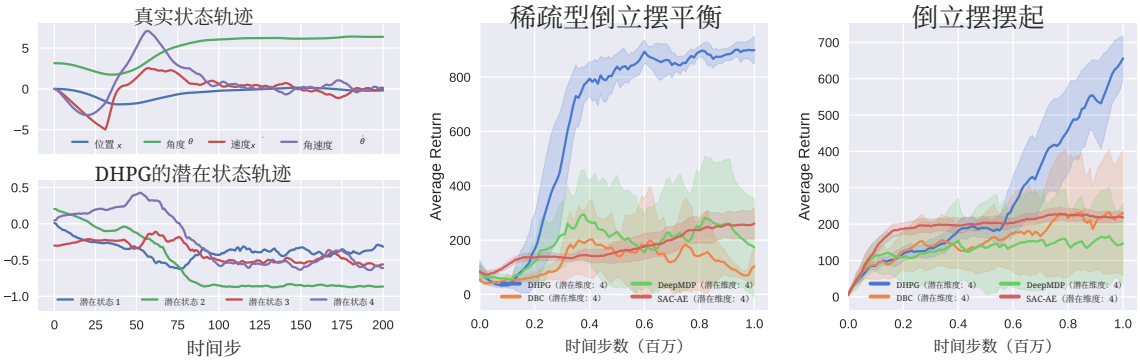
与 Zhang 等人（2020）类似，我们用 W_2度量替代了 Kantorovich (W_1) 度量，因为对于高斯分布，前者有显式公式。最后，我们通过以下损失函数应用定义 11 中连续 MDP 同态映射的条件：

$$\mathcal{L}_{\text{h}}(\phi,\eta,\nu,\rho) = \mathbb{E}_{(s_i,a_i,s_i',r_i)\sim\mathcal{B}}\Big[\big(f_\phi(s_i') - \overline{s}_i'\big)^2 + \big(r_i - \overline{R}_\rho(f_\phi(s_i))\big)^2\Big], \tag{34}$$

其中 $\overline{s}_i' \sim \overline{\tau}_\nu(\cdot|f_\phi(s_i)\, g_\eta(s_i,a_i))$。最终的损失函数为 $\mathrm{L}_{\text{lax}}(\phi,\eta)$+$\mathrm{L}_{\text{h}}(\phi,\eta,\rho,\nu)$。

7 实验

在我们的实验中，我们旨在回答以下关键问题：



(a) 随时间变化的状态轨迹。 (b) 学习曲线。 (c) 学习曲线。

图4：DHPG 从像素中恢复最小MDP的效果。所有方法的潜在空间均限制为4维，与倒立摆真实状态空间的维数相同。(a) 从 Mujoco 获得的真实状态轨迹和 DHPG 的潜在状态轨迹。(b, c) 对10个随机种子取平均的学习曲线。

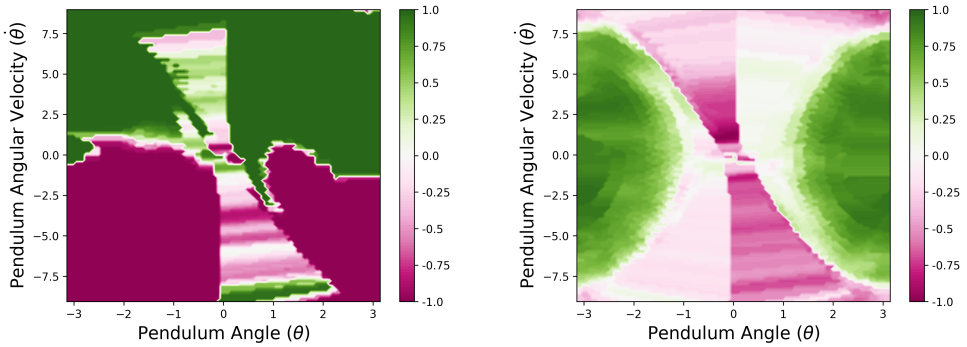
1. 同态策略梯度能否改善策略优化？2. 学习到的表征和抽象MDP具有哪些定性属性?
3. DHPG能否从原始像素观测中学习并恢复最小MDP像?

我们在 DM Control 的像素观测连续控制任务上评估 DHPG。值得注意的是，为了可靠地评估我们的算法与基线方法的性能，并准确反映结果的分布情况，我们遵循 Agarwal 等人（2021）提出的最佳实践，报告了在所有任务上汇总的四分位均值（IQM）和性能剖面，统计结果基于 10 个随机种子。虽然我们的基线结果是通过官方代码获得的，但在可能的情况下[5]，，部分结果可能与原始报告有所不同，这是由于随机种子数量存在差异，同时我们旨在忠实呈现真实的性能分布（Agarwal 等人，2021）。

7.1 DeepMind控制套件

我们将 DHPG 在像素观测上的有效性与 DBC（Zhang等人，2020）、DeepMDP（Gelada 等人，2019）、SAC-AE（Yarats 等人，2021b）以及当前表现最先进的 DrQ-v2（Yarats 等人，2021a）进行比较。所有方法均采用 n 步回报，共享附录D.1中的相同超参数，且所有超参数均基于 DrQ-v2，未进行任何进一步调优。我们承认，由于 DrQ-v2 基于 DDPG，因此与随机 DHPG 相比，我们使用的超参数可能更有利于确定性 DHPG。重要的是，为了与采用图像增强的 DrQ-v2 进行公平比较，我们展示了 DHPG 和其他基线方法的两种变体，分别采用和不采用图像增强。

5. 我们使用 DrQv2、DBC 和 SAC-AE 的官方实现；由于官方代码不可用，我们重新实现了 DeepMDP。完整详情请参见附录 D.2。



(a) 实际最优策略。 (b) 抽象最优策略。

图5：摆杆摆起任务的状态空间中实际最优动作和抽象最优动作的等高线。颜色表示动作价值，状态为 $s = (\theta, \dot{\theta})$。(a)实际最优策略；最优动作的等高线 $a^* = \pi^{\uparrow *}(s)$。(b) 抽象最优策略；抽象最优动作的等高线 $a^* = g_s(\underline{a}^*) = \pi^*(s)$。对于等价状态-动作对，关系 $g_{s_1}(a_1) = g_{s_2}(a_2)$ 成立，且抽象最优策略具有对称性。

DHPG 在像素观测上优于或媲美其他算法，展现了其在表征学习方面的有效性。 图像增强的结果见图3，且完整结果见附录 *C.1*。在14项任务上的汇总结果显示，确定性DHPG优于最先进的DrQ-v2，而随机 DHPG的表现与其相当。尽管总体而言这些方法带来的性能提升有限，但DHPG与DrQ-v2之间的性能对比高度依赖于具体任务（见附录 C.1 中的图12）。对于 Walker Run 或 Cheetah Run 等复杂任务，DHPG的表现与DrQ-v2相当或略差；然而，在具有明显对称性、因而易于学习MDP同态映射的领域中，DHPG优于DrQ-v2。特别是，在 Cartpole 和摆杆等领域中，不使用图像增强的DHPG优于DrQ-v2，展现了其表征学习能力。

确定性DHPG和随机 DHPG的样本效率大致相当，其中确定性DHPG略胜一筹。如第5.3节所述，从理论上讲，确定性策略梯度比随机策略梯度具有更高的样本效率，因为它无需对动作空间进行积分（Silver 等人，2014）。此外，由于提升随机策略存在复杂性，随机DHPG包含更多组件，可能会对学习性能产生负面影响。因此，确定性DHPG的样本效率略高于随机 DHPG。

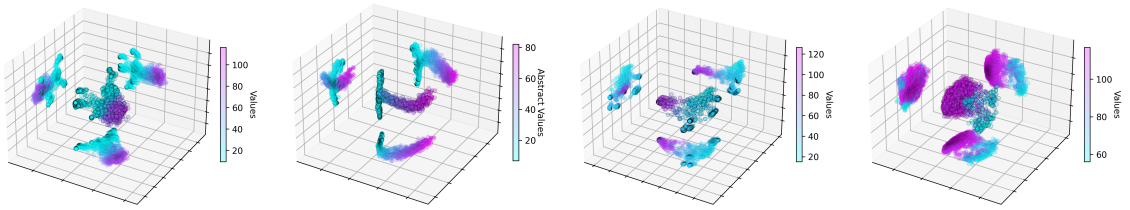
DHPG 能够学习并恢复低维 MDP 像。MDP 同态的一项关键优势是能够表示最小 MDP 像（Ravindran和Barto，2001年），这在从像素观测中学习时尤为重要。为展示这一能力，我们将潜在空间的维度限制为真实系统的维度，并在图4中将 DHPG（不使用图像增强）与基线方法进行了比较。其他方法无法学习这些任务，而 DHPG 则能够成功学习策略和最小低维潜在空间。令人惊讶的是，如图4a所示，潜在状态的轨迹与真实状态的轨迹相似。

学习得到的映射 $h=(f, g_s)$ 呈现出 MDP 同态的性质。我们使用摆锤摆起任务来可视化其学习得到的 MDP 同态，因为该任务的对称性一目了然。如果 $a_1 = -a_2$、$\theta_1 = -\theta_2$ 且 $\dot{\theta}_1 = -\dot{\theta}_2$，则两个状态-动作对（$s_1 = (\theta_1, \dot{\theta}_1)\ a_1$）和（$s_2 = (\theta_2, \dot{\theta}_2)\ a_2$）等价。因此，学习得到的动作表示应通过令 $g_{s_1}(a_1) = g_{s_2}(a_2)$ 来体现这一点。图5a展示了状态空间 S 上最优动作的等高线，图5b则展示了状态空间 S 上最优动作的动作表示 $\bar{a} = g_s(a)$。显然，抽象动作遵循上述等价状态-动作对之间的关系，这表明 $g_s(a)$实际上表示了 MDP 同态映射中的动作编码器。

抽象MDP展示了MDP同态像的性质。为了定性地展示学习联合状态-动作表示的重要性，图6展示了四足行走任务中潜在状态的可视化结果。该任务具有围绕四条腿运动的对称性。有趣的是，尽管 DHPG 的潜在空间（图6a）为每条腿都呈现了不同的状态，但抽象状态编码器 f_ϕ 已将对应的腿（例如，左前腿和右后腿）映射到相同的抽象潜在状态（图6b），因为它们互为某种同态像。显然，DBC 和 DrQ-v2 无法实现这一点。

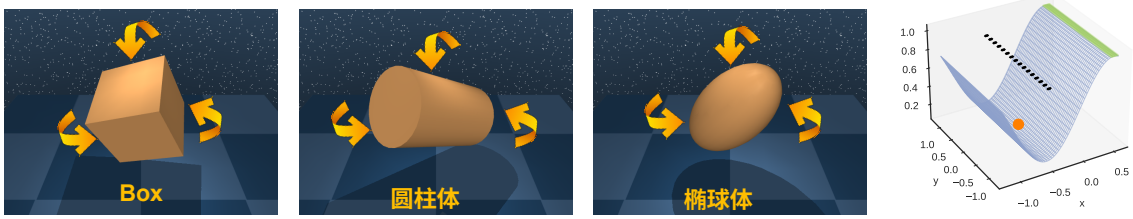
学习到的表征和MDP 同态映射可迁移到同一领域中的新任务。重要的是，依赖奖励的表征学习方法需要考虑学习到的表征在同一领域的新奖励设定下的可迁移性。为确保我们的方法不会妨碍这种迁移，我们进行了实验，将演员、评论家和学习得到的MDP同态映射迁移到同一领域的另一项任务。附录C.3中的结果表明，我们的方法并未损害迁移能力。

更多实验。 我们在附录C.2中研究价值等价性质，将其作为衡量学习得到的MDP同态质量的指标；此外，我们还在附录



（a）DHPG 潜在状态。 (b) DHPG 抽象状态。 (c) DBC 潜在状态。 (d) DrQ-v2 潜在状态。

图 6：基于像素观测的四足行走学习到的表征的主成分分析投影。**(a)** 潜在状态 $s = E_\mu(o)$，**(b)** DHPG 的抽象潜在状态 $\bar{s} = f_\phi(E_\mu(o))$，**(c)** DBC 的潜在状态 $\bar{s} = E_\mu(o)$，以及 **(d)** DrQ-v2。每个点的颜色表示通过 $Q(s,a)$ 或 $Q(\bar{s},a)$ 学得的其对应数值。点也被投影到各个主平面上。DHPG 的同态映射将对应腿部（例如左前腿和右后腿）的潜在状态 **(a)** 映射到相同的抽象潜在状态 **(b)**，这表明 S 中存在清晰的结构。



