Learning Domain Invariant Representations in Goal-conditioned Block MDPs

在目标条件块 MDP 中学习领域不变表示

Beining Han
Beining Han
IIIS, Tsinghua University
清华大学 IIIS
bouldinghan@gmail.com
Chongyi Zheng
Chongyi Zheng
Carnegie Mellon University
卡内基梅隆大学
chongyiz@andrew.cmu.edu
Harris Chan Keiran Paster Michael R. Zhang Jimmy Ba
Harris Chan Keiran Paster Michael R. Zhang Jimmy Ba
University of Toronto & Vector Institute
多伦多大学与向量研究所
{hchan, keirp, michael, jba}@cs.toronto.edu
{hchan, keirp, michael, jba}@cs.toronto.edu

Abstract

摘要

Deep Reinforcement Learning (RL) is successful in solving many complex Markov Decision Processes (MDPs) problems. However, agents often face unanticipated environmental changes after deployment in the real world. These changes are often spurious and unrelated to the underlying problem, such as background shifts for visual input agents. Unfortunately, deep RL agents are usually sensitive to these changes and fail to act robustly against them. This resembles the problem of domain generalization in supervised learning. In this work, we study this problem for goal-conditioned RL agents. We propose a theoretical framework in the Block MDP setting that characterizes the generalizability of goal-conditioned policies to new environments. Under this framework, we develop a practical method PA-SkewFit that enhances domain generalization. The empirical evaluation shows that our goal-conditioned RL agent can perform well in various unseen test environments, improving by 50% over baselines.

深度强化学习 (RL) 在解决许多复杂的马尔可夫决策过程 (MDP) 问题上取得了成功。然而，代理在实际部署后往往面临意想不到的环境变化。这些变化通常是虚假的，与潜在问题无关，例如视觉输入代理的背景变化。不幸的是，深度 RL 代理通常对这些变化敏感，无法对其采取稳健的行动。这类似于监督学习中的领域泛化问题。在本研究中，我们研究了目标条件 RL 代理的这一问题。我们在块 MDP 设置下提出了一个理论框架，该框架描述了目标条件策略在新环境中的可泛化性。在此框架下，我们开发了一种实用方法 PA-SkewFit，以增强领域泛化。实证评估表明，我们的目标条件 RL 代理在各种未见测试环境中表现良好，相较于基线提高了 50% 。

1 Introduction

1 引言

Deep Reinforcement Learning (RL) has achieved remarkable success in solving high-dimensional Markov Decision Processes (MDPs) problems, e.g., Alpha Zero Silver et al. [2017] for Go, DQN Mnih et al. [2015] for Atari games and SAC Haarnoja et al. [2018] for locomotion control. However, current RL algorithms requires massive amounts of trial and error to learn Silver et al. [2017], Mnih et al. [2015], Haarnoja et al. [2018]. They also tend to overfit to specific environments and often fail to generalize beyond the environment they were trained on Packer et al. [2018]. Unfortunately, this characteristic limits the applicability of RL algorithms for many real world applications. Deployed RL agents, e.g. robots in the field, will often face environment changes in their input such as different backgrounds,

lighting conditions or object shapes Julian et al. [2020]. Many of these changes are often spurious and unrelated to the underlying task, e.g. control. However, RL agents trained without experiencing these changes are sensitive to the changes and often perform poorly in practice Julian et al. [2020], Zhang et al. [2020a b].

深度强化学习 (RL) 在解决高维马尔可夫决策过程 (MDP) 问题方面取得了显著成功，例如，Alpha Zero Silver 等人 [2017] 在围棋中的应用，DQN Mnih 等人 [2015] 在雅达利游戏中的应用，以及 SAC Haarnoja 等人 [2018] 在运动控制中的应用。然而，当前的 RL 算法需要大量的试错来学习 Silver 等人 [2017]、Mnih 等人 [2015]、Haarnoja 等人 [2018]。它们还往往会过拟合特定环境，并且通常无法在训练过的环境之外进行泛化 Packer 等人 [2018]。不幸的是，这一特性限制了 RL 算法在许多现实世界应用中的适用性。部署的 RL 代理，例如现场的机器人，通常会面临输入环境的变化，如不同的背景、光照条件或物体形状 Julian 等人 [2020]。这些变化中的许多往往是虚假的，与基础任务 (例如控制) 无关。然而，未经历这些变化的 RL 代理对变化敏感，通常在实践中表现不佳 Julian 等人 [2020]、Zhang 等人 [2020a b]。

In our work, we seek to tackle changing, diverse problems with goal-conditioned RL agents. Goal-conditioned Reinforcement Learning is a popular research topic as its formulation and method is practical for many robot learning problems Marcin et al. [2017], Eysenbach et al. [2020]. In goal-conditioned MDPs, the agent has to achieve a desired goal state g which is sampled from a prior distribution. The agent should be able to achieve not only the training goals but also new test-time goals. Moreover, in practice, goal-conditioned RL agents often receive high-dimensional inputs for both observations and goals Paster et al. [2020], Péré et al. [2018]. Thus, it is important to ensure that the behaviour of goal-conditioned RL agents is invariant to any irrelevant environmental changes in the input at test time. Previous work Zhang et al. [2020a] tries to address these problems via model bisimulation metric Ferns et al. [2011]. These methods aim to acquire a minimal representation which is invariant to irrelevant environment factors. However, as goal-conditioned MDPs are a family of MDPs indexed by the goals, it is inefficient for these methods to acquire the model bisimulation representation for every possible goal, especially in high-dimensional continuous goal spaces (such as images).

在我们的工作中，我们旨在解决具有目标条件的强化学习代理所面临的变化多样的问题。目标条件强化学习是一个热门的研究主题，因为其公式和方法在许多机器人学习问题中都具有实用性 Marcin et al. [2017], Eysenbach et al. [2020]。在目标条件马尔可夫决策过程 (MDPs) 中，代理需要实现一个从先验分布中采样的期望目标状态 g 。代理不仅应该能够实现训练目标，还应该能够实现新的测试时目标。此外，在实践中，目标条件强化学习代理通常会接收高维输入，包括观察和目标 Paster et al. [2020], Péré et al. [2018]。因此，确保目标条件强化学习代理的行为在测试时对输入中的任何无关环境变化是不变的非常重要。之前的工作 Zhang et al. [2020a] 试图通过模型双仿真度量来解决这些问题 Ferns et al. [2011]。这些方法旨在获取对无关环境因素不变的最小表示。然而，由于目标条件马尔可夫决策过程是一类按目标索引的马尔可夫决策过程，因此这些方法为每个可能的目标获取模型双仿真表示是低效的，尤其是在高维连续目标空间 (如图像) 中。

In our work, we instead choose to optimize a surrogate objective to learn the invariant policy. Our main contributions are:

在我们的工作中，我们选择优化一个替代目标以学习不变策略。我们的主要贡献是:

1. We formulate the Goal-conditioned Block MDPs (GBMDPs) to study domain generalization in the goal-conditioned reinforcement learning setting (Section 2), and propose a general theory characterizing how well a policy generalizes to unseen environments (Section 3.1).

1. 我们构建了目标条件块马尔可夫决策过程 (GBMDPs)，以研究目标条件强化学习设置中的领域泛化 (第 2 节)，并提出了一种通用理论，描述策略在未见环境中的泛化能力 (第 3.1 节)。

2. We propose a theoretically-motivated algorithm based on optimizing a surrogate objective, perfect alignment, with aligned data (Section 3.2). We then describe a practical implementation based on Skew-Fit Pong et al. [2020] to achieve the objective (Section 3.3).

2. 我们提出了一种基于优化代理目标的理论驱动算法，完美对齐，使用对齐数据 (第 3.2 节)。然后，我们描述了一种基于 Skew-Fit Pong 等人 [2020] 的实际实现，以达到该目标 (第 3.3 节)。

3. Empirically, our experiments for a sawyer arm robot simulation with visual observations and goals demonstrates that our proposed method achieves state-of-the-art performance compared to data augmentation and bisimulation baselines at generalizing to unseen test environments in goal-conditioned tasks (Section 4).

3. 从经验上看，我们在带有视觉观察和目标的 sawyer 臂机器人仿真中的实验表明，我们提出的方法在目标条件任务中相较于数据增强和双重仿真基线在泛化到未见测试环境方面达到了最先进的性能 (第 4 节)。

2 Problem Formulation

2 问题表述

In this section, we formulate the domain invariant learning problem as solving Goal-conditioned Block MDPs (GBMDPs). This extends previous work on learning invariances Zhang et al. [2020a], Du et al. [2019] to the goal-conditioned setting Kaelbling [1993], Schaul et al. 2015], Marcin et al. [2017].

在本节中，我们将领域不变学习问题表述为解决目标条件块 MDP(GBMDPs)。这扩展了 Zhang 等人 [2020a]、Du 等人 [2019] 在学习不变性方面的先前工作，适用于目标条件设置 Kaelbling [1993]、Schaul 等人 [2015]、Marcin 等人 [2017]。

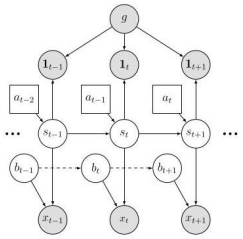


Figure 1: Graphical model for Goal-conditioned Block MDPs (GBMDPs) setting. The agent takes in the goal g and observation x_t , which is produced by the domain invariant state s_t and environmental state b_t , and acts with action a_t . Note that b_t may have temporal dependence indicated by the dashed edge.

图 1: 目标条件块 MDP(GBMDPs) 设置的图形模型。代理接收目标 g 和观察 x_t ，这些由领域不变状态 s_t 和环境状态 b_t 生成，并通过动作 a_t 进行操作。请注意，b_t 可能具有由虚线边表示的时间依赖性。

We consider a family of Goal-conditioned Block MDP environments $M^{\mathcal{E}} = \{(\mathcal{S}, \mathcal{A}, \mathcal{X}^e, \mathcal{T}^e, \mathcal{G}, \gamma) \mid e \in \mathcal{E}\}$ where e stands for the environment index. Each environment consists of shared state space $\mathcal{S}$, shared action space $\mathcal{A}$, observation space $\mathcal{X}^e$, transition dynamic $\mathcal{T}^e$, shared goal space $\mathcal{G} \subset \mathcal{S}$ and the discount factor γ .

我们考虑一系列目标条件块 MDP 环境 $M^{\mathcal{E}} = \{(\mathcal{S}, \mathcal{A}, \mathcal{X}^e, \mathcal{T}^e, \mathcal{G}, \gamma) \mid e \in \mathcal{E}\}$ ，其中 e 代表环境索引。每个环境由共享状态空间 $\mathcal{S}$ 、共享动作空间 $\mathcal{A}$ 、观察空间 $\mathcal{X}^e$ 、转移动态 $\mathcal{T}^e$ 、共享目标空间 $\mathcal{G} \subset \mathcal{S}$ 和折扣因子 γ 组成。

Moreover, we assume that $M^{\mathcal{E}}$ follows the generalized Block structure Zhang et al. [2020a]. The observation $x^e \in \mathcal{X}^e$ is determined by state $s \in \mathcal{S}$ and the environmental factor $b^e \in \mathcal{B}^e$, i.e., $x^e(s, b^e)$ (Figure 6(c)). For brevity, we use $x_t^e(s)$ to denote the observation for domain e at state s and step t . We may also omit t as $x^e(s)$ if we do not emphasize on the step t or the exact environmental factor b_t^e . The transition function is thus consists of state transition $p(s_{t+1} \mid s_t, a_t)$ (also $p(s_0)$), environmental factor transition $q^e\left(b_{t+1}^e \mid b_t^e\right)$. In our work, we assume the state transition is nearly deterministic, i.e., $\forall s, a$, entropy $\mathcal{H}(p(s_{t+1} \mid s_t, a_t)), \mathcal{H}(p(s_0)) \ll 1$, which is quite common in most RL benchmarks and applications Mnih et al. [2015], Greg et al. [2016], Pong et al. [2018]. Most importantly, $\mathcal{X}^{\mathcal{E}} = \cup_{e \in \mathcal{E}} \mathcal{X}^e$ satisfies the disjoint property Du et al. [2019], i.e., each observation $x \in \mathcal{X}^{\mathcal{E}}$ uniquely determines its underlying state s . Thus, the observation space $\mathcal{X}^{\mathcal{E}}$ can be partitioned into disjoint blocks $\mathcal{X}(s), s \in \mathcal{S}$. This assumption prevents the partial observation problem.

