arXiv:2305.00654v2 [cs.LG] 2 May 2023

使用奇异值分解的深度强化学习表征与探索

Yash Chandak [1] **Shantanu Thakoor** [2] **Zhaohan Daniel Guo** [2] **Yunhao Tang** [2] **Remi Munos** [2] **Will Dabney** [2] **Diana Borsa** [2]

摘要

表示学习和探索是任何深度强化学习智能体面临的关键挑战。在这项工作中，我们提出了一种基于奇异值分解的方法，可用于获取能够保留领域底层转移结构的表征。值得关注的是，我们还证明，这些表征能够捕捉状态访问的相对频率，从而无需额外成本即可估算伪计数。为了将这种分解方法扩展到大规模领域，我们提出了一种算法，该算法无需构建转移矩阵，能够利用深度网络，并且支持小批量训练。此外，我们借鉴了预测性状态表征的思想，并将分解方法扩展到部分可观测环境。通过在具有部分可观测领域的多任务设置中开展实验，我们证明，所提出的方法不仅能够在 DM-Lab-30 环境中学习有用的表征（这些环境的输入涉及语言指令、像素图像和奖励等内容），还能够有效应对 DM-Hard-8 环境中的高难度探索任务。

1. 引言

开发能够解决具有复杂观测、部分可观测性和稀疏奖励信号问题的强化学习（RL）方法，是当前一个活跃的研究方向。这类问题带来了有趣的挑战，因为观测通常包含由视觉特征、文本和语音构成的多模态信息。在部分可观测性的情况下，智能体必须查看一系列过去的观测，并学习有用的表示来解决强化学习任务。此外，当奖励信号稀疏时，进行穷举式探索变得不切实际，因此需要谨慎考虑环境中哪些部分应当被探索。也就是说，高效的探索取决于智能体的表示。因此，构建一个同时应对表示学习和探索的统一框架，有助于解决具有挑战性的问题。

表示学习方法通常包含用于预测未来事件的辅- 助任务，可以直接预测（Jader- berg 等人，2016年），也可以在潜在空间中预测（Guo 等人，2020年；2022；Schwarzer 等人，2020年）。然而，在观测内容丰富的场景中，完整重建观测可能并不现实，而在潜在空间中进行重建则容易导致表示坍塌。另一种方法是采用对比损失的辅助任务，以避免表示坍塌（Chen 等人，2020年；Srinivas 等人，2020年），但其效果可能对负样本采样器的选择十分敏感（Zbontar 等人，2021）。

类似地，旨在促进探索的方法通常会采用基于计数的奖励（Bellemare 等人，2016年），或基于重建误差的奖励（Schmidhuber，1991；Pathak 等人，2017年）。尽管这些思路颇具吸引力，但在丰富观测设置中，枚举计数可能颇具挑战性，而重建过程可能会聚焦于无关细节，从而产生误导性的奖励。RND（Burda 等人，2018年）等方法在一定程度上解决了这些问题，但它们与任何表示学习方法完全脱钩。这极大地限制了在智能体的环境模型中开展有针对性探索的可能性。其他相关工作的详细概述将在第6节和附录A中。

在本研究中，我们提出 De-Rex：一种基于分解的表示与探索方法。该方法基于奇异值分解（奇异值分解，SVD），旨在丰富观测／高维环境中训练兼具 实用性的状态表示和用于探索的伪计数估计。

保留结构的表征： De-Rex 使用一种分解过程来学习能够保留转移结构的状态表示，且完全无需构建转移矩阵（第 3节）。此外，我们的方法只需以小批量形式访问转移样本，并且兼容基于深度网络的分解

[1]马萨诸塞大学 [2]DeepMind。通信作者：Yash Chandak <ychandak@cs.umass.edu>。

论文集收录于 40th 国际机器学习会议，美国夏威夷州檀香山。PMLR 202，2023年。版权所有 © 2023年，作者所有。

（第 4节）。这改进了先前的研究：这些研究虽然使用了分解，但仅限于表格设置（Mahadevan，2005年；Mahadevan 和 Maggioni，2007年），或者要求矩阵为对称矩阵（例如拉普拉斯算子（Wu 等人，2018年））；当转移核不对称时，这可能会破坏其底层结构。

用于探索的伪计数：我们证明，这些学习得到的状态表示的范数能够捕捉相对状态访问频率，从而提供一种成本极低的过程来获取伪计数（第3 和4节）。与以往需要利用转移核的谱分解来学习探索选项的工作相比（Machado 等人，2017a；b），这种方法显著简化了探索过程。