(a) Rotate Suite。 (b) 三维山地车。

图7：具有对称性的新型环境。(a) Rotate Suite 是一系列视觉控制任务，目标是沿着三维物体的各轴旋转它，使其达到目标朝向。环境的对称性由物体的对称性决定。(b) 三维山地车沿 y 轴具有连续平移对称性，如黑色虚线所示。橙色点表示汽车，绿色线表示目标位置。

C.4和C.5中分别对DHPG的不同变体进行了消融研究，并分析了{v3}-步回报对我们方法的影响。最后，我们在附录C.6中将我们的方法与基线方法的计算时间进行了比较。

7.2 具有连续对称性的环境

如第5.3节所述，确定性与随机 HPG 定理的关键区别在于，前者要求使用双射动作编码器，而后者对此不作任何结构限制。这一要求意味着，确定性DHPG无法对动作进行抽象，除了重新标记动作之外。对于离散对称性，重新标记动作已足够；而对于具有连续对称性的环境，原则上可以减少其动作维度。为展示随机 DHPG 在动作抽象方面的优势，我们在一系列具有连续对称性的新环境中开展了实验。这些环境已公开提供[6]。

- **Rotate Suite**是一系列基于 DeepMind 控制套件开发的视觉控制任务。在每个环境中，目标都是沿着三维物体的各轴旋转物体，使其达到目标朝向。环境的对称性由物体的对称性决定。因此，方块旋转具有离散对称性，而圆柱体和椭球体的旋转具有连续旋转对称性。图7a展示了这些交互式环境的示例。

- **3D 山地车** 是二维山地车问题（Moore，1990）的扩展版本，其中山地曲线沿 y 轴延伸，形成山地表面。智能体在山地表面上拥有二维动作，而二维山地车的动作空间则是一维。由于这一扩展，该问题沿 y 轴具有连续平移对称性，沿该轴方向的动作是冗余的。图7b展示了三维山地车的表面。

在具有连续对称性的环境中，随机 DHPG 优于确定性DHPG。 结果见图 8 和图 9。值得注意的是，在具有

6. https://github.com/sahandrez/rotate suite

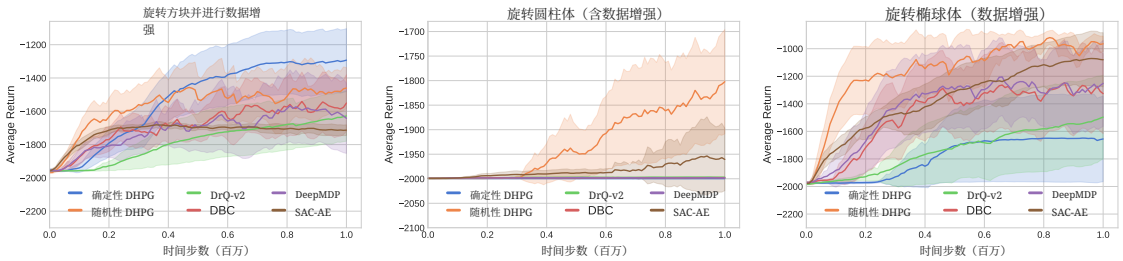


图8：在 10 个随机种子上获得的 Rotate Suite 环境学习曲线。（**a**）方块旋转。（**b**）圆柱体旋转。（**c**）椭球体旋转。阴影区域表示 95% 置信区间。在具有连续对称性的环境中（圆柱体旋转和椭球体旋转），随机 DHPG 的表现优于确定性DHPG。

连续对称性的环境中，随机 DHPG 优于其他基线方法，也优于确定性DHPG。这是因为从理论上讲，随机 DHPG 不会对动作编码器施加任何结构，因此与确定性DHPG相比，能够实现更高程度的动作抽象。

DHPG能够学习出符合三维旋转性质的结构化潜在空间。图10展示了圆柱旋转任务中的潜在状态轨迹可视化结果。每条轨迹均按动作维度着色，通过围绕一个主轴（俯仰、滚转或偏航）施加恒定旋转来采集。有趣的是，DHPG的潜在轨迹完全解耦，并且在潜在空间中呈现出类似三维旋转的形态。然而，其他基线方法的潜在表示均未呈现出这种结构。

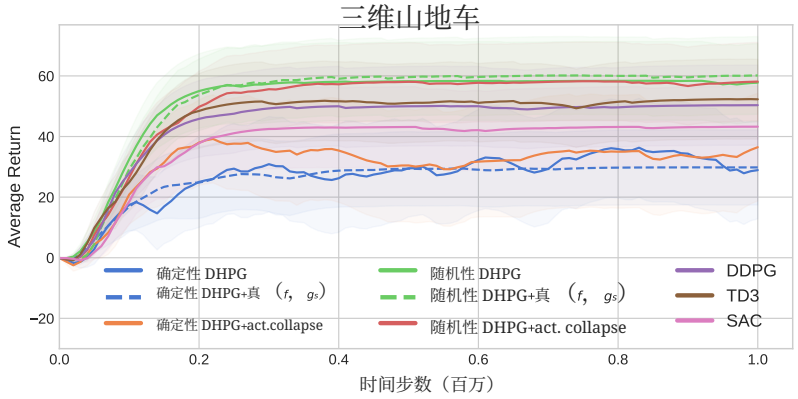
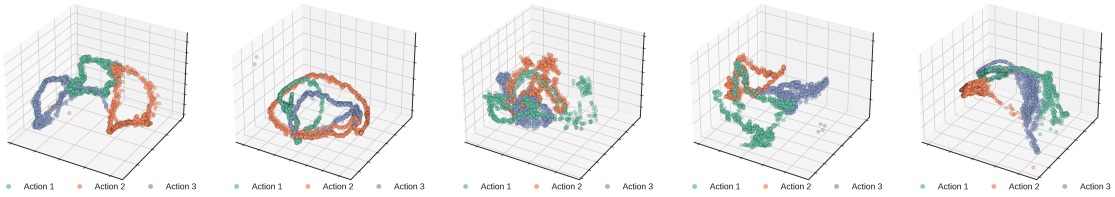


图9：在 50个随机种子上获得的三维山地车环境学习曲线。阴影区域表示 95% 置信区间。由于环境具有连续平移对称性，随机 DHPG 的表现优于确定性DHPG。DHPG+true (f, g_s) 使用真实同态映射，而 DHPG+act. collapse 移除了动作空间，具有一维抽象动作空间，而非原始 MDP 的二维动作空间。



(a) 随机性 DHPG。 (b) 确定性 DHPG。 (c) DrQ-v2。 (d) DBC. (e) DeepMDP。

图10：圆柱旋转任务中的潜在状态轨迹可视化，按动作维度进行颜色编码。动作维度分别对应俯仰、滚转和偏航，每条轨迹均通过绕特定轴施加恒定旋转采集。潜在状态使用主成分分析投影到三维空间。

8 结论

本文利用测度论建立了连续 MDP 同态的新理论，并严格证明了其价值等价性和最优价值等价性。为了直接利用联合状态-动作抽象进行策略优化，我们推导了适用于随机策略和确定性策略的同态策略梯度。重要的是，我们严格证明了，在抽象MDP上应用同态策略梯度，等价于在实际MDP上应用标准策略梯度。基于这些新的理论结果，我们开发了一系列深度演员-评论家算法，这些算法采用随机策略或确定性策略，并能够利用宽松双模拟度量同时学习策略和MDP 同态映射。我们将该算法称为深度同态策略梯度（DHPG）；它在具有挑战性的视觉控制问题中优于强大的基线方法。潜在空间的可视化表明，MDP同态在学习能够保留价值函数的结构化表示方面具有巨大潜力。最后，我们引入了一系列具有连续对称性的环境，进一步展示了我们的算法在连续对称性下进行动作抽象的能力。

我们相信，我们的工作将为马尔可夫决策过程同态在具有挑战性的连续控制问题中的应用开辟新的可能性，使其他强化学习算法也能受益于马尔可夫决策过程同态的抽象能力和同态策略梯度定理。

致谢与经费披露

SRS获得NSERC CGS-D奖学金的支持。完成本研究时，RZ获得了NSERC CGS-M奖学金的支持。PP获得NSERC研究经费的支持。本研究所用的计算资源由Calcul Quebec和加拿大数字研究联盟提供。

参考文献

David Abel、David Hershkowitz 和 Michael Littman。通过近似状态抽象实现近乎最优的行为。载于国际机器学习会议，第 2915–2923 页。PMLR，2016年。David Abel、Dilip Arumugam、Kavosh Asadi、Yuu Jinnai、Michael L Littman 和 Lawson LS Wong。状态抽象作为学徒学习中的压缩。载于*AAAI*人工智能会议论文集，第 33 卷，第 3134–3142 页，2019年。David Abel、Nate Umbanhowar、Khimya Khetarpal、Dilip Arumugam、Doina Precup 和 Michael Littman。价值保持型状态-动作抽象。载于国际人工智能与统计会议，第 1639–1650 页。PMLR，2020年。Rishabh Agarwal、Marlos C Machado、Pablo Samuel Castro 和 Marc G Bellemare。用于强化学习泛化的对比式行为相似性嵌入。载于国际学习表征会议，2020年。Rishabh Agarwal、Max Schwarzer、Pablo Samuel Castro、Aaron C Courville 和 Marc Bellemare。深度强化学习处于统计悬崖边缘。神经信息处理系统进展，34，2021年。Ankesh Anand、Evan Racah、Sherjil Ozair、Yoshua Bengio、Marc-Alexandre Cˆot´e 和 R Devon Hjelm。Atari中的无监督状态表征学习。神经信息处理系统进展，32：8769–8782，2019年。Kai Arulkumaran、Marc Peter Deisenroth、Miles Brundage 和 Anil Anthony Bharath。深度强化学习：简要综述。*IEEE*信号处理杂志，34(6)：26–38，2017年。Aharon Azulay 和 Yair Weiss。为什么深度卷积网络对微小图像变换的泛化能力如此之差？*arXiv*预印本 *arXiv:1805.12177*，2018年。Jimmy Lei Ba、Jamie Ryan Kiros 和 Geoffrey E Hinton。层归一化。*arXiv*预印本 *arXiv:1607.06450*，2016年。Gabriel Barth-Maron、Matthew W Hoffman、David Budden、Will Dabney、Dan Horgan、Dhruva Tb、Alistair Muldal、Nicolas Heess 和 Timothy Lillicrap。分布式分布式确定性策略梯度。*arXiv*预印本 *arXiv:1804.08617*，2018年。Simon Batzner、Albert Musaelian、Lixin Sun、Mario Geiger、Jonathan P Mailoa、Mordechai Kornbluth、Nicola Molinari、Tess E Smidt 和 Boris Kozinsky。用于数据高效且精准的原子间势的 E (3) 等变图神经网络。《自然通讯》，13(1)：1–11，2022年。O Biza、R Platt、JW van de Meent 和 L Wong。通过深度变分推断学习离散状态抽象。近似贝叶斯推断进展，2021年。Ondrej Biza 和 Robert Platt。用于深度学习的基于马尔可夫决策过程同态的在线抽象。载于第*18*届国际自主智能体与多智能体系统会议论文集，2019年。