此外，我们假设 $M^{\mathcal{E}}$ 遵循 Zhang 等人 [2020a] 提出的广义块结构。观察 $x^e \in \mathcal{X}^e$ 由状态 $s \in \mathcal{S}$ 和环境因素 $b^e \in \mathcal{B}^e$ 决定，即 $x^e(s, b^e)$ (图 6(c))。为简便起见，我们用 $x_t^e(s)$ 表示在状态 s 和步骤 t 下的领域 e 的观察。如果我们不强调步骤 t 或确切的环境因素 b_t^e ，也可以省略 t ，用 $x^e(s)$ 表示。因此，转移函数由状态转移 $p(s_{t+1} \mid s_t, a_t)$ (也称为 $p(s_0)$) 和环境因素转移 $q^e\left(b_{t+1}^e \mid b_t^e\right)$ 组成。在我们的研究中，我们假设状态转移是近乎确定性的，即 $\forall s, a$ ，熵 $\mathcal{H}(p(s_{t+1} \mid s_t, a_t)), \mathcal{H}(p(s_0)) \ll 1$ ，这在大多数强化学习基准和应用中是相当常见的 Mnih 等人 [2015]，Greg 等人 [2016]，Pong 等人 [2018]。最重要的是，$\mathcal{X}^{\mathcal{E}} = \cup_{e \in \mathcal{E}} \mathcal{X}^e$ 满足不相交性质 Du 等人 [2019]，即每个观察 $x \in \mathcal{X}^{\mathcal{E}}$ 唯一确定其潜在状态 s 。因此，观察空间 $\mathcal{X}^{\mathcal{E}}$ 可以被划分为不相交的块 $\mathcal{X}(s), s \in \mathcal{S}$ 。这一假设防止了部分观察问题。

The objective function in GBMDP is to learn a goal-conditioned policy $\pi(a \mid x^e, g)$ that maximizes the discounted state density function $J(\pi)$ Eysenbach et al. [2020] across all domains $e \in \mathcal{E}$. In our theoretical analysis, we do not assume the exact form of g to the policy. One can regard $\pi(\cdot \mid x^e, g)$ as a group of RL policies indexed by the goal state g .

GBMDP 中的目标函数是学习一个目标条件策略 $\pi(a \mid x^e, g)$ ，该策略最大化所有领域 $e \in \mathcal{E}$ 的折扣状态密度函数 $J(\pi)$ Eysenbach 等人 [2020]。在我们的理论分析中，我们不假设 g 对策略的确切形式。可

以将 $\pi(\cdot \mid x^e, g)$ 视为一组由目标状态 g 索引的强化学习策略。

$$J(\pi) = \mathbb{E}_{e\sim\mathcal{E}, g\sim\mathcal{G}, \pi}\left[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t p_\pi^e(s_t = g \mid g)\right] = \mathbb{E}_{e\sim\mathcal{E}}\left[J^e(\pi)\right] \tag{1}$$

$p_\pi^e(s_t = g \mid g)$ denotes the probability of achieving goal g under policy $\pi(\cdot \mid x^e, g)$ at step t in domain e . Besides, $e \sim \mathcal{E}$ and $g \sim \mathcal{G}$ refers to uniform samples from each set. As p_π^e is defined over state space, it may differs among environments since policy π takes x^e as input. Fortunately, in a GBMDP, there exist optimal policies $\pi_G(\cdot \mid x^e, g)$ which are invariant over all environments, i.e.,

$p_\pi^e(s_t = g \mid g)$ 表示在领域 e 中，在步骤 t 下，策略 $\pi(\cdot \mid x^e, g)$ 实现目标 g 的概率。此外，$e \sim \mathcal{E}$ 和 $g \sim \mathcal{G}$ 指的是来自每个集合的均匀样本。由于 p_π^e 是在状态空间上定义的，它可能在不同环境中有所不同，因为策略 π 以 x^e 作为输入。幸运的是，在 GBMDP 中，存在在所有环境中不变的最优策略 $\pi_G(\cdot \mid x^e, g)$ ，即，$\pi_G(a \mid x^e(s), g) = \pi_G\left(a \mid x^{e'}(s), g\right), \forall a \in \mathcal{A}, s \in \mathcal{S}, e, e' \in \mathcal{E}$.

During training, the agent has access to training environments $\{e_i\}_{i=1}^N = \mathcal{E}_{\text{train}} \subset \mathcal{E}$ with their environment indices. However, we do not assume that $\mathcal{E}_{\text{train}}$ is i.i.d sampled from $\mathcal{E}$. Thus, we want the goal-conditioned RL agent to acquire the ability to neglect the spurious and unrelated environmental factor b^e and capture the underlying invariant state information. This setup is adopted in many recent works such as in Zhang et al. [2020a] and in domain generalization Koh et al. [2020], Arjovsky et al. [2019] for supervised learning.

在训练过程中，代理可以访问带有环境索引的训练环境 $\{e_i\}_{i=1}^N = \mathcal{E}_{\text{train}} \subset \mathcal{E}$ 。然而，我们并不假设 $\mathcal{E}_{\text{train}}$ 是从 $\mathcal{E}$ 中独立同分布 (i.i.d) 抽样的。因此，我们希望目标条件的强化学习代理能够获得忽略虚假和无关环境因素 b^e 的能力，并捕捉潜在的不变状态信息。这种设置在许多最近的工作中得到了采用，例如 Zhang 等人 [2020a] 和领域泛化 Koh 等人 [2020]、Arjovsky 等人 [2019] 的监督学习中。

3 Method

3 方法

In this section, we propose a novel learning algorithm to solve GBMDPs. First, we propose a general theory to characterize how well a policy π generalizes to unseen test environments after training on $\mathcal{E}_{\text{train}}$. Then, we introduce perfect alignment as a surrogate objective for learning. This objective is supported by the generalization theory. Finally, we propose a practical method to acquire perfect alignment.

在本节中，我们提出了一种新颖的学习算法来解决 GBMDP。首先，我们提出了一种通用理论来描述策略 π 在经过 $\mathcal{E}_{\text{train}}$ 训练后，如何有效地推广到未见的测试环境。然后，我们引入完美对齐作为学习的替代目标。该目标得到了泛化理论的支持。最后，我们提出了一种实用的方法来获得完美对齐。

3.1 Domain Generalization Theory for GBMDP

3.1 GBMDP 的领域泛化理论

In a seminal work, Ben-David et al. Ben-David et al. [2010] shows it is possible to bound the error of a classifier trained on a source domain on a target domain with a different data distribution. Follow-up work extends the theory to the domain generalization setting Sicilia et al. [2021], Albuquerque et al. 2019]. In GBMDP, we can also derive similar theory to characterize the generalization from training environments $\mathcal{E}_{\text{train}}$ to target test environment t . The theory relies on the Total Variation Distance D_{TV} Wikipedia [2021] of two policies π_1, π_2 with input (x^e, g) , which is defined as follows.

在一项开创性的工作中，Ben-David 等人 [2010] 表明，可以在目标域中对源域上训练的分类器的误差进行界定，尽管目标域的数据分布不同。后续的工作将理论扩展到领域泛化的设置 Sicilia 等人 [2021]，Albuquerque 等人 [2019]。在 GBMDP 中，我们也可以推导出类似的理论，以表征从训练环境 $\mathcal{E}_{\text{train}}$ 到目标测试环境 t 的泛化。该理论依赖于两个策略 π_1, π_2 的总变差距离 D_{TV} Wikipedia [2021]，其定义如下。

$$D_{\text{TV}}(\pi_1(\cdot \mid x^e, g) \parallel \pi_2(\cdot \mid x^e, g)) = \sup_{A' \in \sigma(\mathcal{A})} |\pi_1(A' \mid x^e, g) - \pi_2(A' \mid x^e, g)|$$

In the following statements, we denote $\rho(x, g)$ as some joint distributions of goals and observations that $g \sim \mathcal{G}$ and x is determined by $\rho(x \mid g)$. Additionally, we use $\rho_\pi^e(x^e \mid g)$ to denote the discounted

occupancy measure of x^e in environment e under policy $\pi(\cdot \mid x^e, g)$ and refer $\rho_\pi^e(x^e)$ as the marginal distribution. Furthermore, we denote $\epsilon^{\rho(x,g)}(\pi_1 \parallel \pi_2)$ as the average D_{TV} between π_1 and π_2 , i.e., $\epsilon^{\rho(x,g)}(\pi_1 \parallel \pi_2) = \mathbb{E}_{\rho(x,g)}[D_{\mathrm{TV}}(\pi_1(\cdot \mid x, g) \parallel \pi_2(\cdot \mid x, g))]$. This quantity is crucial in our theory as it can characterize the performance gap between two policies (see Appendix C).

在以下陈述中，我们将 $\rho(x,g)$ 表示为一些目标和观察的联合分布，且 $g \sim \mathcal{G}$ 和 x 由 $\rho(x \mid g)$ 决定。此外，我们使用 $\rho_\pi^e(x^e \mid g)$ 表示在策略 $\pi(\cdot \mid x^e, g)$ 下环境 e 中 x^e 的折扣占用度量，并将 $\rho_\pi^e(x^e)$ 称为边际分布。此外，我们将 $\epsilon^{\rho(x,g)}(\pi_1 \parallel \pi_2)$ 表示为 π_1 和 π_2 之间的平均值 D_{TV} ，即 $\epsilon^{\rho(x,g)}(\pi_1 \parallel \pi_2) = \mathbb{E}_{\rho(x,g)}[D_{\mathrm{TV}}(\pi_1(\cdot \mid x, g) \parallel \pi_2(\cdot \mid x, g))]$ 。这个量在我们的理论中至关重要，因为它可以表征两个策略之间的性能差距 (见附录 C)。

Then, similar to the famous $\mathcal{H}\Delta\mathcal{H}$ -divergence Ben-David et al. [2010], Sicilia et al. [2021] in domain adaptation theory, we define $\Pi\Delta\Pi$ -divergence of two joint distributions $\rho(x,g)$ and $\rho(x,g)'$ in terms of the policy class Π :

然后，类似于著名的 $\mathcal{H}\Delta\mathcal{H}$ -散度 Ben-David 等人 [2010]，Sicilia 等人 [2021] 在领域适应理论中，我们定义两个联合分布 $\rho(x,g)$ 和 $\rho(x,g)'$ 的 $\Pi\Delta\Pi$ -散度，基于策略类 Π :

$$d_{\Pi\Delta\Pi}\left(\rho(x,g), \rho(x,g)'\right) = \sup_{\pi,\pi' \in \Pi} \left| \epsilon^{\rho(x,g)}(\pi \parallel \pi') - \epsilon^{\rho(x,g)'}(\pi \parallel \pi') \right|$$

On one hand, $d_{\Pi\Delta\Pi}$ is a distance metric which reflects the distance between two distributions w.r.t function class Π . On the other hand, if we fix these two distributions, it also reveals the quality of the function class Π , i.e., smaller $d_{\Pi\Delta\Pi}$ means more invariance to the distribution change. Finally, we state the following Proposition in which π_G is some optimal and invariant policy.