大规模实验中的性能：基于预测状态表示的思想（Littman 和 Sut- ton，2001；Singh 等人，2003），De-Rex 还提供了一种在部分可观测环境中处理表示学习与探索的方法（第5节）。我们在 DM-Lab-30（Beattie 等人，2016年）和 DM-Hard-8 探索任务（Paine 等人，2019年）的多任务设置中（不使用任务标签），展示了 De-Rex 的有效性。这些领域通过程序生成，具有部分可观测性，且观测具有多模态特征，包括语言指令、像素图像和奖励（第7节）。

2. 预备知识

我们将从马尔可夫决策过程（MDP）的设定开始，并利用奇异值分解（SVD）推导出关键见解。稍后在第 5 节，我们将把这些思路扩展到部分可观测马尔可夫决策过程（POMDP）。

任意向量 x 均视为列向量。任意矩阵 M 均用粗体大写字母表示，而 $\mathbf{M}_{[i,j]}$ 对应于 M 的第 i^{th} 行、第 j^{th} 列。一个马尔可夫决策过程是一个五元组 $(\mathcal{S}, \mathcal{A}, p, r, d_0)$，其中 $\mathcal{S}$ 是有限状态的集合，$\mathcal{A}$ 是有限动作的集合，r 是奖励函数，p是转移函数，d_0 是初始状态分布。策略 π 是一种决策规则，令 $\mathbf{P}_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ 为由 π 诱导的转移矩阵，使得 $\mathbf{P}^{\pi}_{[s,s']} := \sum_{a\in\mathcal{A}} \Pr(s'|s,a)\pi(a|s)$。策略 π 的性能定义为 $J(\pi) := \mathbb{E}_\pi\left[\sum_{t=0}^{\infty}\gamma^t R_t\right]$，其中$\gamma \in [0,1)$，且 R_t 是在时刻 t 与环境交互并采取 π时观测到的奖励。令 Π 为一类（参数化）策略。目标是求得 $\arg\max_{\pi\in\Pi} J(\pi)$。

2.1. 为什么使用奇异值分解?

继原型价值函数的开创性工作之后（Mahadevan，2005年；Mahadevan 和 Maggioni，2007年），通过谱分解获得的表征已被证明不仅能有效减少 Bellman 误差——

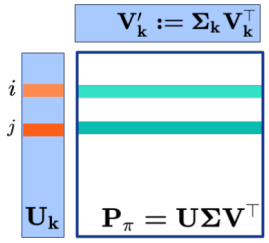

图1。考虑对 $\mathbf{P}_\pi \approx \mathbf{U}_\mathbf{k}\mathbf{V}'_\mathbf{k}$ 进行 top-k 奇异值分解，其中 $\mathbf{V}'_\mathbf{k} := \mathbf{\Sigma}_\mathbf{k}\mathbf{V}_\mathbf{k}^\top \in \mathbb{R}^{k\times|\mathcal{S}|}$。分解后，我们将 $\mathbf{U}_\mathbf{k} \in \mathbb{R}^{|\mathcal{S}|\times k}$ 的各行视为状态的 $k \ll |\mathcal{S}|$ 维表示。直观而言，如果两个状态具有相似的下一状态转移分布（青绿色行 $\mathbf{P}^{\pi}_{[i]}$ 和 $\mathbf{P}^{\pi}_{[j]}$），那么它们的表示（橙色行 $\mathbf{U}^{\mathbf{k}}_{[i]}$ 和 $\mathbf{U}^{\mathbf{k}}_{[j]}$）也会相似。注意，$\mathbf{U}_\mathbf{k}$ 中的各行可以被视为组合公共基向量 $\mathbf{V}'_\mathbf{k}$ 的系数，以最佳估计 $\mathbf{P}_\pi$ 的各行（即 $\mathbf{P}^{\pi}_{[i]} \approx \mathbf{U}^{\mathbf{k}}_{[i]}\mathbf{V}'_\mathbf{k}$ 和 $\mathbf{P}^{\pi}_{[j]} \approx \mathbf{U}^{\mathbf{k}}_{[j]}\mathbf{V}'_\mathbf{k}$）。因此，如果 $\mathbf{P}^{\pi}_{[i]} \approx \mathbf{P}^{\pi}_{[j]}$，那么 $\mathbf{U}^{\mathbf{k}}_{[i]}\mathbf{V}'_\mathbf{k} \approx \mathbf{U}^{\mathbf{k}}_{[j]}\mathbf{V}'_\mathbf{k}$。