George W Bluman 和 Sukeyuki Kumei。《对称性与微分方程》，第 81 卷。Springer Science & Business Media，2013年。R. Blute、J. Desharnais、A. Edalat 和 P. Panangaden。带标记的马尔可夫过程的双模拟。载于第十二届 *IEEE* 计算机科学逻辑研讨会论文集，波兰华沙。，1997a。Richard Blute、Jos´ee Desharnais、Abbas Edalat 和 Prakash Panangaden。带标记的马尔可夫过程的双模拟。载于第十二届 *IEEE* 计算机科学逻辑年会论文集，第 149–158 页。IEEE，1997b。Vladimir Igorevich Bogachev 和 Maria Aparecida Soares Ruas。《测度论》，第 1 卷。Springer，2007 年。T Butler 、J Jakeman 和 Tim Wildey。结合推前测度与贝叶斯规则，为随机反问题构造一致解。《*SIAM* 科学计算杂志》，40(2):A984–A1011，2018年。Hugo Caselles-Dupr´e、Michael Garcia Ortiz 和 David Filliat。基于对称性的解耦表征学习需要与环境交互。《神经信息处理系统进展》，32:4606–4615，2019年。Pablo Samuel Castro。用于计算确定性马尔可夫决策过程中状态相似性的可扩展方法。载于*AAAI*人工智能会议论文集，第 34 卷，第 10069–10076 页，2020 年。Pablo Samuel Castro 和 Doina Precup。利用双模拟进行马尔可夫决策过程中的策略迁移。载于第二十四届 *AAAI* 人工智能会议，2010年。Pablo Samuel Castro 和 Doina Precup。利用双模拟度量自动构造马尔可夫决策过程的时间扩展动作。载于欧洲强化学习研讨会，第 140–152 页。Springer，2011年。Pablo Samuel Castro、Tyler Kastner、Prakash Panangaden 和 Mark Rowland。Mico：通过基于采样的马尔可夫决策过程状态相似性改进表征。神经信息处理系统进展，34，2021年。Yash Chandak、Georgios Theocharous、James Kostas、Scott Jordan 和 Philip Thomas。学习用于强化学习的动作表示。载于国际机器学习会议，第 941–950 页。PMLR，2019年。Haiwei Chen、Shichen Liu、Weikai Chen、Hao Li 和 Randall Hill。用于三维点云分析的等变点网络。载于*IEEE/CVF* 计算机视觉与模式识别会议论文集，第 14514–14523 页，2021年。Taco Cohen 和 Max Welling。群等变卷积网络。载于国际机器学习会议，第 2990–2999 页。PMLR，2016。Taco Cohen、Maurice Weiler、Berkay Kicanaoglu 和 Max Welling。规范等变卷积网络与二十面体卷积神经网络。载于国际机器学习会议，第 1321–1330 页。PMLR，2019年。

Taco S Cohen、Mario Geiger、Jonas K¨ohler 和 Max Welling。球面卷积神经网络。载于国际学习表征会议，2018年。Miles Cranmer、Sam Greydanus、Stephan Hoyer、Peter Battaglia、David Spergel 和 Shirley Ho。拉格朗日神经网络。*arXiv* 预印本 *arXiv:2003.04630*，2020年。Nima Dehmamy、Robin Walters、Yanchen Liu、Dashun Wang 和 Rose Yu。利用李代数卷积网络自动发现对称性。神经信息处理系统进展，34：2503–2515，2021年。J. Desharnais、V. Gupta、R. Jagadeesan 和 P. Panangaden。带标记的马尔可夫系统的度量。载于*CONCUR99* 论文集，《计算机科学讲义》第1664卷。Springer-Verlag，1999年。J. Desharnais、A. Edalat 和 P. Panangaden。带标记的马尔可夫过程的双模拟。信息与计算，179(2)：163–193，2002年12月。Gabriel Dulac-Arnold、Richard Evans、Hado van Hasselt、Peter Sunehag、Timothy Lil- licrap、Jonathan Hunt、Timothy Mann、Theophane Weber、Thomas Degris 和 Ben Coppin。大规模离散动作空间中的深度强化学习。*arXiv* 预印本 *arXiv:1512.07679*，2015年。Stephan Eismann、Raphael JL Townshend、Nathaniel Thomas、Milind Jagota、Bowen Jing 和 Ron O Dror。用于选择蛋白质复合物结构模型的分层旋转等变神经网络。《蛋白质：结构、功能与生物信息学》，89(5)：493–501，2021年。Logan Engstrom、Brandon Tran、Dimitris Tsipras、Ludwig Schmidt 和 Aleksander Madry。一种旋转加一种平移便已足够：用简单变换欺骗卷积神经网络。2017年。SM Ali Eslami、Danilo Jimenez Rezende、Frederic Besse、Fabio Viola、Ari S Morcos、Marta Garnelo、Avraham Ruderman、Andrei A Rusu、Ivo Danihelka、Karol Gregor 等。神经场景表征与渲染。《科学》，360(6394)：1204–1210，2018年。Linxi Fan、Guanzhi Wang、De-An Huang、Zhiding Yu、Li Fei-Fei、Yuke Zhu 和 Animashree Anandkumar。Secant：通过自专家克隆实现视觉策略的零样本泛化。载于国际机器学习会议，第3088–3099页。PMLR，2021年。Norm Ferns、Prakash Panangaden 和 Doina Precup。有限马尔可夫决策过程的度量。载于第二十届人工智能中的不确定性会议论文集，第162–169页，2004年7月。Norm Ferns、Prakash Panangaden 和 Doina Precup。无限状态空间马尔可夫决策过程的度量。载于第二十一届人工智能中的不确定性会议论文集，第201–208页，2005a年。Norm Ferns、Prakash Panangaden 和 Doina Precup。无限状态空间马尔可夫决策过程的度量。载于第二十一届人工智能中的不确定性会议论文集，第201–208页，2005b年。

Norm Ferns、Pablo Samuel Castro、Doina Precup 和 Prakash Panangaden。马尔可夫决策过程中状态相似性的计算方法。载于第二十二届人工智能中的不确定性会议论文集，第 174–181 页，2006年。Norm Ferns、Prakash Panangaden 和 Doina Precup。连续马尔可夫决策过程的双模拟度量。《*SIAM*计算杂志》，40(6):1662–1714，2011年。Marc Finzi、Max Welling 和 Andrew Gordon Wilson。为任意矩阵群构造等变多层感知机的实用方法。载于国际机器学习会议，第 3318–3328 页。PMLR，2021年。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年。Octavian-Eugen Ganea、Xinyuan Huang、Charlotte Bunne、Yatao Bian、Regina Barzilay、Tommi S Jaakkola 和 Andreas Krause。用于端到端刚性蛋白质对接的独立 SE(3) 等变模型。载于国际学习表征会议，2021年。Carles Gelada、Saurabh Kumar、Jacob Buckman、Ofir Nachum 和 Marc G Bellemare。DeepMDP：学习用于表征学习的连续潜在空间模型。载于国际机器学习会议，第 2170–2179 页。PMLR，2019年。Robert Givan、Thomas Dean 和 Matthew Greig。马尔可夫决策过程中的等价概念与模型最小化。《人工智能》，147(1-2):163–223，2003。Samuel Greydanus、Misko Dzamba 和 Jason Yosinski。哈密顿神经网络。神经信息处理系统进展，32，2019。Christopher Grimm、Andr´e Barreto、Satinder Singh 和 David Silver。基于模型的强化学习的价值等价原理。神经信息处理系统进展，33:5541–5552，2020年。Christopher Grimm、Andr´e Barreto、Greg Farquhar、David Silver 和 Satinder Singh。适当价值等价。神经信息处理系统进展，34:7773–7786，2021年。David Ha 和 J¨urgen Schmidhuber。世界模型。*arXiv* 电子预印本，第 arXiv–1803 页，2018年。Tuomas Haarnoja、Aurick Zhou、Pieter Abbeel 和 Sergey Levine。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。载于国际机器学习会议，第 1861–1870 页。PMLR，2018年。Danijar Hafner、Timothy Lillicrap、Jimmy Ba 和 Mohammad Norouzi。梦境控制：通过潜在想象学习行为。载于国际学习表征会议，2019a。Danijar Hafner、Timothy Lillicrap、Ian Fischer、Ruben Villegas、David Ha、Honglak Lee 和 James Davidson。学习用于基于像素规划的潜在动力学。载于国际机器学习会议，第 2555–2565 页。PMLR，2019b。

Nicklas Hansen 和 Xiaolong Wang。通过软数据增强实现强化学习中的泛化。载于 *2021*年*IEEE*机器人与自动化国际会议 *(ICRA)*，第 13611–13617 页。IEEE，2018年。Nicklas Hansen、Rishabh Jangir、Yu Sun、Guillem Aleny`a、Pieter Abbeel、Alexei A Efros、Lerrel Pinto 和 Xiaolong Wang。部署期间的自监督策略适应。载于国际学习表征会议，2020年。Peter Henderson、Riashat Islam、Philip Bachman、Joelle Pineau、Doina Precup 和 David Meger。值得关注的深度强化学习。载于*AAAI*人工智能会议论文集，第 32 卷，2018年。Irina Higgins、David Amos、David Pfau、Sebastien Racaniere、Loic Matthey、Danilo Rezende 和 Alexander Lerchner。迈向解耦表征的定义。*arXiv*预印本 *arXiv:1812.02230*，2018年。

Irina Higgins、Peter Wirnsberger、Andrew Jaegle 和 Aleksandar Botev。Symetric：衡量从视觉推断出的习得哈密顿动力学的质量。神经信息处理系统进展，34，2021年。

R Devon Hjelm、Alex Fedorov、Samuel Lavoie-Marchildon、Karan Grewal、Phil Bachman、Adam Trischler 和 Yoshua Bengio。通过互信息估计与最大化学习深层表征。载于国际表征学习会议，2018年。

Max Jaderberg、Volodymyr Mnih、Wojciech Marian Czarnecki、Tom Schaul、Joel Z Leibo、David Silver 和 Koray Kavukcuoglu。利用无监督辅助任务进行强化学习。*arXiv*预印本 *arXiv:1611.05397*，2016。Rico Jonschkowski 和 Oliver Brock。利用机器人先验学习状态表征。《自主机器人》，39(3):407–428，2015。John Jumper、Richard Evans、Alexander Pritzel、Tim Green、Michael Figurnov、Olaf Ron-ˇneberger、Kathryn Tunyasuvunakool、Russ Bates、Augustin Z´ıdek、Anna Potapenko 等。利用 AlphaFold 实现高精度蛋白质结构预测。《自然》，596(7873):583–589，2021。Jari Kaipio 和 Erkki Somersalo。统计与计算逆问题，第160卷。Springer Science & Business Media，2006。Mete Kemertas 和 Tristan Aumentado-Armstrong。迈向稳健的双模拟度量学习。神经信息处理系统进展，34，2021。Mete Kemertas 和 Allan Douglas Jepson。基于双模拟度量的近似策略迭代。《机器学习研究汇刊》，2022。B Ravi Kiran、Ibrahim Sobh、Victor Talpaert、Patrick Mannion、Ahmad A Al Sallab、Senthil Yogamani 和 Patrick P´erez。用于自动驾驶的深度强化学习：综述。*IEEE*智能交通系统汇刊，23(6):4909–4926，2021。

Risi Kondor、Zhen Lin 和 Shubhendu Trivedi。Clebsch–Gordan 网络：一种完全基于傅里叶空间的球面卷积神经网络。神经信息处理系统进展，第31卷，2018年。Serge Lang。微分流形与黎曼流形，第160卷。Springer Science & Business Media，2012年。K. G. Larsen 和 A. Skou。通过概率测试实现双模拟。信息与计算，94：1–28，1991a。Kim G Larsen 和 Arne Skou。通过概率测试实现双模拟。信息与计算，94(1)：1–28，1991b。Charline Le Lan、Marc G Bellemare 和 Pablo Samuel Castro。强化学习中的度量与连续性。载于*AAAI*人工智能会议论文集，第35卷，第8261–8269页，2021年。Yann LeCun、Bernhard Boser、John Denker、Donnie Henderson、Richard Howard、Wayne Hubbard 和 Lawrence Jackel。利用反向传播网络识别手写数字。神经信息处理系统进展，第2卷，1989年。Yann LeCun、Yoshua Bengio 等。用于图像、语音和时间序列的卷积网络。脑理论与神经网络手册，3361(10)：1995，1995年。Kimin Lee、Kibok Lee、Jinwoo Shin 和 Honglak Lee。网络随机化：一种用于深度强化学习泛化的简单技术。*arXiv*预印本 *arXiv:1910.05396*，2019年。Lihong Li、Thomas J Walsh 和 Michael L Littman。迈向马尔可夫决策过程状态抽象的统一理论。*ISAIM*，4：5，2006年。Timothy P Lillicrap、Jonathan J Hunt、Alexander Pritzel、Nicolas Heess、Tom Erez、Yuval Tassa、David Silver 和 Daan Wierstra。利用深度强化学习实现连续控制。*arXiv*预印本 *arXiv:1509.02971*，2015年。Hao Liu 和 Pieter Abbeel。APS：基于后继特征的主动预训练。载于国际机器学习会议，第6736–6747页。PMLR，2021年。Shikun Liu、Andrew Davison 和 Edward Johns。通过元辅助学习实现自监督泛化。神经信息处理系统进展，第32卷，2019年。Clare Lyle、Mark Rowland、Georg Ostrovski 和 Will Dabney。辅助任务对表征动态的影响。载于国际人工智能与统计会议，第1–9页。PMLR，2021年。Anuj Mahajan 和 Theja Tulabandhula。强化学习中函数逼近的对称性学习。*arXiv*预印本 *arXiv:1706.02999*，2017年。R. Milner。通信系统演算，计算机科学讲义第92卷。Springer-Verlag，1980年。