一方面，$d_{\Pi\Delta\Pi}$ 是一种距离度量，反映了两个分布相对于函数类 Π 的距离。另一方面，如果我们固定这两个分布，它也揭示了函数类 Π 的质量，即，较小的 $d_{\Pi\Delta\Pi}$ 意味着对分布变化的更强不变性。最后，我们陈述以下命题，其中 π_G 是某个最优且不变的策略。

Proposition 1 (Informal). For any $\pi \in \Pi$, we consider the occupancy measure $\{\rho_\pi^{e_i}(x^{e_i}, g)\}_{i=1}^N$ for training environments and $\rho_{\pi_G}^t(x^t, g)$ for the target environment. For simplicity, we use ϵ^{e_i} as the abbreviation of $\epsilon^{\rho_\pi^{e_i}(x^{e_i},g)}, \epsilon^t$ as $\epsilon^{\rho_{\pi_G}^t(x^t,g)}$ and $\delta = \max_{e_i, e_i' \in \mathcal{E}_{\text{train}}} d_{\Pi\Delta\Pi}\left(\rho_\pi^{e_i}(x^{e_i}, g), \rho_\pi^{e_i'}\left(x^{e_i'}, g\right)\right)$. Let

命题 1(非正式)。对于任何 $\pi \in \Pi$ ，我们考虑训练环境的占用测度 $\{\rho_\pi^{e_i}(x^{e_i}, g)\}_{i=1}^N$ 和目标环境的 $\rho_{\pi_G}^t(x^t, g)$ 。为了简化，我们使用 ϵ^{e_i} 作为 $\epsilon^{\rho_\pi^{e_i}(x^{e_i},g)}, \epsilon^t$ 的缩写，表示 $\epsilon^{\rho_{\pi_G}^t(x^t,g)}$ 和 $\delta = \max_{e_i, e_i' \in \mathcal{E}_{\text{train}}} d_{\Pi\Delta\Pi}\left(\rho_\pi^{e_i}(x^{e_i}, g), \rho_\pi^{e_i'}\left(x^{e_i'},\right.\right.$ 。设

$$\lambda = \frac{1}{N}\sum_{i=1}^N \epsilon^{e_i}(\pi^* \parallel \pi_G) + \epsilon^t(\pi^* \parallel \pi_G), \ \pi^* = \arg\min_{\pi' \in \Pi} \sum_{i=1}^N \epsilon^{e_i}(\pi' \parallel \pi_G)$$

Then, we have

然后，我们有

$$J^t(\pi_G) - J^t(\pi) \leq \frac{1}{N}\sum_{i=1}^N \epsilon^{e_i}(\pi \parallel \pi_G) + \lambda + \delta + \min_{\rho(x,g) \in B} d_{\Pi\Delta\Pi}\left(\rho(x,g), \rho_{\pi_G}^t\left(x^t, g\right)\right) \tag{2}$$

where B is a characteristic set of joint distributions determined by $\mathcal{E}_{\text{train}}$ and policy class Π .

其中 B 是由 $\mathcal{E}_{\text{train}}$ 和策略类 Π 确定的联合分布的特征集。

The formal statement and the proof are shown in Appendix C.2. Generally speaking, the first term of the right hand side in Eq. 2) quantifies the performance of $\bar{\pi}$ in the N training environments. λ quantifies the optimality of the policy class Π over all environments. δ reflects how the policy class Π can reflect the difference among $\{\rho_\pi^{e_i}(x^{e_i}, g), e_i \in \mathcal{E}_{\text{train}}\}$, which should be small if the policy class is invariant. The last term characterizes the distance between training environment and target environment and will be small if the training environments are diversely distributed.

正式的陈述和证明在附录 C.2 中展示。一般来说，方程 2) 右侧的第一项量化了 $\bar{\pi}$ 在 N 训练环境中的表现。λ 量化了策略类 Π 在所有环境中的最优性。δ 反映了策略类 Π 如何反映 $\{\rho_\pi^{e_i}(x^{e_i}, g), e_i \in \mathcal{E}_{\text{train}}\}$ 之间的差异，如果策略类是不变的，这个差异应该很小。最后一项表征了训练环境和目标环境之间的距离，如果训练环境分布多样，这个距离将会很小。

Many works on domain generalization of supervised learning Ben-David et al. [2010], Liu et al. [2019], Sicilia et al. [2021], Albuquerque et al. [2019], Akuzawa et al. [2019] spend much effort in discussing the trade-offs among different terms similar to the ones in Eq. (2), e.g., minimizing δ may increase λ Akuzawa et al. [2019], and in developing sophisticated techniques to optimize the bound, e.g. distribution matching Louizos et al. [2016], Li et al. [2018], Jin et al. [2020] or adversarial learning Liu et al. [2019].

许多关于监督学习领域泛化的研究 Ben-David et al. [2010], Liu et al. [2019], Sicilia et al. [2021], Albuquerque et al. [2019], Akuzawa et al. [2019] 花费了大量精力讨论类似于公式 (2) 中不同项之间的权衡，例如，最小化 δ 可能会增加 λ Akuzawa et al. [2019]，并开发复杂的技术来优化界限，例如分布匹配 Louizos et al. [2016], Li et al. [2018], Jin et al. [2020] 或对抗学习 Liu et al. [2019]。

Different from their perspectives, in GBMDPs, we propose a simple but effective criteria to minimize the bound. From now on, we only consider the policy class $\Pi = \Pi_\Phi = \{w(\Phi(x), g), \forall w\}$. Usually, Φ will be referred as an encoder which maps $x \in \mathcal{X}^{\mathcal{E}}$ to some latent representation $z = \Phi(x)$. We will also use the notation $z(s) = \Phi(x(s))$ if we do not emphasize on the specific environment.

与他们的观点不同，在 GBMDPs 中，我们提出了一种简单但有效的标准来最小化界限。从现在开始，我们只考虑策略类 $\Pi = \Pi_\Phi = \{w(\Phi(x), g), \forall w\}$ 。通常，Φ 将被称为编码器，它将 $x \in \mathcal{X}^{\mathcal{E}}$ 映射到某种潜在表示 $z = \Phi(x)$ 。如果我们不强调特定环境，我们还将使用符号 $z(s) = \Phi(x(s))$ 。

Definition 1 (Perfect Alignment). An encoder is called a perfect alignment encoder Φ w.r.t environment set E if $\forall e, e' \in E$ and $\forall s, s' \in \mathcal{S}, \Phi(x^e(s)) = \Phi\left(x^{e'}(s')\right)$ if and only if $s = s'$.

定义 1(完美对齐)。如果编码器相对于环境集 E 是完美对齐编码器 Φ ，当且仅当 $\forall e, e' \in E$ 和 $\forall s, s' \in \mathcal{S}, \Phi(x^e(s)) = \Phi\left(x^{e'}(s')\right)$ 时，成立 $s = s'$ 。

As illustrated in Figure 5, an encoder is in perfect alignment if it maps two observations of the same underlying state s to the same latent encoding $z(s)$ while also preventing meaningless embedding, i.e., mapping observations of different states to the same z . We believe perfect alignment plays an important role in domain generalization for goal-conditioned RL agents. Specifically, it can minimize the bound of Eq. 2 as follows.

如图 5 所示，如果编码器将同一潜在状态 s 的两个观察映射到相同的潜在编码 $z(s)$ ，同时防止无意义的嵌入，即将不同状态的观察映射到相同的 z ，则该编码器处于完美对齐状态。我们相信完美对齐在目标条件强化学习代理的领域泛化中发挥着重要作用。具体而言，它可以如下最小化公式 2 的界限。

Proposition 2 (Informal). If the encoder Φ is a perfect alignment over $\mathcal{E}_{\text{train}}$, then

命题 2(非正式)。如果编码器 Φ 在 $\mathcal{E}_{\text{train}}$ 上是完美对齐的，那么

$$J^t(\pi_G) - J^t(\pi) \leq \underbrace{\frac{1}{N}\sum_{i=1}^{N} \epsilon^{e_i}(\pi \parallel \pi_G)}_{(E)} + \underbrace{\epsilon^t(\pi^* \parallel \pi_G) + d_{\Pi_\Phi} \Delta\Pi_\Phi\left(\widetilde{\rho}(x, g), \rho^t_{\pi_G}\left(x^t, g\right)\right)}_{(t)} \tag{3}$$

where $\widetilde{\rho}(x, g)$ and π^* are defined in Proposition 1 (also Appendix C).

其中 $\widetilde{\rho}(x, g)$ 和 π^* 在命题 1(也见附录 C) 中定义。

In Appendix C. 3, we formally prove Proposition 2 when Φ is a (η, ψ) -perfect alignment, i.e., Φ is only near perfect alignment. The proof shows that the generalization error bound is minimized on the R.H.S of Eq. (3) when Φ asymptotically becomes an exact perfect alignment encoder. Therefore, in our following method, we aim to learn a perfect alignment encoder via aligned sampling (Section 3.2).

在附录 C.3 中，我们在 Φ 是 (η, ψ) -完美对齐时正式证明了命题 2，即 Φ 仅是近乎完美对齐。证明表明，当 Φ 渐近地变为一个精确的完美对齐编码器时，R.H.S 的 Eq. (3) 上的泛化误差界限被最小化。因此，在我们接下来的方法中，我们旨在通过对齐采样 (第 3.2 节) 学习一个完美对齐编码器。

For the remaining terms in the R.H.S of Eq. (3), we find it hard to quantify them task agnostically, as similar difficulties also exist in the domain generalization theory of supervised learning Sicilia et al. [2021]. Fortunately, we can derive upper bounds for the remaining terms under certain assumptions and we observe that these upper bounds are significantly reduced via our method in the experiments (Section 4). The(E)term represents how well the learnt policy π approximates the optimal invariant policy on the training environments and is reduced to almost zero via RL (Table 1). For the (t) term, we show that an upper bound of(t)is proportion to the invariant quality of $\bar{\Phi}$ on the target environment. Moreover, we find that learning a perfect alignment encoder over $\mathcal{E}_{\text{train}}$ empirically improves the invariant quality over other unseen environments(t)(Figure 4). Thus, this(t)term upperbound is reduced by learning perfect alignment. Please refer to Appendix C. 4 for more details.

对于 Eq. (3) 的 R.H.S 中剩余的项，我们发现很难以任务无关的方式量化它们，因为在监督学习的领域泛化理论中也存在类似的困难 Sicilia et al. [2021]。幸运的是，我们可以在某些假设下推导出剩余项的上界，并且我们观察到这些上界在实验中通过我们的方法显著降低 (第 4 节)。(E) 项表示学习到的策略 π 在训练环境中对最优不变策略的近似程度，并通过强化学习 (表 1) 几乎降低到零。对于 (t) 项，我们展示了 (t) 的上界与目标环境中 $\bar{\Phi}$ 的不变质量成正比。此外，我们发现，在 $\mathcal{E}_{\text{train}}$ 上学习一个完美对齐编码器在经验上改善了其他未见环境上的不变质量 (t)(图 4)。因此，通过学习完美对齐，这个 (t) 项的上界被降低。有关更多细节，请参见附录 C.4。

Based on the theory we derived in this subsection, we adopt perfect alignment as the heuristic to address GBMDPs in our work. In the following subsections, we propose a practical method to acquire a perfect alignment encoder over the training environments.