Robin Milner。《通信与并发》。Prentice-Hall，1989a。Robin Milner。《通信与并发》，第84卷。Prentice hall Englewood Cliffs，1989b。Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Alex Graves、Ioannis Antonoglou、Daan Wierstra 和 Martin Riedmiller。使用深度强化学习玩 Atari。*arXiv* 预印本 *arXiv:1312.5602*，2013年。Arnab Kumar Mondal、Vineet Jain、Kaleem Siddiqi 和 Siamak Ravanbakhsh。Eqr：面向数据高效强化学习的等变表征。载于国际机器学习会议，第15908–15926页。PMLR，2022年。Andrew William Moore。用于机器人控制的高效基于记忆的学习。技术报告，剑桥大学计算机实验室，1990年。Emmy Noether。不变变分问题。《输运理论与统计物理》，1（3）：186–207，1971年。1918年发表的原德文论文的英文译本。David Park。标题未知。巴特洪内夫并发语义研讨会幻灯片，1981年。Jung Yeon Park、Ondrej Biza、Linfeng Zhao、Jan Willem van de Meent 和 Robin Walters。为等变世界模型学习对称嵌入。*arXiv* 预印本 *arXiv:2204.11371*，2022年。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，32：8026–8037，2019年。Zhuoran Qiao、Anders S Christensen、Matthew Welborn、Frederick R Manby、Anima Anandkumar 和 Thomas F Miller III。Unite：用于量子化学的幺正多体张量等变网络。*arXiv* 预印本 *arXiv:2105.14655*，2021年。Robin Quessard、Thomas D Barrett 和 William R Clements。学习动力学环境的群结构与解耦表征。*arXiv* 预印本 *arXiv:2002.06991*，2020年。Srividhya Rajendran 和 Manfred Huber。使用近似部分策略同态学习技能的泛化与复用。载于*2009 IEEE*系统、人类与控制论国际会议，第2239–2244页。IEEE，2009年。Balaraman Ravindran。《强化学习中抽象的代数方法》。马萨诸塞大学阿默斯特分校，2004年。Balaraman Ravindran 和 Andrew G Barto。马尔可夫决策过程中的对称性与模型最小化，2001年。

Balaraman Ravindran 和 Andrew G Barto。相对化选项：选择正确的变换。载于第*20*届国际机器学习会议论文集（*ICML-03*），第608–615页，2003年。Balaraman Ravindran 和 Andrew G Barto。近似同态：马尔可夫决策过程非精确最小化的框架，2004年。Sahand Rezaei-Shoshtari、Rosie Zhao、Prakash Panangaden、David Meger 和 Doina Precup。连续 MDP 同态与同态策略梯度。载于神经信息处理系统进展，2022年。Brian Sallans 和 Geoffrey E Hinton。基于因子化状态和动作的强化学习。机器学习研究杂志，5：1063–1088，2004年。Vıctor Garcia Satorras、Emiel Hoogeboom 和 Max Welling。E(n) 等变图神经网络。载于国际机器学习会议，第9323–9332页。PMLR，2021年。Sahil Sharma、Aravind Suresh、Rahul Ramesh 和 Balaraman Ravindran。学习因子化策略和动作价值函数：用于深度强化学习的因子化动作空间表示。*arXiv*预印本 *arXiv:1705.07269*，2017年。David Silver、Guy Lever、Nicolas Heess、Thomas Degris、Daan Wierstra 和 Martin Riedmiller。确定性策略梯度算法。载于国际机器学习会议，第387–395页。PMLR，2014年。Samarth Sinha、Ajay Mandlekar 和 Animesh Garg。S4RL：用于机器人学离线强化学习的出奇简单的自监督方法。载于第五届机器人学习年会，2021年。

Meir Smorodinsky。《style id='1'>遍历理论熵，第214卷。Springer，2006年。

Vishal Soni 和 Satinder Singh。利用同态在连续强化学习领域之间迁移选项。载于第*21*届全国人工智能会议论文集-第*1*卷，第494–499页，2006年。Jonathan Sorg 和 Satinder Singh。通过软同态进行迁移。载于第*8*届国际自主智能体与多智能体系统会议论文集-第*2*卷，第741–748页，2009年。Michael Spivak。流形上的微积分：高等微积分经典定理的现代方法。CRC出版社，2018年。Adam Stooke、Kimin Lee、Pieter Abbeel 和 Michael Laskin。将表征学习与强化学习解耦。载于国际机器学习会议，第9870–9879页。PMLR，2021年。Richard S Sutton 和 Andrew G Barto。强化学习：入门。麻省理工学院出版社，2018年。

Richard S Sutton、Doina Precup 和 Satinder Singh。介于MDP与半MDP之间：强化学习中时间抽象的框架。《人工智能》，112 (1-2):181–211，1999年。Richard S Sutton、David A McAllester、Satinder P Singh 和 Yishay Mansour。具有函数逼近的强化学习策略梯度方法。载于神经信息处理系统进展，第1057–1063页，2000年。Jonathan Taylor、Doina Precup 和 Prakash Panagaden。近似MDP同态中的性能损失界。神经信息处理系统进展，21，2008年。Guy Tennenholtz 和 Shie Mannor。动作的自然语言。载于国际机器学习会议，第6196–6205页。PMLR，2019年。Nathaniel Thomas、Tess Smidt、Steven Kearnes、Lusann Yang、Li Li、Kai Kohlhoff 和 Patrick Riley。张量场网络：适用于三维点云的旋转和平移等变神经网络。*arXiv*预印本 *arXiv:1802.08219*，2018年。Valentin Thomas、Jules Pondard、Emmanuel Bengio、Marc Sarfati、Philippe Beaudoin、Marie-Jean Meurs、Joelle Pineau、Doina Precup 和 Yoshua Bengio。独立可控因素。*arXiv*预印本 *arXiv:1708.01289*，2017年。Emanuel Todorov、Tom Erez 和 Yuval Tassa。Mujoco：用于基于模型控制的物理引擎。载于*2012*年*IEEE/RSJ*智能机器人与系统国际会议，第5026–5033页。IEEE，2012年。Elise van der Pol、Thomas Kipf、Frans A Oliehoek 和 Max Welling。对MDP同态的可规划近似：动作等变性。载于第*19*届国际自主智能体与多智能体系统会议论文集，第1431–1439页，2020a年。Elise van der Pol、Daniel Worrall、Herke van Hoof、Frans Oliehoek 和 Max Welling。马尔可夫决策过程同态网络：强化学习中的群对称性。神经信息处理系统进展，33，2020b年。Elise van der Pol、Herke van Hoof、Frans A Oliehoek 和 Max Welling。多智能体马尔可夫决策过程同态网络。载于国际学习表征会议，2021年。Veeravalli S Varadarajan。博雷尔空间的自同构群。美国数学会汇刊，109(2):191–220，1963年。Bastiaan S Veeling、Jasper Linmans、Jim Winkens、Taco Cohen 和 Max Welling。用于数字病理学的旋转等变卷积神经网络。载于医学图像计算与计算机辅助干预国际会议，第210–218页。Springer，2018年。Dian Wang、Robin Walters 和 Robert Platt。SO(2)等变强化学习。载于国际学习表征会议，2021年。

Maurice Weiler、Mario Geiger、Max Welling、Wouter Boomsma 和 Taco S Cohen。三维可转向卷积神经网络：在体积数据中学习旋转等变特征。神经信息处理系统进展，31，2018年。

William Whitney、Rajat Agarwal、Kyunghyun Cho 和 Abhinav Gupta。动力学感知嵌入。载于国际学习表征会议，2019年。

Ronald J Williams。用于联结主义强化学习的简单统计梯度跟随算法。机器学习，8(3):229–256，1992 年。

Marysia Winkels 和 Taco S Cohen。使用等变卷积神经网络进行 CT扫描 中的肺结节检测。医学图像分析，55:15–26，2019年。

Alicia P Wolfe 和 Andrew G Barto。《用于寻找可复用马尔可夫决策过程同态的决策树方法》。载于人工智能全国会议论文集，体积 21，第 530 页。美国加利福尼亚州门洛帕克；美国马萨诸塞州剑桥；英国伦敦；AAAI出版社；麻省理工学院出版社；1999年，2006年a。

Alicia Peregrin Wolfe 和 Andrew G Barto。使用马尔可夫决策过程同态定义对象类型和选项。载于*ICML-06*机器学习结构化知识迁移研讨会论文集。Citeseer，2006年b。

Denis Yarats、Ilya Kostrikov 和 Rob Fergus。图像增强就是你所需要的一切：对基于像素的深度强化学习进行正则化。载于国际学习表征会议，2020年。

Denis Yarats、Rob Fergus、Alessandro Lazaric 和 Lerrel Pinto。掌握视觉连续控制：改进的数据增强强化学习。载于国际学习表征会议，2021年a。

Denis Yarats、Amy Zhang、Ilya Kostrikov、Brandon Amos、Joelle Pineau 和 Rob Fergus。提高基于图像的无模型强化学习的样本效率。载于*AAAI*人工智能会议论文集，第35卷，第10674–10681页，2021年b。

Amy Zhang、Rowan Thomas McAllister、Roberto Calandra、Yarin Gal 和 Sergey Levine。《无需重建的强化学习不变表征学习》。载于国际学习表征会议，2020年。

Brian D Ziebart。基于最大因果熵原理对有目的的自适应行为进行建模。卡内基梅隆大学，2010年。

Brian D Ziebart、Andrew Maas、J Andrew Bagnell 和 Anind K Dey。最大熵逆强化学习。载于第 *23* 届全国人工智能会议论文集—第 *3* 卷，第 1433–1438 页，2008 年。

附录 A. 假设与条件

我们的同态策略梯度定理的推导针对连续状态与动作空间。因此，我们假设实际MDP M 及其在 MDP 同态映射 h 下的 MDP 同态像 M 满足以下正则性条件。这些条件主要基于确定性策略梯度定理的正则性条件（Silver 等人，2014 年）：

正则性条件 1：$\tau_a(s'|s)$、$\nabla_a\tau_a(s'|s)$、$\tau_a(s'|\bar{s})$、$\nabla_a\tau_a(s'|\bar{s})$、$R(s,a)$、$\nabla_a R(s,a)$、$R(\bar{s},\bar{a})$、$\nabla_a R(\bar{s},\bar{a})$、$\pi_\theta^{\uparrow}(s)$、$\nabla_\theta\pi_\theta^{\uparrow}(s)$、$\pi_\theta(s)$、$\nabla_\theta\pi_\theta(s)$、$p_1(s)$以及 $p_1(s)$关于所有参数和变量 $\bar{s}$、$\bar{s}$、a、a、s'和 s'均连续。

正则性条件 2：存在一个 b和 L，使得 $\sup_s p_1(s) < b$，$\sup_s p_1(s) < b$，$\sup_{a,s,s'} \tau_a(s'|\bar{s}) < b$，$\sup_{a,s,s'} \tau_a(s'|s) < b$，$\sup_{a,s} R(\bar{s},\bar{a}) < b$，$\sup_{a,s} R(s,a) < b$，$\sup_{a,s,s'} \|\nabla_a \tau_a(s'|\bar{s})\| < L$，$\sup_{a,s,s'} \|\nabla_a \tau_a(s'|s)\| < L$，$\sup_{s,a} \|\nabla_a R(\bar{s},\bar{a})\| < L$，$\sup_{s,a} \|\nabla_a R(s,a)\| < L$。

正则性条件 3：动作映射 $g_s(a)$是一个局部微分同胚（定义 30）。因此，它相对于 a 是连续的，并且相对于 a是局部双射的。此外，$\nabla_a g_s(a)$关于参数 a连续，并且存在一个 L，使得 $\sup_{s,\ a} \|\nabla_a g_s(a)\| < L$。