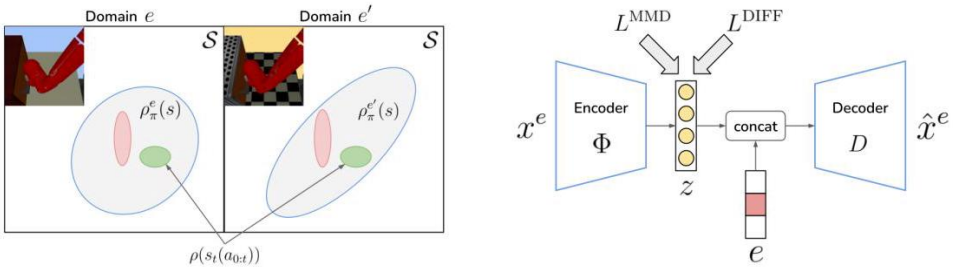
基于我们在本小节中推导的理论，我们采用完美对齐作为启发式方法来解决我们的工作中的 GB-MDPs。在接下来的小节中，我们提出了一种实用的方法来获取训练环境中的完美对齐编码器。

3.2 Learning Domain Invariant via Aligned Sampling

3.2 通过对齐采样学习领域不变性

First, we discuss about the if condition on perfect alignment encoder Φ , i.e., $\forall s, \Phi\left(x^{e}(s)\right)=\Phi\left(x^{e^{\prime}}(s)\right)$. The proposed method is based on aligned sampling. In contrast, most RL algorithms use observation-dependent sampling from the environment, e.g., 6-greedy or Gaussian distribution policies Haarnoja et al. [2018], Fujimoto et al. [2018], Pong et al. [2020], Mnih et al. [2015]. However,

首先，我们讨论完美对齐编码器的 if 条件 Φ ，即 $\forall s, \Phi\left(x^{e}(s)\right)=\Phi\left(x^{e^{\prime}}(s)\right)$ 。所提出的方法基于对齐采样。相比之下，大多数强化学习算法使用依赖于观察的环境采样，例如 6-greedy 或高斯分布策略 Haarnoja 等 [2018]，Fujimoto 等 [2018]，Pong 等 [2020]，Mnih 等 [2015]。然而，



(a) Illustration of Aligned Sampling (b) Overall structure

Figure 2: (a): Illustration of Aligned Sampling. Square represents the whole state space $\mathcal{S}$, gray area represents the distribution $\rho_{\pi}^{e}(s)$ in two different environments. Small colored areas are the aligned state distribution generated by aligned sampling in Section 3.2 (b): Overall VAE structure in our PA-SF. Encoder maps x^e to the latent embedding z and decoder D reconstructs the observations with z and index e.L^{MMD} and L^{DIFF} denote the two losses in Section 3.2

图 2: (a): 对齐采样的示意图。方框代表整个状态空间 $\mathcal{S}$ ，灰色区域代表两个不同环境中的分布 $\rho_{\pi}^{e}(s)$ 。小的彩色区域是通过第 3.2 节中的对齐采样生成的对齐状态分布 (b): 我们的 PA-SF 中的整体 VAE 结构。编码器将 x^e 映射到潜在嵌入 z ，解码器 D 使用 z 重构观察，索引 e.L^{MMD} 和 L^{DIFF} 表示第 3.2 节中的两个损失。

with observation-dependent sampling, occupancy measures $\rho_{\pi}^{e}(s), \forall e \in \mathcal{E}_{\text{train}}$ will be different. Thus, simply aligning the latent representation of these observations will fail to produce a perfect alignment encoder Φ .

使用依赖于观察的采样，占用度量 $\rho_{\pi}^{e}(s), \forall e \in \mathcal{E}_{\text{train}}$ 将会不同。因此，仅仅对齐这些观察的潜在表示将无法产生完美对齐编码器 Φ 。

Thus, we propose a novel strategy for data collection called aligned sampling. First, we randomly select a trajectory (e.g., from replay buffer etc.), denoted as $\{x_0^e, a_0, x_1^e, a_1, \ldots, x_T^e\}$ from environment e . The set of corresponding states along this trajectory are denoted as $\{s_t^e(a_{0:t})\}_{t=0}^{T}$. Second, we take the same action sequence $a_{0:T}$ in another domain e' to get another trajectory $\left\{x_0^{e'}, a_0, x_1^{e'}, a_1, \ldots, x_T^{e'}\right\}$ (so as $\left\{s_t^{e'}(a_{0:t})\right\}_{t=0}^{T}$). We refer to the data collected by aligned sampling from all training environments as aligned data. These aligned observations $\{x_t^{e_i}(a_{0:t})\}, \forall e_i \in \mathcal{E}_{\text{train}}$ are stored in an aligned buffer $\mathcal{R}_{\text{align}}$ corresponding to the aligned action sequence $a_{0:t}$.

因此，我们提出了一种称为对齐采样的新颖数据收集策略。首先，我们随机选择一个轨迹 (例如，从重放缓冲区等)，记作 $\{x_0^e, a_0, x_1^e, a_1, \ldots, x_T^e\}$ 来自环境 e 。沿此轨迹的对应状态集合记作 $\{s_t^e(a_{0:t})\}_{t=0}^{T}$ 。其次，我们在另一个领域 e' 中采取相同的动作序列 $a_{0:T}$ 以获得另一条轨迹 $\left\{x_0^{e'}, a_0, x_1^{e'}, a_1, \ldots, x_T^{e'}\right\}$ (同

样适用于 $\left\{s_t^{e'}\left(a_{0:t}\right)\right\}_{t=0}^{T}$)。我们将从所有训练环境中通过对齐采样收集的数据称为对齐数据。这些对齐观察 $\left\{x_t^{e_i}\left(a_{0:t}\right)\right\}, \forall e_i \in \mathcal{E}_{\text{train}}$ 被存储在与对齐动作序列 $a_{0:t}$ 相对应的对齐缓冲区 $\mathcal{R}_{\text{align}}$ 中。

Under the definition of GBMDP, we have $\forall t \in [0:T], s \in \mathcal{S}, \rho\left(s_t^e\left(a_{0:t}\right)\right) = \rho\left(s_t^{e'}\left(a_{0:t}\right)\right)$, i.e., the same state distribution. Therefore, we can use MMD loss Gretton et al. [2008] to match distribution of $\Phi\left(x^e\left(s\right)\right)$ for the aligned data. More specifically, in each iteration, we sample a mini-batch of B aligned observations of every training environment $e_i \in \mathcal{E}_{\text{train}}$ from $\mathcal{R}_{\text{align}}$, i.e., $\mathcal{B}_{\text{align}} = \left\{x^{e_i}\left(s_t^{e_i}\left(a_{0:t}^b\right)\right), \forall e_i \in \mathcal{E}_{\text{train}}\right\}_{b=1}^{B}$. Then we use the following loss as a computationally efficient approximation of the MMD metric Zhao and Meng [2015], Louizos et al. [2016].

在 GBMDP 的定义下，我们有 $\forall t \in [0:T], s \in \mathcal{S}, \rho\left(s_t^e\left(a_{0:t}\right)\right) = \rho\left(s_t^{e'}\left(a_{0:t}\right)\right)$ ，即相同的状态分布。因此，我们可以使用 MMD 损失 Gretton 等 [2008] 来匹配对齐数据的 $\Phi\left(x^e\left(s\right)\right)$ 的分布。更具体地说，在每次迭代中，我们从 $\mathcal{R}_{\text{align}}$ 中抽取每个训练环境 $e_i \in \mathcal{E}_{\text{train}}$ 的一小批 B 对齐观察，即 $\mathcal{B}_{\text{align}} = \left\{x^{e_i}\left(s_t^{e_i}\left(a_{0:t}^b\right)\right), \forall e_i \in \mathcal{E}_{\text{train}}\right\}_{b=1}^{B}$ 。然后我们使用以下损失作为 MMD 度量的计算高效近似 Zhao 和 Meng [2015]，Louizos 等 [2016]。

$$L^{\text{MMD}}\left(\Phi\right) = \mathbb{E}_{e,e' \sim \mathcal{E}_{\text{train}}, \mathcal{B}_{\text{align}}}\left[\left\| \frac{1}{B}\sum_{b=1}^{B}\psi\left(\Phi\left(x^e\left(s_t^e\left(a_{0:t}^b\right)\right)\right)\right) - \frac{1}{B}\sum_{b=1}^{B}\psi\left(\Phi\left(x^{e'}\left(s_t^{e'}\left(a_{0:t}^b\right)\right)\right)\right) \|_2^2 \right\|^2\right]$$

where ψ is a random expansion function.

其中 ψ 是一个随机扩展函数。

In Figure 2(a), we illustrate the intuition of the above approach. When the transition is nearly deterministic, the entropy for $\rho\left(s_t^e\left(a_{0:t}\right)\right)$ is much smaller, i.e., $\mathcal{H}\left(\rho\left(s_t^e\left(a_{0:t}\right)\right)\right) \ll \mathcal{H}\left(\rho_\pi^e\left(s_t\right)\right)$. Thus, $\rho\left(s_t^e\left(a_{0:t}\right)\right)$ can be regarded as small patches in $\mathcal{S}$. We use the MMD loss L^{MMD} to match the latent representation $\left\{\Phi\left(x^e\left(s\right)\right), s \sim \rho\left(s_t^e\left(a_{0:t}\right)\right)\right\}, \forall e \in \mathcal{E}_{\text{train}}$ together. As a consequence, we should achieve an encoder Φ that is more aligned. We discuss the theoretical property of L^{MMD} in detail in Appendix C. 5

在图 2(a) 中，我们说明了上述方法的直觉。当转移几乎是确定性的时，$\rho\left(s_t^e\left(a_{0:t}\right)\right)$ 的熵要小得多，即 $\mathcal{H}\left(\rho\left(s_t^e\left(a_{0:t}\right)\right)\right) \ll \mathcal{H}\left(\rho_\pi^e\left(s_t\right)\right)$ 。因此，$\rho\left(s_t^e\left(a_{0:t}\right)\right)$ 可以被视为 $\mathcal{S}$ 中的小补丁。我们使用 MMD 损失 L^{MMD} 来将潜在表示 $\left\{\Phi\left(x^e\left(s\right)\right), s \sim \rho\left(s_t^e\left(a_{0:t}\right)\right)\right\}, \forall e \in \mathcal{E}_{\text{train}}$ 结合在一起。因此，我们应该实现一个更对齐的编码器 Φ 。我们在附录 C.5 中详细讨论了 L^{MMD} 的理论性质。

However, simply minimizing L^{MMD} may violate the only if condition for perfect alignment. For example, a trivial solution for $L^{\text{MMD}} = 0$ is mapping all observations to some constant latent. To ensure that $\Phi\left(x^e\left(s\right)\right) = \Phi\left(x^{e'}\left(s'\right)\right)$ only if $s = s'$, we additionally use the difference loss L^{DIFF} as follows.

然而，简单地最小化 L^{MMD} 可能会违反完美对齐的仅当条件。例如，$L^{\text{MMD}} = 0$ 的一个平凡解是将所有观察映射到某个常量潜变量。为了确保 $\Phi\left(x^e\left(s\right)\right) = \Phi\left(x^{e'}\left(s'\right)\right)$ 仅当 $s = s'$ 时，我们还使用差异损失 L^{DIFF} 如下所示。

$$L^{\text{DIFF}}\left(\Phi\right) = -\mathbb{E}_{e \sim \mathcal{E}_{\text{train}}, x^e, \widetilde{x}^e \in \mathcal{R}^e}\left\|\Phi\left(x^e\right) - \Phi\left(\widetilde{x}^e\right)\right\|_2^2$$

where $\mathcal{R}^e$ refers to the replay buffer of environment e . Clearly, minimizing L^{DIFF} encourages dispersed latent representations over all states $s \in \mathcal{S}$.

其中 $\mathcal{R}^e$ 指的是环境 e 的重放缓冲区。显然，最小化 L^{DIFF} 鼓励在所有状态 $s \in \mathcal{S}$ 上分散的潜在表示。

We refer to the combination $\alpha_{\text{MMD}} L^{\text{MMD}} + \alpha_{\text{DIFF}} L^{\text{DIFF}}$ as our perfect alignment loss L^{PA} . Note that L^{PA} resembles contrastive learning Chen et al. [2020], Laskin et al. [2020a]. Namely, observations of aligned data from $\mathcal{R}_{\text{align}}$ are positive pairs while observations sampled randomly from a big replay buffer are negative pairs. We match the latent embedding of positive pairs via the MMD loss while separating negative pairs via the difference loss. As discussed in Section 3.1, we believe this latent representation will improve generalization to unseen target environments.

我们将组合 $\alpha_{\text{MMD}} L^{\text{MMD}} + \alpha_{\text{DIFF}} L^{\text{DIFF}}$ 称为我们的完美对齐损失 L^{PA} 。请注意，L^{PA} 类似于对比学习 Chen 等人 [2020]，Laskin 等人 [2020a]。即，从 $\mathcal{R}_{\text{align}}$ 中对齐数据的观察是正对，而从一个大型重放缓冲区随机抽样的观察是负对。我们通过 MMD 损失匹配正对的潜在嵌入，同时通过差异损失分离负对。如第 3.1 节所讨论，我们相信这种潜在表示将改善对未见目标环境的泛化能力。

3.3 Perfect Alignment for Skew-Fit

3.3 偏斜拟合的完美对齐

In Section 4, we will train goal-conditioned RL agents with perfect alignment encoder using the Skew-Fit algorithm Pong et al. [2020]. Skew-Fit is typically designed for visual-input agents which learn a goal-conditioned policy via purely self-supervised learning.

在第 4 节中，我们将使用偏斜拟合算法 Pong 等人 [2020] 训练具有完美对齐编码器的目标条件强化学习代理。偏斜拟合通常是为视觉输入代理设计的，这些代理通过纯自监督学习学习目标条件策略。

First, Skew-Fit trains a β -VAE with observations collected online to acquire a compact and meaningful latent representation for each state, i.e., $z(s)$ from the image observations $x(s)$. Then, Skew-Fit optimizes a SAC Haarnoja et al. [2018] agent in the goal-conditioned setting over the latent embedding of the image observation and goal, $\pi(a \mid z, g)$. The reward function is the negative of l_2 distance between the two latent representation $z(s)$ and $z(g)$, i.e., $r(s, g) = -\parallel z(s) - z(g) \parallel_2$. Furthermore, to improve sample efficiency, Skew-Fit proposes skewed sampling for goal-conditioned exploration.

首先，Skew-Fit 训练一个 β -VAE，通过在线收集的观察数据，为每个状态获取一个紧凑且有意义的潜在表示，即来自图像观察的 $z(s)$ $x(s)$ 。然后，Skew-Fit 在潜在嵌入的图像观察和目标上优化一个 SAC Haarnoja 等 [2018] 代理，以实现目标条件设置 $\pi(a \mid z, g)$ 。奖励函数是两个潜在表示之间的 l_2 距离的负值 $z(s)$ 和 $z(g)$ ，即 $r(s, g) = -\parallel z(s) - z(g) \parallel_2$ 。此外，为了提高样本效率，Skew-Fit 提出了倾斜采样以进行目标条件探索。

In our algorithm, Perfect Alignment for Skew-Fit (PA-SF), the encoder Φ is optimized via both β -VAE losses as Pong et al. [2020], Nair et al. [2018] and L^{PA} loss to ensure meaningful and perfectly aligned latent representation.

在我们的算法中，Skew-Fit 的完美对齐 (PA-SF)，编码器 Φ 通过 β -VAE 损失，如 Pong 等 [2020]、Nair 等 [2018] 以及 L^{PA} 损失进行优化，以确保有意义且完美对齐的潜在表示。

$$L(\Phi, D) = L^{\mathrm{RECON}}\left(x^e, \widehat{x^e}\right) + \beta D_{\mathrm{KL}}\left(q_\Phi\left(z \mid x^e\right) \parallel p(z)\right) + \alpha_{\mathrm{MMD}} L^{\mathrm{MMD}} + \alpha_{\mathrm{DIFF}} L^{\mathrm{DIFF}} \tag{4}$$

In addition, we use both aligned sampling and observation-dependent sampling. Aligned sampling provides aligned data but hurts sample-efficiency while observation-dependent sampling is exploration-efficient but fails to ensure alignment. In practice, we find that collecting a small portion (15% of all data collected) of aligned data in $\mathcal{R}_{\text{align}}$ is enough for perfect alignment via L^{PA} .

此外，我们使用对齐采样和观察依赖采样。对齐采样提供对齐数据，但会影响样本效率，而观察依赖采样则具有探索效率，但无法确保对齐。在实践中，我们发现在 $\mathcal{R}_{\text{align}}$ 中收集一小部分 (占所有收集数据的 15%) 的对齐数据就足以通过 L^{PA} 实现完美对齐。

Additionally, inspired by Louizos et al. [2016], we also change the β -VAE structure to what is shown in Figure 2(b), since in GBMDP data are collected from N training environments and thus, the identity Gaussian distribution is no longer a proper fit for prior. The encoder Φ maps $x^e(s)$ to some latent representation $z(s)$ while the decoder D takes both $z(s)$ and the environment index e as input to reconstruct $\widehat{x^e}(s)$. Note that by using both L^{PA} and L^{RECON} , we require static environmental factors in $\mathcal{E}_{\text{train}}$ (unnecessary for testing environments) for a stable optimization. In future work, we will address the limit from β -VAE by training two latent representations simultaneously to stabilize the optimization for generality.

此外，受到 Louizos 等人 [2016] 的启发，我们还将 β -VAE 结构更改为图 2(b) 中所示的形式，因为在 GBMDP 中，数据是从 N 训练环境中收集的，因此，单位高斯分布不再适合作为先验。编码器 Φ 将 $x^e(s)$ 映射到某些潜在表示 $z(s)$ ，而解码器 D 则将 $z(s)$ 和环境索引 e 作为输入来重构 $\widehat{x^e}(s)$ 。请注意，通过同时使用 L^{PA} 和 L^{RECON} ，我们要求在 $\mathcal{E}_{\text{train}}$ 中具有静态环境因素 (在测试环境中不必要)，以实现稳定的优化。在未来的工作中，我们将通过同时训练两个潜在表示来解决 β -VAE 的限制，以稳定优化以实现通用性。

4 Experiments

4 实验

In this section, we conduct experiments to evaluate our PA-SF algorithms. The experiments are based on multiworld Pong et al. [2018]. Our empirical analysis tries to answer the following questions: (1) How well does PA-SF perform in solving GBMDP problems? (2) How does each component proposed in Section 3 contribute to the performance?

在本节中，我们进行实验以评估我们的 PA-SF 算法。实验基于多世界 Pong 等人 [2018]。我们的实证分析试图回答以下问题:(1)PA-SF 在解决 GBMDP 问题方面表现如何？(2) 第 3 节中提出的每个组件对性能的贡献如何？

4.1 Comparative Evaluation

4.1 比较评估

In this subsection, we aim to answer the question (1) by comparing our proposed PA-SF method with vanilla Skew-Fit and several other baselines that attempt to acquire invariant policies for RL agents.

在本小节中，我们旨在通过将我们提出的 PA-SF 方法与普通的 Skew-Fit 以及其他几个试图为 RL 代理获取不变策略的基线进行比较，来回答问题 (1)。

Baselines Current methods for obtaining robust policies can be characterized into two categories: (1) data augmentation and (2) model bisimulation.

基线当前获取鲁棒策略的方法可以分为两类:(1) 数据增强和 (2) 模型双重仿真。

1. Data Augmentation. Recent work Stone et al. [2021] tries to use data augmentation to prevent the RL agents from distractions. We implement the most widely accepted data augmentation methods RAD Laskin et al. [2020b] upon Skew-Fit (Skew-Fit + RAD) as a baseline. Note that our PA-SF method does not use any data augmentation and is parallel with this kind of techniques.

1. 数据增强。最近的研究 Stone et al. [2021] 尝试使用数据增强来防止强化学习代理的干扰。我们在 Skew-Fit 的基础上实现了最广泛接受的数据增强方法 RAD Laskin et al. [2020b](Skew-Fit + RAD)。请注意，我们的 PA-SF 方法不使用任何数据增强，并与这类技术并行。

2. Model Bisimulation Ferns et al. [2011]. These methods utilize bisimulation metrics to learn a minimal but sufficient representation which will neglect irrelevant features of Block MDPs. We include MISA Zhang et al. [2020a] and DBC Zhang et al. [2020b] in our comparison as they are the most successful implementations for high-dimensional tasks. Moreover, in the goal-conditioned setting, we use an oracle state-goal distance $-\parallel s-g\parallel_2$ as rewards for these two algorithms in GBMDP. In contrast, our PA-SF method does not have such information.

2. 模型双模拟 Ferns et al. [2011]。这些方法利用双模拟度量来学习一个最小但足够的表示，从而忽略 Block MDPs 的无关特征。我们在比较中包括了 MISA Zhang et al. [2020a] 和 DBC Zhang et al. [2020b]，因为它们是高维任务中最成功的实现。此外，在目标条件设置中，我们使用一个 oracle 状态-目标距离 $-\parallel s-g\parallel_2$ 作为这两个算法在 GBMDP 中的奖励。相比之下，我们的 PA-SF 方法没有这样的信息。

Table 1: Evaluation of PA-SF and baselines on four control tasks. We report the mean and one standard deviation on each task (lower metric is better).

表 1:PA-SF 和基线在四个控制任务上的评估。我们报告每个任务的均值和一个标准差 (较低的指标更好)。

	Algorithm	Reach	Door	Push	Pickup
		Hand distance (35K)	Angle difference (150K)	Puck distance (400K)	Object distance (280K)
Test Avg	Skew-Fit	0.111 ± 0.010	0.194 ± 0.018	0.086 ± 0.004	0.037 ± 0.006
	Skew-Fit + RAD	0.105 ± 0.010	0.162 ± 0.030	0.082 ± 0.008	0.040 ± 0.004
	MISA	0.239 ± 0.0142	0.255 ± 0.027	0.099 ± 0.006	0.043 ± 0.004
	DBC	0.185 ± 0.037	0.320 ± 0.033	0.095 ± 0.006	0.045 ± 0.002
	PA-SF(Ours)	0.076 ± 0.005	0.106 ± 0.015	0.069 ± 0.005	0.028 ± 0.004
Train Avg	PA-SF(Ours)	0.067 ± 0.005	0.058 ± 0.074	0.060 ± 0.005	0.020 ± 0.008
	Oracle Skew-Fit	0.055 ± 0.010	0.057 ± 0.012	0.054 ± 0.006	0.020 ± 0.006

	算法	到达	门	推	拾取
		手距 (35K)	角度差 (150K)	冰球距离 (400K)	物体距离 (280K)
测试平均值	偏斜拟合	0.111 ± 0.010	0.194 ± 0.018	0.086 ± 0.004	0.037 ± 0.006
	偏斜拟合 + RAD	0.105 ± 0.010	0.162 ± 0.030	0.082 ± 0.008	0.040 ± 0.004
	MISA	0.239 ± 0.0142	0.255 ± 0.027	0.099 ± 0.006	0.043 ± 0.004
	DBC	0.185 ± 0.037	0.320 ± 0.033	0.095 ± 0.006	0.045 ± 0.002
	PA-SF(我们的)	0.076 ± 0.005	0.106 ± 0.015	0.069 ± 0.005	0.028 ± 0.004
训练平均值	PA-SF(我们的)	0.067 ± 0.005	0.058 ± 0.074	0.060 ± 0.005	0.020 ± 0.008
	预言者偏斜拟合	0.055 ± 0.010	0.057 ± 0.012	0.054 ± 0.006	0.020 ± 0.006

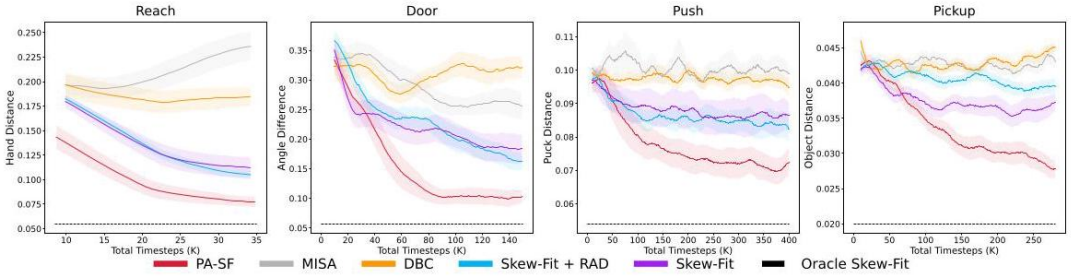


Figure 3: Learning curve of all algorithms on average across test environments for each task. All curves show the mean and one standard deviation (a half for Pickup to show clearly) of 7 seeds.