附录 B. 数学工具

本节简要介绍测度论和微分几何中的若干数学概念。我们只明确介绍与第 4 节和第 5 节中的证明直接相关或有所涉及的概念；如需更全面的概述，读者可参阅 Bogachev 和 Ruas（2007年）、Lang（2012年）及 Spivak（2018年）等教材。

B.1 度量空间和拓扑学

配备了度量的集合称为度量空间，通常写作一对，典型形式为 (X,d)。给定一个度量，可以定义基础分析中的标准概念：收敛序列、序列极限、柯西序列和连续函数。如果每个柯西序列都收敛，我们就称该度量空间是完备的。在真度量空间中，收敛序列的极限是唯一的，但在伪度量空间中则不一定。度量空间之间的函数$f:(X,d)\to(Y,d')$称为非扩张，如果：

$$\forall x,x'\in X, d'(f(x),f(x'))\le d(x,x').$$

如果存在某个数 $c\in(0,\ 1)$，使得：则称函数为压缩的。

$$\forall x,x'\in X, d'(f(x),f(x'))< c\cdot d(x,x').$$

度量空间的基本定理称为巴拿赫不动点定理，其内容如下：

定理 20 如果 f 是从一个完备度量空间 X 到其自身的压缩映射，那么 f存在一个唯一的不动点，即存在唯一的点 $x_0\in X$，使得 $f(x_0)=x_0$。

我们假设读者熟悉拓扑学的基本概念：开集与闭集、基与子基、序列收敛和函数连续性。

如果一个拓扑空间可以赋予一个生成其拓扑的度量，且由此得到的度量空间是完备的，那么称该拓扑空间是完全可度量的。如果一个拓扑空间包含一个可数稠密子集——也就是说，该拓扑空间中的每个非空子集都至少包含该子集中的一个元素——那么称该拓扑空间是可分的。波兰空间是指既可分又完全可度量的拓扑空间。波兰空间具有“理想的性质”，主要用于描述集合论和测度论领域。

拓扑空间的另一个基本概念是紧致性。拓扑空间 X 的开覆盖是一个开子集族，其并集包含 X的全部元素。覆盖的子覆盖是覆盖中开集的一个子族，并且同样覆盖 X。如果每个开覆盖都有有限子覆盖，则称拓扑空间是紧致的。在度量空间中，紧集恰好是闭集和有界集。若每个点都包含于某个开集，而该开集又包含于某个紧集之中，则称空间是局部紧的。

B.2 测度论

测度论试图将长度、面积、体积或质量的概念推广到比人们在几何学中最初接触到的简单子集更为复杂的子集上。在实直线等常见情形中，不可能在所有子集上定义测度，同时又使人们关于长度的通常直觉得以保留。在实直线上，不存在一种定义于所有子集上的测度，并且该测度赋予所有区间以其长度。因此，人们需要选择一些性质良好的子集族，并在其上定义测度。

定义 21（σ-代数）给定集合 X， σ-代数是 X 上的子集族 Σ，满足：*1）* $X \in \Sigma$；*2）* $A \in \Sigma$ 蕴含 $A^c \in \Sigma$ （对补集封闭）；以及 *3）*如果（A_i）$_{i\in\mathbb{N}}$ 满足对于所有 $i \in \mathbb{N}$都有 $A_i \in \Sigma$，那么 $\cup_{i\in\mathbb{N}} A_i \in \Sigma$ （对可数并封闭）。有序组（X， Σ）是一个可测空间。

配备 σ 代数的集合称为<样式 id='2'>可测空间。空间的 σ 代数规定了哪些集合可以定义测度；在概率论中——以及在我们的应用场景中—— σ 代数表示一组可赋予概率的事件。

对于任意一族子集，都存在包含该族的<样式 id='1'>最小 σ代数：这称为由该族生成的 σ代数。在度量空间中，由开集生成的 σ代数称为<样式 id='6'>博雷尔 σ代数。

给定一个 σ代数，就可以定义测度。

<样式 *id='1'*>定义 *22*定义在 X 上、基于 σ 代数 Σ 的（概率）<样式 *id='3'*>测度 μ 是一个函数$\mu : \Sigma \to [0, \infty]$ *(*$\mu : \Sigma \to [0, 1]$*)*，满足：

- $\mu(\varnothing) = 0$
- 如果$\{A_i\}_{i\in\mathbb{N}}$ 是 Σ 中两两不交子集的可数族，那么 $\mu(\cup_i A_i) = \sum_{i\in\mathbb{N}} \mu(A_i)$。

对于概率测度，我们要求 $\mu(X) = 1$。

起关键作用的函数称为可测函数。

定义 23 两个可测空间 $(X\Sigma)$ 和(Y,Λ)之间的一个可测函数或映射是一个满足以下条件的函数 $f:X\to Y$：对于任意 $B\in\Lambda$, $f^{-1}(B)\in\Sigma$。

我们可以基于原像上的测度，在可测函数的像上定义测度。这便导出了推前测度和变量替换公式的定义；对于证明中通过可测函数切换积分域而言，这一公式至关重要。

定义 24（推前测度）设 (X_1,Σ_1) 和 (X_2,Σ_2) 是两个可测空间，$f:X_1\to X_2$ 是一个可测映射，$\mu:\Sigma_1\to[0,\infty]$ 是 X_1上的一个测度。则关于 f 的 μ 的推前测度，记作 $f_*(\mu):\Sigma_2\to[0,\infty]$，定义为：

$$(f_*(\mu))(B)=\mu(f^{-1}(B))\ \forall\ B\in\Sigma_2.$$

定理 25（变量替换）可测空间 X_2 上的可测函数 g 关于 $f_*(\mu)$ 可积，当且仅当函数 $g\circ f$ 关于 μ 可积；在此情况下，两个积分相等：

$$\int_{X_2} g d(f_*(\mu))=\int_{X_1} g\circ f d\mu.$$

B.3 流形

微分流形将如何在曲面上进行微积分形式化。与向量空间不同，在任意拓扑空间中，我们无法定义点的加法；我们需要额外的结构。因此，策略是定义拓扑空间中一些“看起来像”向量空间中相应部分的“补丁”，然后将它们粘合起来。这促成了微分流形或光滑流形的定义。“光滑”是无穷次可微 或 C^∞的同义词。

定义 26 一个 n维光滑（或微分）流形 是一个拓扑空间 M[7]，它配备了一个称为坐标图的有序对族，$\{(U_i,\ \phi_i)|i\in\mathcal{A}\}$ 其中：

- 每个 U_i 都是 M的开子集，
- 每个 $\phi_i:U_i\to\mathbf{R}^n$ 都是 U_i 与像 $V_i:=\phi_i(U_i)$之间的同胚 ，
- $\{U_i\}$ 构成 M的一个开覆盖。

此外，还必须满足以下相容条件：若 $U_i\cap U_j\neq\varnothing$，则映射

$$\phi_j\circ\phi_i^{-1}\Big|_{\phi_i(U_i\cap U_j)}:\phi_i(U_i\cap U_j)\to\phi_j(U_i\cap U_j)$$

是无穷次可微的，记作 C^∞。一组相容图册称为图册。

7. 假设其为仿紧且豪斯多夫。

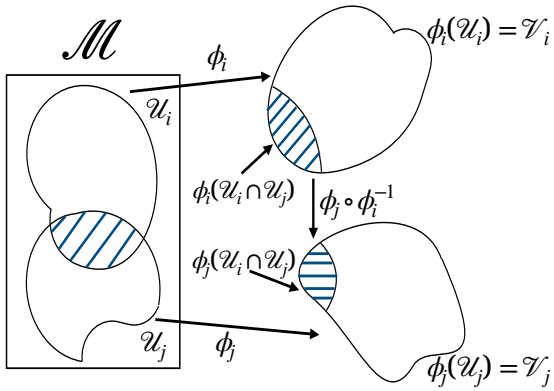


图11：坐标图上的相容条件。

请注意，最后一个条件指的是从$\mathbf{R}^n$ 中的一个开集到$\mathbf{R}^n$中的另一个开集的映射；因此，由于$\mathbf{R}^n$ 是一个向量空间，该条件的含义很明确。图 11 的示意图说明了这一条件的含义。现在，一切都可以用坐标图来定义。

定义 27 *A* 光滑函数 $f : M \to \mathbf{R}$ 是满足以下条件的函数：对于任意坐标图(U, ϕ)，映射 $f \circ \phi^{-1} : \mathbf{R}^n \to \mathbf{R}^8$ 是光滑的。

流形 M 和 M' 之间的光滑映射可以用类似的方式定义。

定义 28 光滑映射 $f : M \to M'$是一个函数，对于M的任意坐标图(U, ϕ) 和 $M_{\prime}$的任意坐标图(U', ϕ')，映射 $\phi' \circ \psi \circ \phi^{-1} : \mathbf{R}^n \to \mathbf{R}^{n'}$都是光滑的。

光滑映射和光滑函数自动都是连续的
。微分流形具有一种称为微分同胚的同构概念。

定义 29 两个流形之间的光滑映射称为微分同胚，如果它是双射，且逆映射也光滑。

定义 30（局部微分同胚）设 M 和 N 是可微流形。若函数 $f : M \to N$ 满足：对于每个点 $x \in M$，都存在一个包含 x的开集 U，使得 $f(U)$是 N中的开集，且 $f|_U : U \to f(U)$是微分同胚，则称其为局部微分同胚。

一旦建立了光滑流形的结构，就可以定义导数算子的概念。曲线的切向量也可以用微分来定义。点 x 处的切向量 t 应被视为方向导数。可以借助多元微积分中的标准结果证明，某一点处的切向量集合构成一个 n维向量空间。这个向量空间记作 T_x，称为切空间，位于 x处。

在点 x 处公理化切向量这一概念的最简洁方式如下。设 F 为定义在 M上的光滑实值函数集合。

8. 实际上，这只在 $\phi(U)$上有定义，而不是在整个$\mathbf{R}^n$ 上有定义，但若总是写出正确的限制条件，记号就会过于繁杂。

定义 31 给定 M 上的一点 x，我们将映射$t : \mathcal{F} \to \mathbf{R}$称为 x 处的切向量，其中满足：$\forall a$、$b \in \mathbf{R}$且 $\forall f$、$g \in \mathcal{F}$:

1. $t(af + bg) = at(f) + bt(g)$,
2. $t(fg) = f(x)t(g) + g(x)t(f)$.

由此立即可知，如果 f是常值函数，那么 $t(f) = 0$。请注意，第二个条件明确提到了点 x。

光滑映射 $\psi : M \to N$在相应点的切空间之间诱导出一个映射。映射 ψ 在 x 处的微分 是如下定义的线性映射 $\mathrm{d}\psi : T_x M \to T_{\psi(x)} N$。设 g 是 $\psi(x)$的一个邻域上的光滑函数，且 t 是 x处的一个切向量。我们需要在 $\psi(x)$处定义一个切向量，因此它应能够作用于 g。我们定义

$$\mathrm{d}\psi(t)(g) := t(g \circ \psi).$$

以下定理是基础性的，并用于策略梯度定理的证明。

定理 32（流形上的逆函数定理） 如果 $f : M \to N$ 是一个光滑映射，并且其微分 $df_x : T_x M \to T_{f(x)} N$ 在点 $x \in M$处是一个同构，那么 f 在 x处是一个局部微分同胚。

定理 33（流形上的链式法则）如果 $f : M \to N$和 $g : N \to O$ 是流形之间的光滑映射，那么：

$$d(g \circ f)_x = dg_{f(x)} \circ df_x.$$

附录 C：完整结果

如第 7 节所述，我们在 DM Control 中的连续控制任务（像素观测）以及自定义环境上评估 DHPG。重要的是，为了可靠地将我们的算法与基线方法进行比较，并准确呈现结果的分布，我们遵循 Agarwal 等人（2021）提出的最佳实践，报告在所有任务上、基于 10 个随机种子汇总得到的四分位均值（IQM）和性能剖面。虽然我们的基线结果是使用官方代码得出的，但在可能的情况下，部分结果可能与原文报告的结果有所不同，这是因为随机种子数量不同，而且我们力求如实呈现真实的性能分布（Agarwal等人，2021年）。