图 3: 所有算法在每个任务的测试环境中的平均学习曲线。所有曲线显示了 7 个种子的均值和一个标准差 (为了清晰起见，Pickup 的标准差为一半)。

Environments We evaluate PA-SF and all baselines on a set of GBMDP tasks based on multiworld benchmark Pong et al. [2018], which is widely used to evaluate the performance of visual input goal-conditioned algorithms. We use the following four basic tasks Nair et al. [2018], Pong et al. [2020]: Reach, Door, Pickup and Push. In GBMDP, we create different environments with various backgrounds, desk surfaces, and object appearances. During testing, we also create environments with unseen video backgrounds to mimic environmental factor transitions $q^e\left(b^e_{t+1} \mid b^e_t\right)$. This makes policy generalization more challenging. Please refer to Appendix E for a full description of our experiment setup and implementation details of the baselines and our algorithm.

我们在一组基于多世界基准的 GBMDP 任务上评估 PA-SF 和所有基线，这些基准由 Pong 等人 [2018] 提出，广泛用于评估视觉输入目标条件算法的性能。我们使用以下四个基本任务 Nair 等人 [2018], Pong 等人 [2020]: 到达、门、拾取和推。在 GBMDP 中，我们创建具有不同背景、桌面表面和物体外观的不同环境。在测试期间，我们还创建具有未见视频背景的环境，以模拟环境因素的转变 $q^e\left(b^e_{t+1} \mid b^e_t\right)$ 。这使得策略泛化变得更加具有挑战性。有关我们实验设置和基线及我们算法的实现细节的完整描述，请参见附录 E。

Results In Table 1, we show the final average performance of each algorithm on unseen test environments $\mathcal{E}_{\text{test}}$. The corresponding learning curves are shown in Figure 3 . This metric shows the generalizability of each RL agent. All these agents are allowed to collect data from $\mathcal{E}_{\text{train}}$ $(N = 3)$ with static environmental factors. Our PA-SF achieves SOTA performance on all tasks. On testing environments, we achieve a relative reduction around 40% to 65% of the corresponding metrics over vanilla Skew-Fit w.r.t the optimal metric possible (Oracle Skew-Fit). Oracle Skew-Fit refers to the performance of a Skew-Fit algorithm trained directly on the single environment (and not simultaneously on all $\mathcal{E}_{\text{train}}$).

在表 1 中，我们展示了每个算法在未见测试环境上的最终平均性能 $\mathcal{E}_{\text{test}}$ 。相应的学习曲线如图 3 所示。该指标显示了每个 RL 代理的泛化能力。所有这些代理都被允许从 $\mathcal{E}_{\text{train}}$ $(N = 3)$ 中收集数据，环境因素保持静态。我们的 PA-SF 在所有任务上都达到了 SOTA 性能。在测试环境中，我们在相应指标上相对于普通 Skew-Fit 实现了大约 40% 到 65% 的相对减少，针对可能的最佳指标 (Oracle Skew-Fit)。Oracle Skew-Fit 指的是在单一环境上直接训练的 Skew-Fit 算法的性能 (而不是同时在所有 $\mathcal{E}_{\text{train}}$ 上训练)。

Other invariant policy learning methods perform sluggishly on all tasks. For DBC and MISA, we hypothesize that they struggle for goal-conditioned problems since the model bisimulation metric is defined for a single MDP. In GBMDPs, this means acquiring a set of encoders Φ_g that achieves model bisimulation for every possible g and is thus inefficient for learning. By design, our method is not susceptible to this issue as the perfect alignment is a universal invariant representation for all goals. Data augmentation via RAD provides marginal improvement over the vanilla Skew-Fit. Nevertheless, we believe developing adequate data augmentation techniques for GBMDPs is an important research problem and is orthogonal with our method.

其他不变政策学习方法在所有任务上表现缓慢。对于 DBC 和 MISA，我们假设它们在目标条件问题上表现不佳，因为模型双重仿真度量是为单个 MDP 定义的。在 GBMDPs 中，这意味着需要获取一组编码器 Φ_g ，以实现对每个可能的 g 的模型双重仿真，因此在学习上效率低下。根据设计，我们的方法不易受到此问题的影响，因为完美对齐是所有目标的通用不变表示。通过 RAD 进行的数据增强对普通 Skew-Fit 提供了边际改进。尽管如此，我们认为为 GBMDPs 开发适当的数据增强技术是一个重要的研究问题，并且与我们的方法是正交的。

Additionally, we also show the performance of PA-SF on the training environments in Table II PA-SF is still comparable and as sample-efficient as Skew-Fit in the training environments. This supports the

claim that the(E)term in the R.H.S of Eq. (3) is reduced to almost zero via RL training in practice.

此外，我们还展示了 PA-SF 在训练环境中的性能，如表 II 所示。PA-SF 在训练环境中仍然与 Skew-Fit 相当，并且样本效率相当。这支持了这样的说法: 在实际中，方程 (3) 右侧的 (E) 项通过 RL 训练几乎减少到零。

4.2 Design Evaluation

4.2 设计评估

In this subsection, we conduct comprehensive analysis on the design of PA-SF to interpret how well it carries out the theoretical framework discussed in Section 3.1 and Section 3.2

在本小节中，我们对 PA-SF 的设计进行了全面分析，以解释它在多大程度上执行了第 3.1 节和第 3.2 节中讨论的理论框架。

To begin with, we show the learning curves in Figure 4 of different ablations of PA-SF in the Door environment during both training and testing. To understand the roles of L^{DIFF} and L^{MMD} , PA-SF (w/o D) excludes L^{DIFF} and PA-SF (w/o MD) excludes both losses I Noticing that PA-SF (w/o MD) is equivalent to the Skew-Fit algorithm with our proposed VAE structure (Figure 2(b)). We also add PA-SF (w/o AS) which excludes aligned sampling. represent the mean and one standard deviation across 7 seeds.

首先，我们展示了在 Door 环境中 PA-SF 不同消融实验的学习曲线，如图 4 所示，涵盖了训练和测试阶段。为了理解 L^{DIFF} 和 L^{MMD} 的作用，PA-SF (w/o D) 排除了 L^{DIFF} ，而 PA-SF (w/o MD) 排除了两个损失 I 注意到 PA-SF (w/o MD) 等同于具有我们提出的 VAE 结构的 Skew-Fit 算法 (图 2(b))。我们还添加了 PA-SF (w/o AS)，它排除了对齐采样。表示 7 个种子的均值和一个标准差。

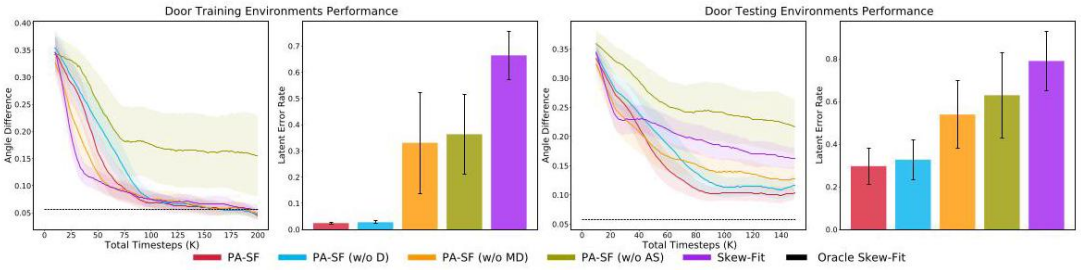


Figure 4: Ablation of PA-SF and visualization of the latent representation via LER metric. All curves

图 4:PA-SF 的消融实验及通过 LER 指标可视化潜在表示。所有曲线

Additionally, we also quantify the quality of the latent representation $\Phi\left(x^{e}\left(s\right)\right)$ in Figure 4 via the metric Latent Error Rate (LER). LER is defined as the average over environment set $E \in \{\mathcal{E}_{\text{train}}, \mathcal{E}_{\text{test}}\}$ as follows:

此外，我们还通过潜在错误率 (LER) 指标在图 4 中量化潜在表示的质量 $\Phi\left(x^{e}\left(s\right)\right)$ 。LER 定义为环境集 $E \in \{\mathcal{E}_{\text{train}}, \mathcal{E}_{\text{test}}\}$ 的平均值，如下所示:

$$\operatorname{Err}\left(\Phi\right) = \mathbb{E}_{e\sim E, s\sim\mathcal{S}}\left[\frac{\left\|\Phi\left(x^{e}\left(s\right)\right) - \Phi\left(x^{e_0}\left(s\right)\right)\right\|_2}{\left\|\Phi\left(x^{e}\left(s\right)\right)\right\|_2}\right]$$

In general, the smaller $\operatorname{Err}\left(\Phi\right)$ is, the closer the encoder Φ is to perfect alignment over the environments E . We first focus on the discussion about training performance.

一般来说，$\operatorname{Err}\left(\Phi\right)$ 越小，编码器 Φ 在环境 E 上的完美对齐程度就越高。我们首先关注训练性能的讨论。

1. Φ achieves the if condition of perfect alignment over $\mathcal{E}_{\text{train}}$ via L^{MMD} as the LER value of PA-SF and PA-SF (w/o D) is almost 0 . While without MMD loss, PA-SF (w/o MD) and Skew-Fit struggle with large LER value despite achieving good training performance. Furthermore, the comparison between PA-SF and PA-SF (w/o AS) demonstrates the importance of using aligned data in the MMD loss (Otherwise, the matching is inherently erroneous).

1. Φ 实现了在 $\mathcal{E}_{\text{train}}$ 上完美对齐的必要条件，因为 PA-SF 和 PA-SF(不带 D) 的 LER 值几乎为 0。尽管 PA-SF(不带 MD) 和 Skew-Fit 在训练性能良好的情况下，但在没有 MMD 损失的情况下却面临较

大的 LER 值。此外，PA-SF 与 PA-SF(不带 AS) 之间的比较展示了在 MMD 损失中使用对齐数据的重要性 (否则，匹配本质上是错误的)。

2. The only if condition, i.e., $\Phi\left(x^{e}\left(s\right)\right)=\Phi\left(x^{e'}\left(s'\right)\right)$ only if $s=s'$, is also achieved empirically by visualizing the reconstruction of the VAE (Figure 9 in Appendix D) and we believe this is satisfied by both the difference loss and the reconstruction loss. Under the only if condition, the SAC Haarnoja et al. [2018] trained on the latent space achieves the optimal performance. In contrast, PA-SF (w/o AS) fails to learn well on the training environments as its latent representation is mixed over different states.

2. 唯一的必要条件，即 $\Phi\left(x^{e}\left(s\right)\right)=\Phi\left(x^{e'}\left(s'\right)\right)$ 仅当 $s=s'$ 时，也通过可视化 VAE 的重构 (附录 D 中的图 9) 得到经验验证，我们认为这由差异损失和重构损失共同满足。在唯一必要条件下，SAC Haarnoja 等人 [2018] 在潜在空间上训练时达到了最佳性能。相比之下，PA-SF(不带 AS) 在训练环境中学习效果不佳，因为其潜在表示在不同状态之间混合。

Second, we focus on the generalization performance on target domains t , i.e., term(t)in Eq. (3). We observe the following:

其次，我们关注目标领域 t 的泛化性能，即方程 (3) 中的项 (t)。我们观察到以下几点:

1. As shown by the learning curve of test environments, the target domain performance of different ablations match that of the LER metric: SkewFit, PA-SF (w/o AS) > PA-SF (w/o MD) > PA-SF (w/o D) > PA-SF. During training, these ablations have almost the same performance, except PA-SF (w/o AS). This indicates that the increased test performance indeed comes from the improved representation quality of the encoder Φ , i.e., more aligned. This supports our claim at the end of Section 3.1 and the upper bound analysis on the(t)term in Appendix C.4, that the increased invariant property of Φ produces better domain generalization performance.

1. 从测试环境的学习曲线可以看出，不同消融实验的目标领域性能与 LER 指标相匹配:SkewFit, PA-SF (不含 AS) > PA-SF (不含 MD) > PA-SF (不含 D) > PA-SF。在训练过程中，这些消融实验的性能几乎相同，除了 PA-SF (不含 AS)。这表明，测试性能的提高确实来自于编码器 Φ 的表示质量提升，即更加对齐。这支持了我们在第 3.1 节末尾的论点以及附录 C.4 中关于 (t) 项的上界分析，即 Φ 的不变性增强特性产生了更好的领域泛化性能。

2. In test environment ablations, the LER is reduced significantly on methods with L^{MMD} . This supports our claim that a perfect alignment encoder on training environments also improves the encoder's invariant property on unseen environments. In addition, by encouraging dispersed latent representation, the difference loss L^{DIFF} also plays a role in reducing LER during testing. This supports the necessity of both losses for generalization.

2. 在测试环境的消融实验中，具有 L^{MMD} 的方法的 LER 显著降低。这支持了我们的论点，即在训练环境中完美对齐的编码器也改善了编码器在未见环境中的不变性。此外，通过鼓励分散的潜在表示，差异损失 L^{DIFF} 在测试过程中也在降低 LER 方面发挥了作用。这支持了两种损失对泛化的必要性。

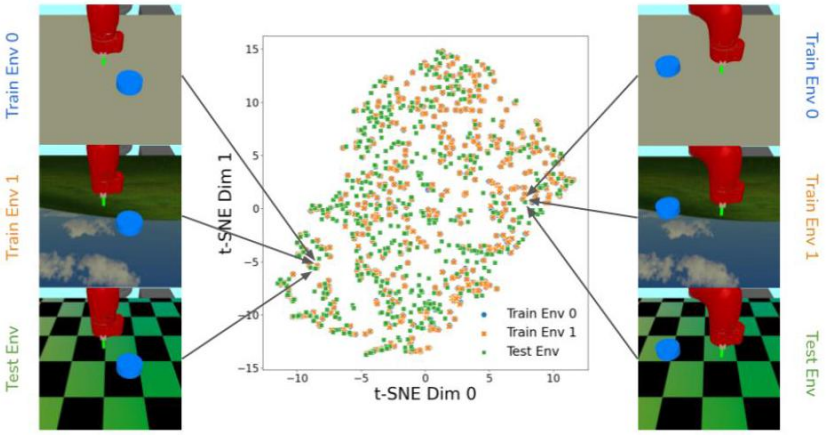


Figure 5: t-SNE visualization of the latent space $\Phi\left(x^{e}\right)$ trained with PA-SF for three environments: 2 training and 1 testing of Push as well as instances visualization.

[1] A single L^{DIFF} is not useful here.
[1] 单个 L^{DIFF} 在这里没有用处。

图 5: 使用 PA-SF 训练的三个环境的潜在空间 $\Phi(x^e)$ 的 t-SNE 可视化:2 个训练环境和 1 个测试环境的 Push 以及实例可视化。

We observe the similar results in other tasks as well (Appendix D). Here, we also visualize the latent space by t-SNE plot to illustrate the perfect alignment on task Push. Dots in training environments are matched perfectly and the corresponding test environment dot is approximately near as expected.

我们在其他任务中也观察到了类似的结果 (附录 D)。在这里，我们还通过 t-SNE 图可视化潜在空间，以说明任务 Push 上的完美对齐。训练环境中的点完全匹配，对应的测试环境点也如预期般接近。

5 Related Work

5 相关工作

Goal-conditioned RL: Goal-conditioned RL Kaelbling [1993], Schaul et al. [2015] removes the need for complicated reward shaping by only rewarding agents for reaching a desired set of goal states. RIG Nair et al. [2018] is the seminal work for visual-input, reward-free learning in goal-conditioned MDPs. Skew-Fit Pong et al. [2020] improves over RIG Nair et al. [2018] in training efficiency by ensuring the behavioural goals used to explore are diverse and have wide state coverage. However, Skew-Fit has its own limitation in understanding the semantic meaning of the goal-conditioned task. To acquire more meaningful goal's and observation's latent representation, several approaches apply inductive biases or seek human feedback. ROLL Wang et al. [2020] applies object extraction methods under strong assumptions, while WSC Lee et al. [2020] uses weak binary labeled data as the reward function. Others explore the same goal-conditioned RL problem via hindsight experience replay Marcin et al. [2017], Ren et al. [2019], Ghosh et al. [2019], unsupervised reward learning Péré et al. [2018], inverse dynamics models Paster et al. [2020], C-learning Eysenbach et al. [2020], goal generation Florensa et al. [2018], Nair and Finn [2019], Pitis et al. [2020], goal-conditioned forward models Nair et al. [2020], and hierarchical RL Li et al., Nachum et al. [2018], Zhang et al. [2020c], Hou et al. [2020]. Our study focus on learning goal-conditioned policies that is invariant of spurious environmental factors. We aim to learn a policy that can generalize to visual goals in unseen test environments.

目标条件强化学习: 目标条件强化学习 Kaelbling [1993]，Schaul 等 [2015] 通过仅对代理达到期望的目标状态集给予奖励，从而消除了对复杂奖励塑造的需求。RIG Nair 等 [2018] 是目标条件马尔可夫决策过程 (MDP) 中视觉输入、无奖励学习的开创性工作。Skew-Fit Pong 等 [2020] 在训练效率上改进了 RIG Nair 等 [2018]，通过确保用于探索的行为目标多样且具有广泛的状态覆盖。然而，Skew-Fit 在理解目标条件任务的语义意义方面存在自身的局限性。为了获得更有意义的目标和观察的潜在表示，几种方法应用了归纳偏见或寻求人工反馈。ROLL Wang 等 [2020] 在强假设下应用了对象提取方法，而 WSC Lee 等 [2020] 则使用弱二元标记数据作为奖励函数。其他研究通过事后经验重放 Marcin 等 [2017]，Ren 等 [2019]，Ghosh 等 [2019]，无监督奖励学习 Péré 等 [2018]，逆动态模型 Paster 等 [2020]，C-learning Eysenbach 等 [2020]，目标生成 Florensa 等 [2018]，Nair 和 Finn [2019]，Pitis 等 [2020]，目标条件前向模型 Nair 等 [2020]，以及层次强化学习 Li 等，Nachum 等 [2018]，Zhang 等 [2020c]，Hou 等 [2020] 探索相同的目标条件强化学习问题。我们的研究重点是学习对虚假环境因素不变的目标条件策略。我们的目标是学习一种能够在未见测试环境中对视觉目标进行泛化的策略。

Learning Invariants in RL: Robustness to domain shifts is crucial for real-world applications of RL. Zhang et al. [2020a b], Gelada et al. [2019] implement the model-bisimulation framework Ferns et al. [2011] to acquire a minimal but sufficient representation for solving the MDP problem. However, model-bisimulation for high-dimension problems typically requires domain-invariant and dense rewards. These assumptions do not hold in GBMDPs. Contrastive Metric Embeddings (CME) Agarwal et al. [2021] instead uses π^* -bisimulation metric but it also requires extra information of the optimal policy. Another line of work tries to address these issues via self-supervised learning. Stone et al. [2021] tests multiple data augmentation methods including RAD Laskin et al. [2020b] and DrQ Kostrikov et al. [2020] to boost the robustness of the representation as well as the policy. Our work can also apply data augmentation in practice. However, we find that RAD is not very helpful in the Skew-Fit framework. Additionally, Hansen et al. [2020], Bodnar et al. [2020] use self-supervised correction during real-world adaptation like sim2real transfer but these methods are incompatible for domain generalization.

在强化学习中的学习不变性: 对领域转变的鲁棒性对于强化学习在现实世界应用中的重要性至关重要。Zhang 等人 [2020a b]，Gelada 等人 [2019] 实现了模型-双重仿真框架 Ferns 等人 [2011]，以获取解决 MDP 问题的最小但足够的表示。然而，高维问题的模型-双重仿真通常需要领域不变和密集的奖励。这些假设在 GBMDPs 中并不成立。对比度度量嵌入 (CME) Agarwal 等人 [2021] 则使用 π^* -双重仿真度量，但它也需要额外的最优策略信息。另一条研究方向试图通过自监督学习来解决这些问题。Stone 等人

[2021] 测试了多种数据增强方法，包括 RAD Laskin 等人 [2020b] 和 DrQ Kostrikov 等人 [2020]，以增强表示和策略的鲁棒性。我们的工作在实践中也可以应用数据增强。然而，我们发现 RAD 在 Skew-Fit 框架中并不太有帮助。此外，Hansen 等人 [2020]，Bodnar 等人 [2020] 在现实世界适应期间使用自监督校正，例如 sim2real 转移，但这些方法与领域泛化不兼容。

6 Conclusion

6 结论

In this paper, we study the problem of learning invariant policies in Goal-conditioned RL agents. The problem is formulated as a GBMDP, which is an extension of Goal-conditioned MDPs and Block MDPs where we want the agent's policy to generalize to unseen test environments after training on several training environments.

在本文中，我们研究了在目标条件强化学习代理中学习不变策略的问题。该问题被表述为 GBMDP，这是目标条件 MDP 和块 MDP 的扩展，我们希望代理的策略在经过多个训练环境的训练后能够泛化到未见过的测试环境。

As supported by the generalization bound for GBMDP, we propose a simple but effective heuristic, i.e., perfect alignment which we can minimize the bound asymptotically and benefit the generalization. To learn a perfect alignment encoder, we propose a practical method based on aligned sampling. The method resembles contrastive learning: matching latent representation of aligned data via MMD loss and dispersing the whole latent representations via the DIFF loss. Finally, we propose a practical implementation Perfect Alignment for Skew-Fit (PA-SF) by adding the perfect alignment loss to Skew-Fit and changing the VAE structure to handle GBMDPs.