我们使用 DrQv2、DBC 和 SAC-AE 的官方实现；由于官方代码不可用，我们重新实现了 DeepMDP；有关基线方法的更多详情，请参阅附录 D.2。

C.1 DeepMind 控制套件

图 12–13 展示了在 16 项 DeepMind 控制套件任务上使用像素观测获得的完整结果，以补充第 7.1 节的结果。本基准测试不包括需要过多探索和大量时间步的领域（例如杂技演员、游泳者和人形机器人），也不包括目标在视觉上较小的领域（例如困难触及任务和困难手指转动任务）。

图 14–15 和图 16–17 分别展示了 14 项任务上的性能剖面和汇总指标（Agarwal等人，2021年）；由于在 100 万个时间步内，所有算法均未取得合理的性能，因此在 RLiable 评估中移除了跳跃者跳跃和 Walker Run。

在图 12、14 和 16 中，所有方法均采用图像增强；而在图 13、15 和 17 中，所有方法均不采用图像增强。

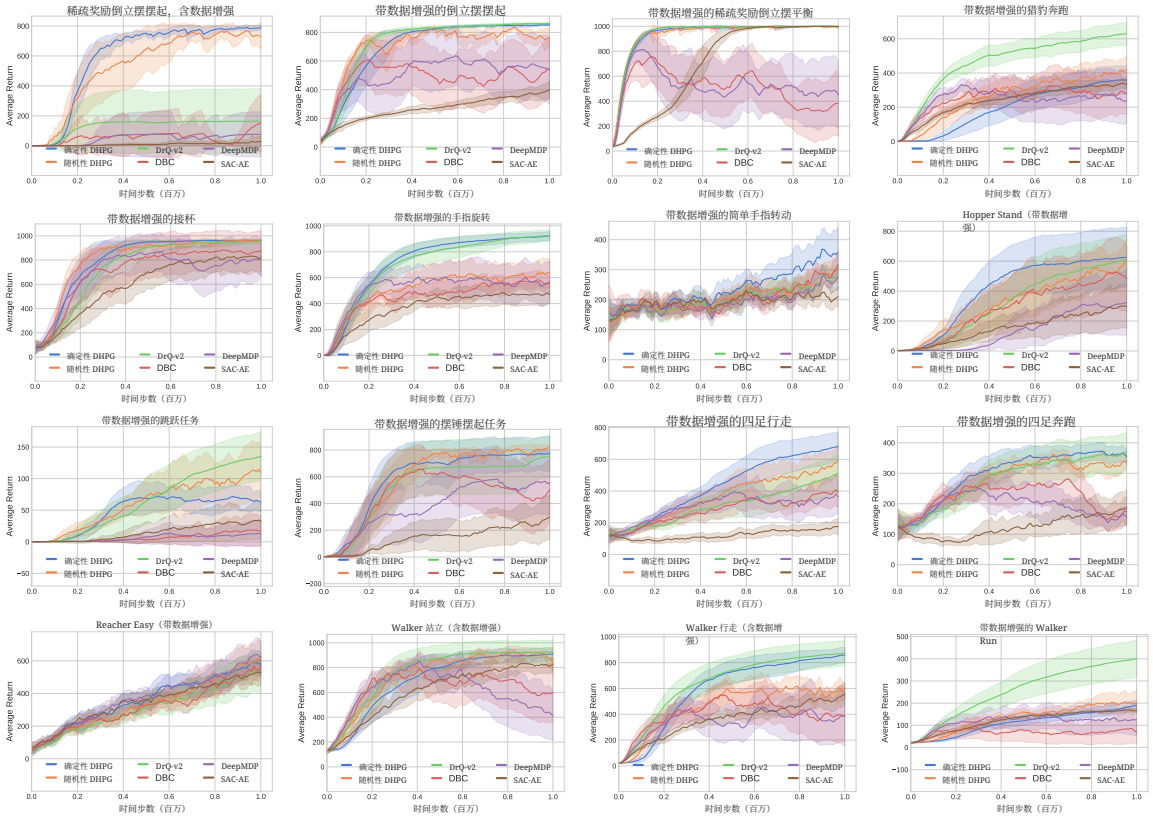


图12：采用像素观测的 16 个 DM 控制任务的学习曲线。所有方法均使用 图像增强。均值表现基于 10 个随机种子计算，阴影区域表示 95% 置信区间。为便于观察，所有图表均进行了统一的平滑处理。

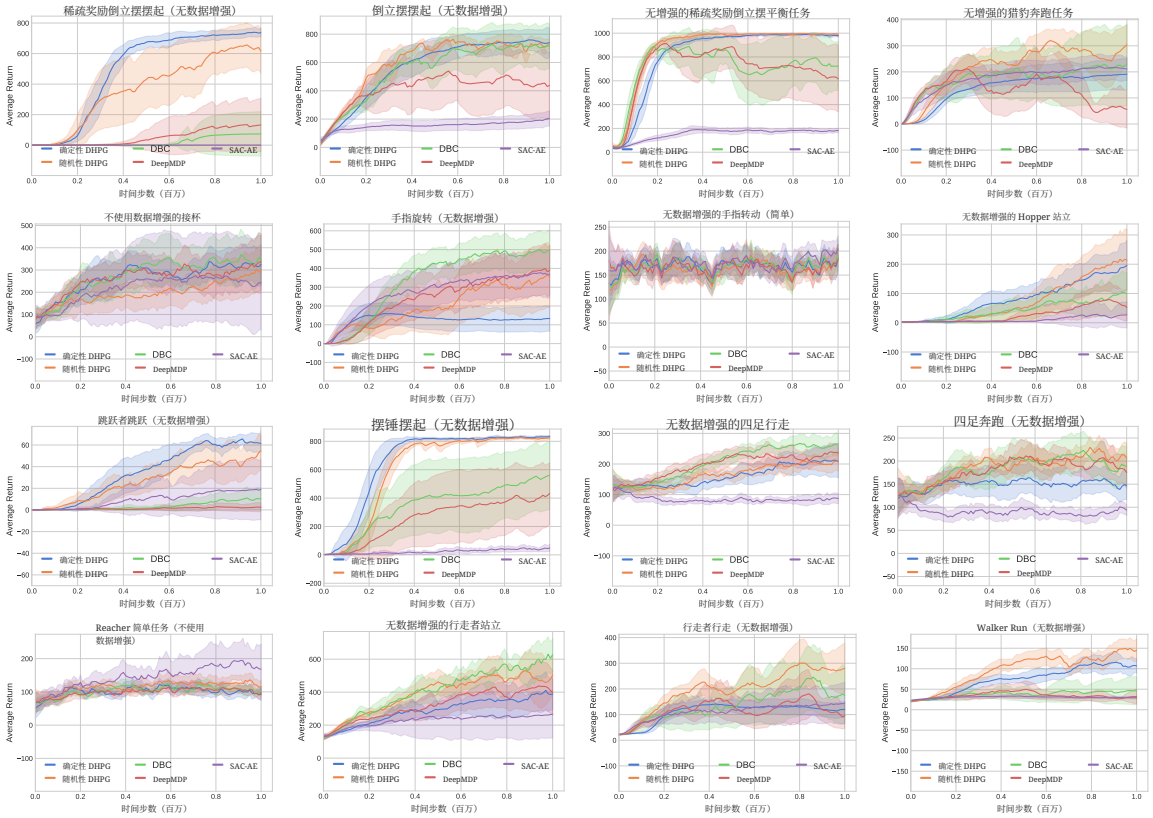
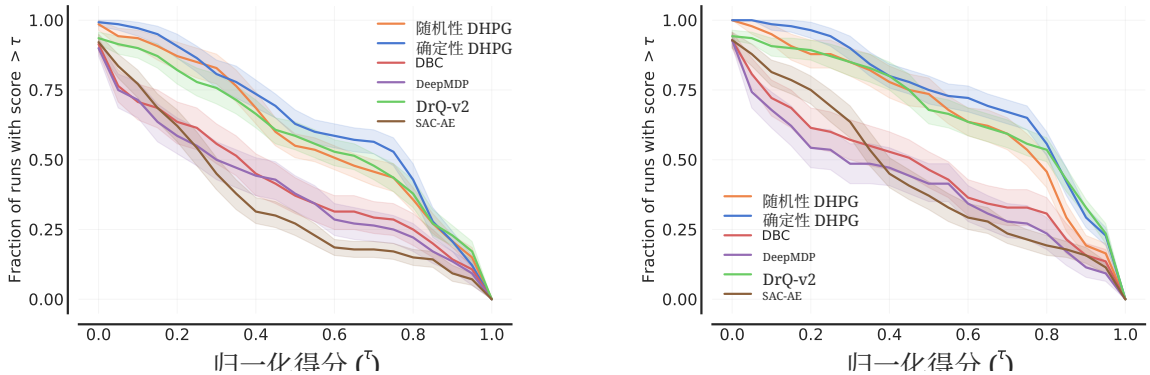
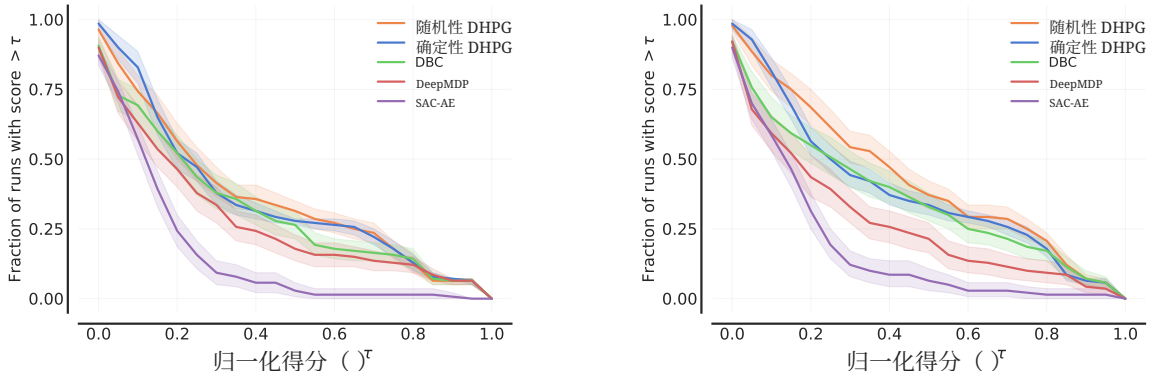


图13：16 个采用像素观测的 DM 控制任务的学习曲线。所有方法均**未使用**图像增强。均值性能基于 10 个随机种子计算，阴影区域表示 95% 置信区间。为使图示更清晰，所有曲线均进行了统一光滑处理。



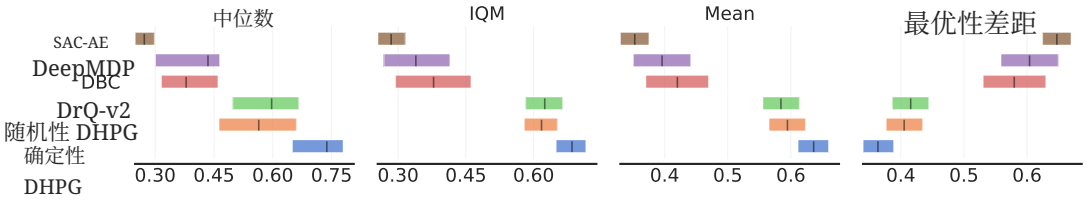
(a) 50 万步基准测试。 (b) 100万步基准测试。

图14：基于14项任务和10个随机种子的像素观测性能剖面，分别为500k步时(a)和1m步时(b)。所有方法**均使用**图像增强。阴影区域表示95% 置信区间。

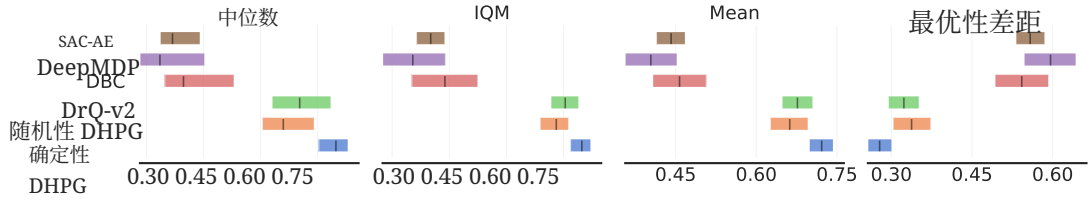


（a）50万步基准测试。 (b) 100 万步基准测试。

图15：基于14项任务和10个随机种子的像素观测性能剖面，分别对应500k步(a)和1m步(b)。所有方法**均未使用**图像增强。阴影区域表示95% 置信区间。