根据 GBMDP 的泛化界限，我们提出了一种简单而有效的启发式方法，即完美对齐，我们可以渐近地最小化界限并有利于泛化。为了学习一个完美对齐编码器，我们提出了一种基于对齐采样的实用方法。该方法类似于对比学习: 通过 MMD 损失匹配对齐数据的潜在表示，并通过 DIFF 损失分散整个潜在表示。最后，我们通过将完美对齐损失添加到 Skew-Fit 中并改变 VAE 结构以处理 GBMDP，提出了一个实用实现——完美对齐以适应偏斜 (PA-SF)。

The empirical evaluation shows that our method is the SOTA algorithm and achieves a remarkable increase in test environments' performance over other methods. We also compare our algorithm with several ablations and analyze the representation quantitatively. The results support our claims in the theoretical analysis that perfect alignment criteria is effective and that we can effectively optimize the criteria with our proposed method. We believe the perfect alignment criteria will enable applications in diverse problem settings and offers interesting directions for future work, such as extensions to other goal-conditioned learning frameworks Eysenbach et al. [2020], Paster et al. [2020].

实证评估表明，我们的方法是 SOTA 算法，并在测试环境的性能上显著超过其他方法。我们还将我们的算法与几种消融进行比较，并定量分析表示。结果支持我们在理论分析中的主张，即完美对齐标准是有效的，并且我们可以有效地使用我们提出的方法优化该标准。我们相信完美对齐标准将使其在多样的问题设置中得到应用，并为未来的工作提供有趣的方向，例如扩展到其他目标条件学习框架 Eysenbach 等 [2020]，Paster 等 [2020]。

Acknowledgements

致谢

We are grateful for the feedback from anonymous reviewers. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute <http://www.vectorinstitute.ai/partners>.

我们感谢匿名评审者的反馈。准备本研究所使用的资源部分由安大略省、加拿大政府通过 CIFAR 提供，以及赞助 Vector Institute 的公司 <http://www.vectorinstitute.ai/partners>。

References

参考文献

David Silver, Julian Schrittwieser, Karen Simonyan, Aj Antonoglou, Joannis abd Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. Mastering the game of go without human knowledge. Nature, 550, 2017.

David Silver, Julian Schrittwieser, Karen Simonyan, Aj Antonoglou, Joannis abd Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, 和 Demis Hassabis. 在没有人类知识的情况下掌握围棋游戏。自然, 550, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc Bellemare, Alex Graves, Martin Riedmiller, Andreas Fidjeland, Georg strovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan umaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. Nature, 518, 2015.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc Bellemare, Alex Graves, Martin Riedmiller, Andreas Fidjeland, Georg strovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan umaran, Daan Wierstra, Shane Legg, 和 Demis Hassabis. 通过深度强化学习实现人类水平的控制。自然, 518, 2015.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 1861-1870, Stockholmsmässan, Stockholm Sweden, 10-15 Jul 2018. PMLR.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, 和 Sergey Levine. 软演员-评论家: 具有随机演员的离线最大熵深度强化学习。在第 35 届国际机器学习会议论文集中，机器学习研究论文集第 80 卷，页 1861-1870，瑞典斯德哥尔摩 Stockholmsmässan，2018 年 7 月 10-15 日。PMLR.

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning. CoRR, abs/1810.12282, 2018. URL

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, 和 Dawn Song. 评估深度强化学习中的泛化能力。CoRR, abs/1810.12282, 2018. URLhttp://arxiv.org/abs/1810.12282

Ryan Julian, Benjamin Swanson, Gaurav S Sukhatme, Sergey Levine, Chelsea Finn, and Karol Hausman. Efficient adaptation for end-to-end vision-based robotic manipulation. arXiv preprint arXiv:2004.10190, 2020.

Ryan Julian, Benjamin Swanson, Gaurav S Sukhatme, Sergey Levine, Chelsea Finn, 和 Karol Hausman. 基于视觉的机器人操作的高效适应. arXiv 预印本 arXiv:2004.10190, 2020.

Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, and Doina Precup. Invariant causal prediction for block MDPs. In International Conference on Machine Learning, pages 11214-11224. PMLR, 2020a.

Amy Zhang, Clare Lyle, Shagun Sodhani, Angelos Filos, Marta Kwiatkowska, Joelle Pineau, Yarin Gal, 和 Doina Precup. 针对块 MDP 的不变因果预测. 在国际机器学习会议上，页码 11214-11224. PMLR, 2020a.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. arXiv preprint arXiv:2006.10742, 2020b.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, 和 Sergey Levine. 学习无重构的强化学习不变表示. arXiv 预印本 arXiv:2006.10742, 2020b.

Andrychowicz Marcin, Wolski Filip, Ray Alex, Schneider Jonas, Fong Rachel, Welinder Peter, McGrew Bob, Tobin Josh, Abbeel Pieter, and Zaremba Wojciech. Hindsight experience replay. In Advances in Neural Information Processing Systems, 2017.

Benjamin Eysenbach, Ruslan Salakhutdinov, and Sergey Levine. C-learning: Learning to achieve goals via recursive classification. arXiv preprint arXiv:2011.08909, 2020.

Keiran Paster, Sheila A McIlraith, and Jimmy Ba. Planning from pixels using inverse dynamics models. arXiv preprint arXiv:2012.02419, 2020.

Alexandre Péré, Sébastien Forestier, Olivier Sigaud, and Pierre-Yves Oudeyer. Unsupervised learning of goal spaces for intrinsically motivated goal exploration. In International Conference on Learning Representations, 2018.

Norm Ferns, Prakash Panangaden, and Doina Precup. Bisimulation metrics for continuous markov decision processes. SIAM Journal on Computing, 40(6):1662-1714, 2011.

Vitchyr Pong, Murtaza Dalal, Steven Lin, Ashvin Nair, Shikhar Bahl, and Sergey Levine. Skew-fit: State-covering self-supervised reinforcement learning. In International Conference on Machine Learning, pages 7783-7792. PMLR, 2020.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient RL with rich observations via latent state decoding. In International Conference on Machine Learning, pages 1665-1674. PMLR, 2019.

Leslie Pack Kaelbling. Learning to achieve goals. In IJCAI, pages 1094-1099. Citeseer, 1993.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In International conference on machine learning, pages 1312-1320. PMLR, 2015.

Brockman Greg, Cheung Vicki, Pettersson Ludwig, Schneider Jonas, Schulman John, Tang Jie, and Zaremba Wojciech. Openai gym, 2016.

Vitchyr Pong, Murtaza Dalal, Steven Lin, and Ashvin Nair. multiworld. https://github.com/vitchyr/multiworld, 2018.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, et al. Wilds: A benchmark of in-the-wild distribution shifts. arXiv preprint arXiv:2012.07421, 2020.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. arXiv preprint arXiv:1907.02893, 2019.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. Machine learning, 79(1):151-175, 2010.

Anthony Sicilia, Xingchen Zhao, and Seong Jae Hwang. Domain adversarial neural networks for domain generalization: When it works and how to improve. arXiv preprint arXiv:2102.03924, 2021.

Isabela Albuquerque, João Monteiro, Mohammad Darvishi, Tiago H Falk, and Ioannis Mitliagkas. Generalizing to unseen domains via distribution matching. arXiv preprint arXiv:1911.00804, 2019.

Wikipedia. Total variation distance of probability measures - Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Total%20variation% 20distance%20of%20probability%20measures&oldid=10 2021. [Online; accessed 24-May-2021].

Hong Liu, Mingsheng Long, Jianmin Wang, and Michael Jordan. Transferable adversarial training: A general approach to adapting deep classifiers. In International Conference on Machine Learning, pages 4013-4022. PMLR, 2019.

Kei Akuzawa, Yusuke Iwasawa, and Yutaka Matsuo. Adversarial invariant feature learning with accuracy constraint for domain generalization. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 315-331. Springer, 2019.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S Zemel. The variational fair autoencoder. In ICLR, 2016.

Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In Proceedings of the European Conference on Computer Vision (ECCV), pages 624-639, 2018.

Xin Jin, Cuiling Lan, Wenjun Zeng, and Zhibo Chen. Feature alignment and restoration for domain generalization and adaptation. arXiv preprint arXiv:2006.12009, 2020.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In International Conference on Machine Learning, pages 1587-1596, 2018.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel method for the two-sample problem. Journal of Machine Learning Research, 1:1-10, 2008.

Ji Zhao and Deyu Meng. Fastmmd: Ensemble of circular discrepancy for efficient two-sample test. Neural computation, 27(6):1345-1372, 2015.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597-1607. PMLR, 2020.

Michael Laskin, Aravind Srinivas, and Pieter Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. In International Conference on Machine Learning, pages 5639-5650. PMLR, 2020a.

Ashvin Nair, Vitchyr Pong, Murtaza Dalal, Shikhar Bahl, Steven Lin, and Sergey Levine. Visual reinforcement learning with imagined goals. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, pages 9209-9220, 2018.

Austin Stone, Oscar Ramirez, Kurt Konolige, and Rico Jonschkowski. The distracting control suite-a challenging benchmark for reinforcement learning from pixels. arXiv preprint arXiv:2101.02722, 2021.

Michael Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. arXiv preprint arXiv:2004.14990, 2020b.

Yufei Wang, Gautham Narayan Narasimhan, Xingyu Lin, Brian Okorn, and David Held. Roll: Visual self-supervised reinforcement learning with object reasoning. arXiv preprint arXiv:2011.06777, 2020.

Lisa Lee, Benjamin Eysenbach, Ruslan Salakhutdinov, Chelsea Finn, et al. Weakly-supervised reinforcement learning for controllable behavior. arXiv preprint arXiv:2004.02860, 2020.

Zhizhou Ren, Kefan Dong, Yuan Zhou, Qiang Liu, and Jian Peng. Exploration via hindsight goal generation. Advances in Neural Information Processing Systems, 32:13485-13496, 2019.

Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals via iterated supervised learning. arXiv e-prints, pages arXiv-1912, 2019.

Carlos Florensa, David Held, Xinyang Geng, and Pieter Abbeel. Automatic goal generation for reinforcement learning agents. In International conference on machine learning, pages 1515-1528. PMLR, 2018.

Suraj Nair and Chelsea Finn. Hierarchical foresight: Self-supervised learning of long-horizon tasks via visual subgoal generation. In International Conference on Learning Representations, 2019.

Silviu Pitis, Harris Chan, Stephen Zhao, Bradly Stadie, and Jimmy Ba. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning. In International Conference on Machine Learning, pages 7750-7761. PMLR, 2020.

Suraj Nair, Silvio Savarese, and Chelsea Finn. Goal-aware prediction: Learning to model what matters. In International Conference on Machine Learning, pages 7207-7219. PMLR, 2020.

Siyuan Li, Lulu Zheng, Jianhao Wang, and Chongjie Zhang. Learning subgoal representation with slow dynamics.

Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Near-optimal representation learning for hierarchical reinforcement learning. In International Conference on Learning Representations, 2018.

Lunjun Zhang, Ge Yang, and Bradly C Stadie. World model as a graph: Learning latent landmarks for planning. arXiv preprint arXiv:2011.12491, 2020c.

Zhimin Hou, Jiajun Fei, Yuelin Deng, and Jing Xu. Data-efficient hierarchical reinforcement learning for robotic assembly control applications. IEEE Transactions on Industrial Electronics, 2020.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In International Conference on Machine Learning, pages 2170-2179. PMLR, 2019.

Rishabh Agarwal, Marlos C Machado, Pablo Samuel Castro, and Marc G Bellemare. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. arXiv preprint arXiv:2101.05265, 2021.

Ilya Kostrikov, Denis Yarats, and Rob Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. arXiv preprint arXiv:2004.13649, 2020.

Nicklas Hansen, Yu Sun, Pieter Abbeel, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. arXiv preprint arXiv:2007.04309, 2020.

Cristian Bodnar, Karol Hausman, Gabriel Dulac-Arnold, and Rico Jonschkowski. A geometric perspective on self-supervised policy adaptation. arXiv preprint arXiv:2011.07318, 2020.

John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International conference on machine learning, pages 1889-1897. PMLR, 2015.

Christopher JCH Watkins and Peter Dayan. Q-learning. Machine learning, 8(3-4):279-292, 1992.