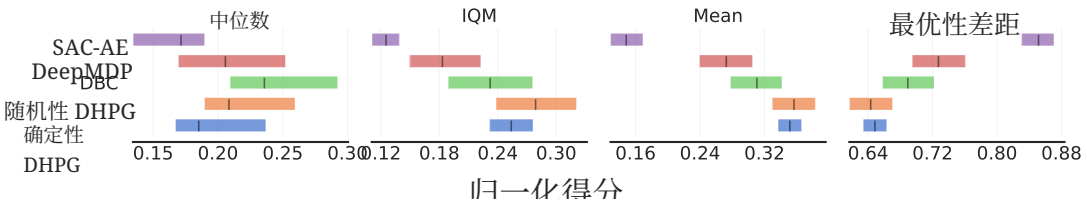


(a) 50 万步基准测试。

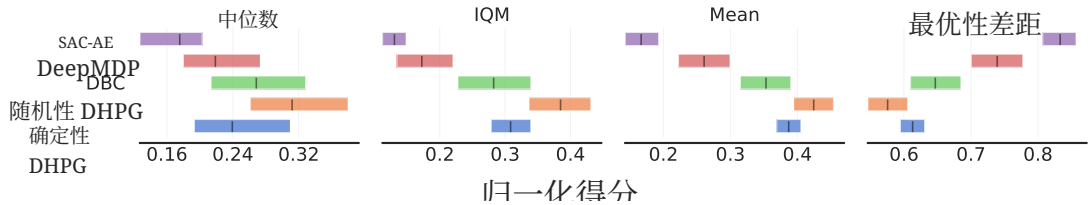


(b) 100 万步基准测试。

图16：基于 10 个随机种子的 14项任务，在 50 万步(a)和 100 万步(b)时，像素观测的汇总指标及其 95% 置信区间。所有方法**均采用**图像增强。



(a) 50 万步基准测试。

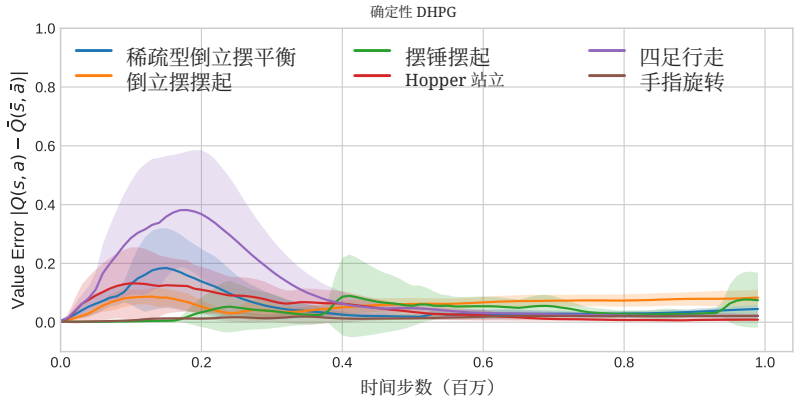


(b) 100万步基准测试。

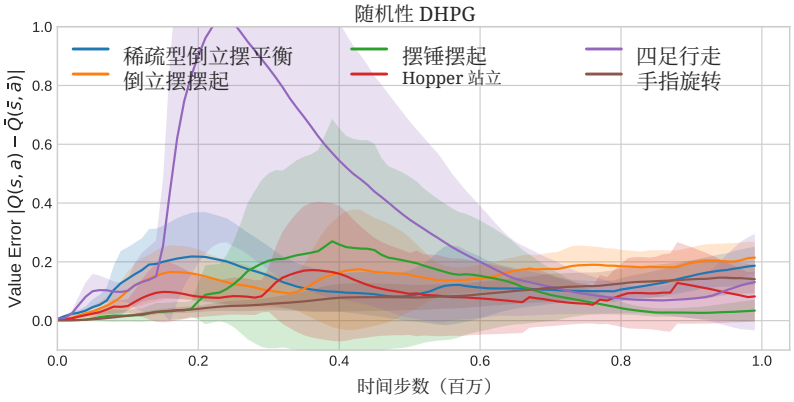
图17：基于14项任务和10个随机种子的像素观测汇总指标及95% 置信区间，分别对应50万步(a)和100万步(b)。所有方法均**未使用**图像增强。

C.2 实践中的价值等价性质

我们可以用实际MDP与抽象MDP的评论家之间的价值等价性，作为衡量学习得到的MDP同态质量的指标，因为这两个评论家并非直接通过训练来最小化这一距离，而是通过学习得到的MDP同态映射获得等价的价值。图18展示了$|Q(s,\ a)–Q(s,\ a)|$在训练过程中的归一化平均绝对误差，表明这一性质在实践中成立。不出所料，对于同态映射容易学习的低维任务（例如倒立摆），误差比复杂任务（例如四足机器人）更早降低。但重要的是，在所有情况下，误差都会随时间推移而下降，并在训练结束时处于合理范围内，这意味着连续MDP同态满足定义 11中的条件。



(a) 确定性DHPG的价值等价。

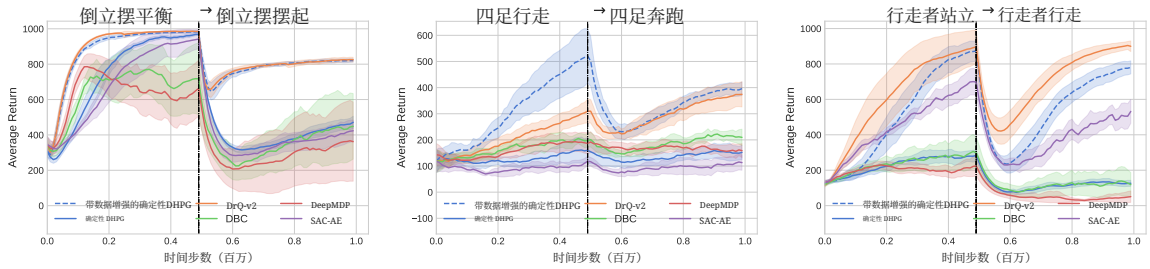


(b) 随机 DHPG的价值等价。

图18：归一化平均绝对误差 $|Q(s,a) - Q(s,a)|$，作为不同任务从像素观测进行训练期间价值等价性质的测度。误差根据回放缓冲区中的样本测量，并按价值函数的范围进行归一化。误差取 10 个随机种子的均值，阴影区域表示 95% 置信区间。

C.3 迁移学习实验

如第 7 节所述，迁移实验旨在确保使用 MDP同态不会损害迁移能力。我们在这些实验中选择确定性DHPG。图19展示了一系列迁移场景的学习曲线，其中评论家、演员和表征被迁移到同一领域内的新任务。DHPG与其他方法具有相同的迁移能力。



(a) 倒立摆。 (b) 四足机器人。 (c) 行走机器人。

图19：像素观测下迁移实验的学习曲线。在 500k 时间步处，所有组件都被迁移到同一领域内的新任务。均值表现是对 10 个随机种子取平均得到的，阴影区域表示 95% 置信区间。为便于观察，所有图表均经过一致的光滑处理。

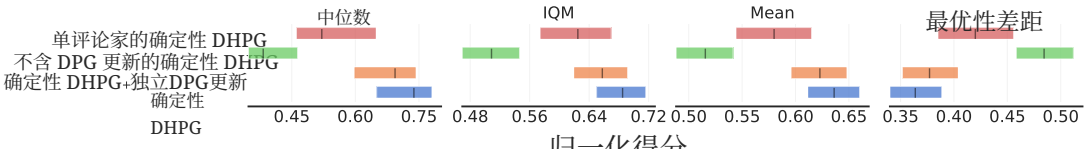
C.4 HPG 与标准策略梯度相结合的消融研究

我们对 HPG 与用于行动者更新的标准策略梯度的组合进行了消融研究。由于提升策略通常可以通过解析方式获得，因此确定性 DHPG 算法通常更简单。我们评估了确定性 DHPG 的四种变体的性能（所有变体均使用图像增强）：

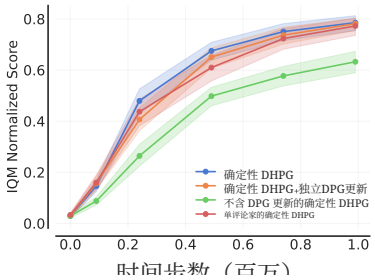
1. **DHPG：** 将 HPG 和 DPG 的梯度相加，并基于梯度之和执行一次演员更新。这就是本文中始终使用的确定性 DHPG 算法。2. **采用独立 DPG 更新的 DHPG：** 分别使用 HPG 和 DPG 的梯度来更新演员。3. **不进行 DPG 更新的 DHPG：** 仅使用 HPG 来更新演员。4. **采用单评论家的 DHPG：** 训练一个评论家网络，以学习实际 MDP 和抽象 MDP 的价值。因此，使用 HPG 和 DPG 来更新演员。

图 21 展示了在 16 项采用像素观测的 DeepMind 控制套件任务上获得的学习曲线，图 20 则展示了 RLiable（Agarwal等人，2021年）的评估指标。总体而言，将 HPG 和 DPG 的梯度相加（变体 1），与分别进行 HPG 和 DPG 更新（变体 2）相比，梯度估计的方差更低。有趣的是，在学习MDP同态较为容易的简单任务（例如倒立摆和摆杆）中，不含 DPG 的 DHPG 变体（变体 3）表现良好，甚至优于其他变体，这表明我们的方法能够有效地仅利用

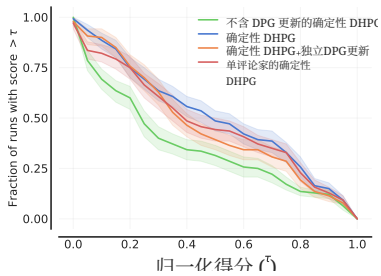
抽象MDP 来更新实际MDP 的策略。然而，对于较复杂的任务（例如行走机器人），还需要通过 DPG 使用实际MDP 进行策略优化。最后，在实际MDP 和抽象马尔可夫决策过程之间共用单评论家（变体 4），可以提高对称马尔可夫决策过程中的样本效率，但由于两个 MDP 之间的误差界较大，在非对称马尔可夫决策过程中可能会导致性能下降，$\|Q^{\pi^{\uparrow}}(s,\ a) - Q^{\pi}(s,\ a)\|$ (Taylor 等人，2008年)。



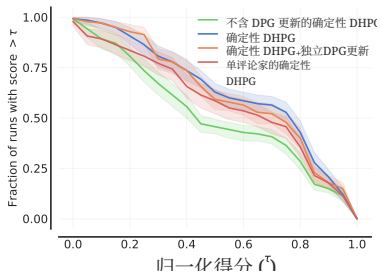
(a) 500k 时的汇总指标。



(b) 样本效率。



(c) 250k 时的性能剖面。



(d) 500k 时的性能剖面。

图 20：HPG 与 DPG 组合的消融研究。基于像素观测的 RLiable 评估指标，在 14 项任务上对 10 个随机种子取平均。500k 步时的汇总指标(a)，用于比较样本效率的 IQM 分数随步数变化的曲线(b)，250k 步时的性能剖面(c)，500k 步时的性能剖面(d)。阴影区域表示 95% 置信区间。

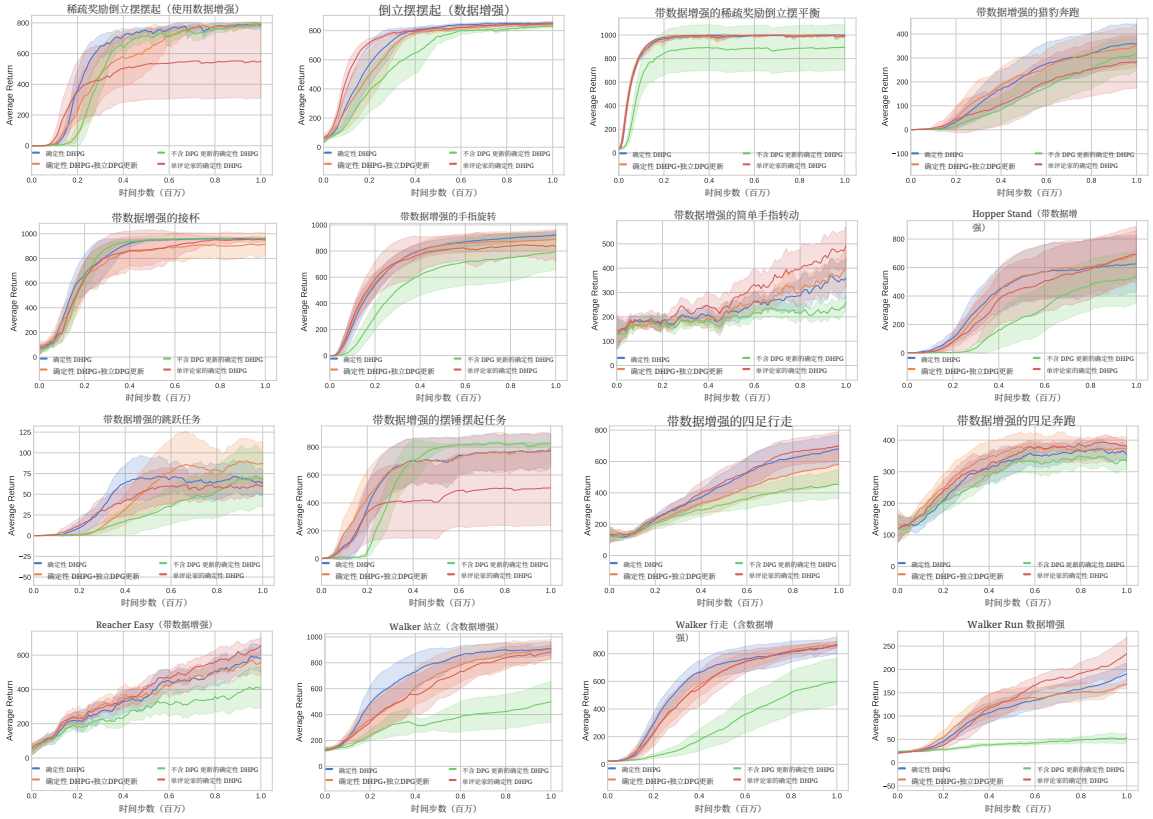
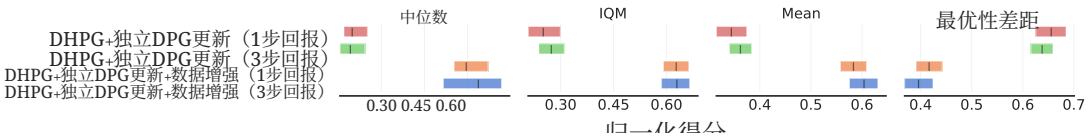


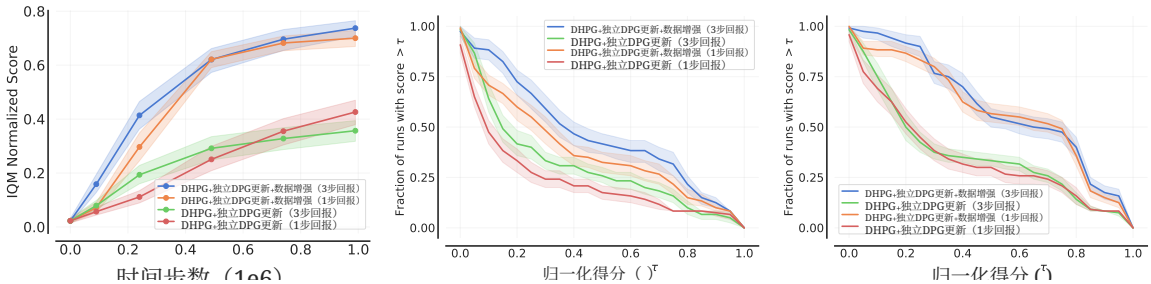
图 21：HPG 与 DPG 组合的消融研究。16 项采用像素观测的 DM 控制任务的学习曲线。均值性能取自 10 个随机种子，阴影区域表示 95% 置信区间。为便于观察，所有图表均经过统一光滑处理。

C.5 n步回报的消融研究

我们对 DHPG 所采用的 n步回报进行了消融研究。与附录 C.4 中的消融研究类似，我们选择确定性DHPG，因为它通常更简单，且其提升策略可以通过解析方式获得。图22展示了针对像素观测，采用1步和3步回报的 DHPG 的 RLiable（Agarwal等人，2021年）评估指标。我们展示了采用和不采用图像增强时， n步回报对 DHPG 的影响。总体而言， n步回报似乎能改善训练的早期阶段。对于不采用图像增强的 DHPG，1步回报的最终性能优于3步回报，这或许表明，使用 n步回报会使学习MDP同态变得更加困难。



(a) 500k 步时的汇总指标。



(b) 样本效率。 (c) 250k 步时的性能剖面。(d)500k 步时的性能剖面。

图22： n步回报的消融研究。对 12 个任务、10 个随机种子上的像素观测取平均后得到的 RLiable 评估指标。1m 步时的汇总指标(a)、用于比较样本效率的 IQM 分数随步数的变化(b)、250k 步时的性能剖面(c)，以及 500k 步时的性能剖面(d)。阴影区域表示 95% 置信区间。

C.6 计算时间比较

图23比较了我们的方法与基线方法的计算成本。横轴表示墙钟时间（小时）。由于我们的方法不需要进行图像重建，因此其计算效率高于 SAC-AE 和 DeepMDP。然而，与 DrQ-v2 相比，双模拟计算、HPG 更新以及策略提升损失（针对随机 DHPG）增加了我们方法的计算成本。

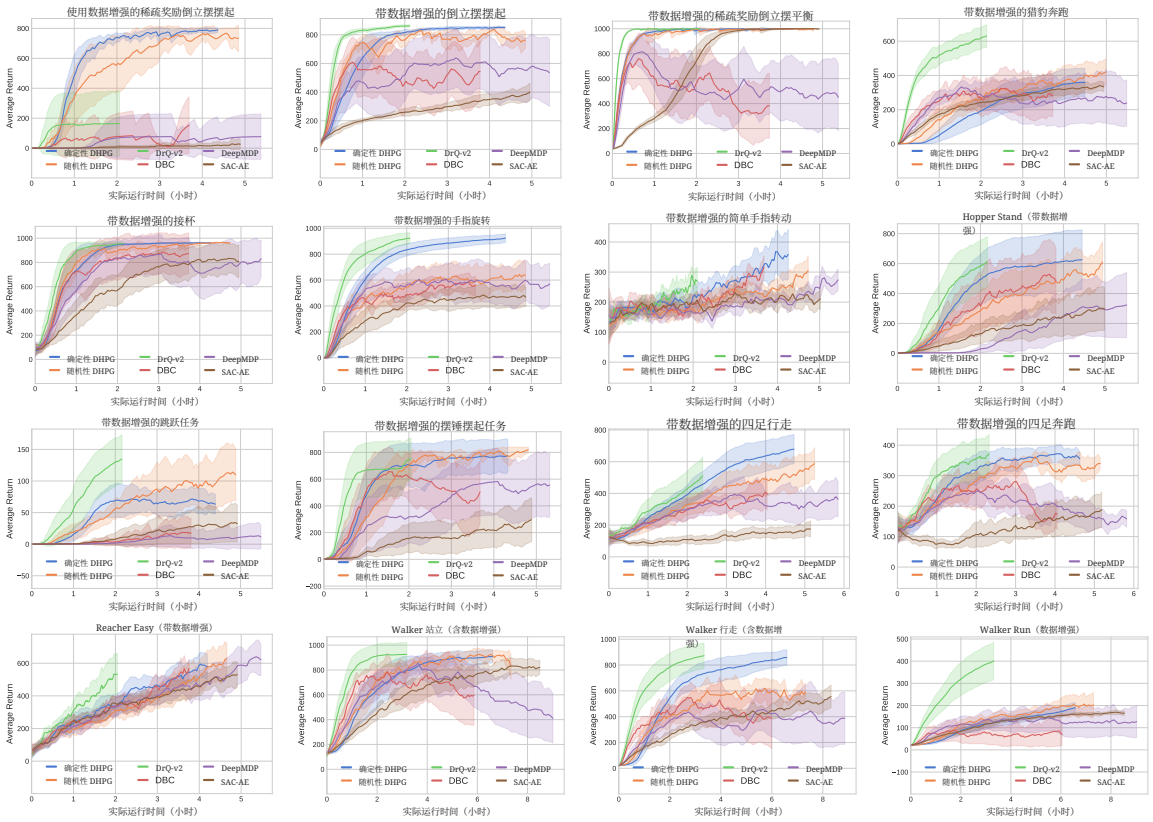


图 23：16 项 DM 控制任务在像素观测下的学习曲线。横轴为**墙钟时间**，单位为小时。所有方法**均使用**图像增强。性能均值基于 10 个随机种子计算，阴影区域表示 95% 置信区间。为清晰起见，所有曲线均采用统一方式进行平滑处理。

附录 D. 实现细节

D.1 超参数

第 6 节中的算法 1 展示了用于像素观测的随机和确定性 DHPG 算法的伪代码。

在 DHPG 的图像增强版本以及所有基线方法中，我们使用 DrQ（Yarats等人（2020年））的图像增强方法，该方法只是对像素观测进行随机平移。首先，将 84×84 张图像填充 4 个像素（通过重复边界像素），然后随机选取一个 84×84 裁剪区域，从而将原始图像渲染为平移 ± 4 个像素后的图像。与 Yarats等人（2021a）类似，我们还会在平移后的图像上进行双线性插值，将每个像素值替换为最近的四个像素值的平均值。我们的代码已公开发布于 https://github.com/sahandrez/homomorphic policy gradient。

我们使用 PyTorch（Paszke et al., 2019）实现了我们的方法，实验结果是在云计算服务上的 A100 GPU 上，使用 Python v3.8.10、PyTorch v1.10.0、CUDA 11.4 和 Mujoco 2.1.1（Todorov et al., 2012）获得的。表 1 列出了我们实验中使用的超参数。这些超参数均改编自 DrQ-v2（Yarats et al., 2021a），且未进行任何进一步的超参数调优。除 walker 领域外，所有算法和任务均使用相同的超参数设置；对于 walker 领域，与 DrQ-v2（Yarats et al., 2021a）类似，我们使用 n 步回报和 $n = 1$ 小批量大小 512。

核心强化学习组件（演员和评论家网络），以及 DHPG 的组件（状态和动作编码器、转移模型和奖励模型）均为采用 ReLU 激活函数、包含一个维度为 256 的隐藏层的多层感知机网络。图像编码器基于 DrQ-v2 的架构，而 DrQ-v2 本身基于 SAC-AE（Yarats 等，

表 1：我们实验中使用的超参数。

超参数	设置
学习率	1e−4
优化器	Adam
n步回报	3
小批量大小	256
演员更新频率 d	2
目标网络更新频率	2
目标网络软更新 τ	0.01
目标策略平滑标准差。裁剪 c	0.3
特征维度	50
动作重复	2
帧堆叠	3
隐藏层维度。	256
回放缓冲区容量	10^6
折扣 γ	0.99
种子帧	4000
探索步数	2000
探索标准差调度	linear(1.0, 0.1, 1e6)

2021b），由 $32 \times 3 \times 3$ 个卷积层组成，使用 ReLU 作为激活函数，后接一个带层归一化（Ba 等，2016）和 tanh 激活函数的单层全连接神经网络。卷积层的步幅均为 1，第一层除外，其步幅为 2。带图像重建的基线模型的图像解码器基于 SAC-AE（Yarats 等，2021b），由一个单层全连接神经网络和 $32 \times 32 \times 3$ 个转置卷积层组成，并使用 ReLU 激活函数。转置卷积层的步幅均为 1，最后一层除外，其步幅为 2。

D.2 基准实现

所有基线方法均已提交至补充材料中。我们使用了 DBC 和 SAC-AE 的官方实现。DeepMDP 没有公开可用的代码，因此我们使用了 DBC 官方代码库中提供的实现。如第 7 节所述，我们运行了两个版本的基线方法，分别采用和不采用图像增强。采用图像增强的变体使用了与附录D.1中所述 DrQ-v2 相同的图像增强方法。