（Behzadian 等，2019）还能稳定离策略学习（Ghosh 和 Bellemare，2020）。Lyle 等人近期的研究（2021；2022b）尤其值得关注：他们表明，在理想化条件下，仅使用时序差分学习进行价值函数估计所产生的状态特征，能够捕获 $\mathbf{P}_\pi$ 的特征基中前 k 个特征向量所张成的子空间。如下文所示，$\mathbf{P}_\pi$ 的这前 k 个特征向量与 $(\mathbf{I}-\gamma\mathbf{P}_\pi)^{-1}$ 的前 k 个特征向量完全相同。因此，这些基向量对于获得状态表示特别有用，因为价值函数关于 $(\mathbf{I}-\gamma\mathbf{P}_\pi)^{-1}$ 是线性的（Sutton 和 Barto，2018）。也就是说，$v^\pi = (\mathbf{I}-\gamma\mathbf{P}_\pi)^{-1}R_\pi$，其中 $R_\pi \in \mathbb{R}^{|\mathcal{S}|}$ 是一个向量，其各个状态对应的元素均为单步期望奖励 $\mathbf{E}_\pi[R|s]$。

定理 1. 若 $\mathbf{P}_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ 的特征值均为实数且互不相同，[1] 则对于任意 k，$\mathbf{P}_\pi$ 的前 k 个特征向量与 $(\mathbf{I}-\gamma\mathbf{P}_\pi)^{-1}$ 相同。

所有结果的证明均推迟至附录B。由此自然引出一个问题：我们能否显式分解$\mathbf{P}_\pi$以助力表示学习？要回答这个问题，定理1从实践角度来看可能很有用，因为估计和分解$\mathbf{P}_\pi$ 可能比对$(\mathbf{I}-\gamma\mathbf{P}_\pi)^{-1}$更容易。然而遗憾的是，当$\mathbf{P}_\pi$ 不可对角化时，$\mathbf{P}_\pi$ 的特征分解通常可能不存在，例如当$\mathbf{P}_\pi$ 不是对称矩阵时。

因此，我们转而考虑 $\mathbf{P}_\pi$ 的奇异值分解（SVD）。奇异值分解是特征分解的推广，而且始终存在（VanLoan，1976）。也就是说，始终存在正交矩阵$\mathbf{U} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$、$\mathbf{V} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$以及实值对角矩阵 $\mathbf{\Sigma} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$，使得$\mathbf{P}_\pi = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$。在图1中，我们直观地说明了 $\mathbf{P}_\pi$ 的前 k 项 SVD 分解所捕捉的内容：

[1]为简化“前 k 个”特征值的选择，我们假设特征值均为实数且互不相同。

$\mathbf{P}_\pi \approx \mathbf{U_k}\mathbf{\Sigma_k}\mathbf{V_k}^\top$，其中$\mathbf{U_k} \in \mathbb{R}^{|\mathcal{S}|\times k}$和$\mathbf{V_k} \in \mathbb{R}^{|\mathcal{S}|\times k}$是对应于前 k 个奇异值 $\mathbf{\Sigma_k} \in \mathbb{R}^{k\times k}$ 的前 k 个左、右奇异向量。我们基于这一洞见，开发了一种可扩展的 SVD 方法；它不仅限于表示学习，还可用于获得驱动探索的内在奖励。

在强化学习之外，人们还提出了不同的矩阵分解方法，例如核主成分分析（Schölkopf 等，1997年）、深度分解（De Handschutter 等，2021；Handschutter 等，2021）和 EigenGame（Gemp 等，2020年；Deng 等，2022）。下文中，我们将从这些方法中汲取灵感，但重点关注强化学习设置中的挑战，即非对称的（转移）矩阵、探索奖励的构造以及部分可观测性。

3. 表格设置

为了阐明核心见解并为我们的方法奠定基础，我们将首先考虑表格设置。在第 4 节中，我们将进一步拓展这些见解，开发使用函数近似的可扩展方法。

要精确分解$\mathbf{P}_\pi$，需要与所有状态对应的转移。这显然不切实际，因为智能体很少能够访问所有可能的状态。因此，我们考虑分解$\mathbf{P}_\pi$的加权版本。设 $\mathcal{D} \coloneqq \{(s,\ a,\ r,\ s')_i\}_{i=1}^n$ 是使用 π收集的包含 n 个转移元组的数据集。设 $d \in \Delta(\mathcal{S})$ 表示状态 s 的分布，观测到的转移元组即从该分布中采样得到；并设 $\mathbf{D} \coloneqq \mathrm{diag}(d) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$。在下文中，我们将考虑分解加权矩阵 $\mathbf{DP}_\pi$（当 d 服从均匀分布时，$\mathbf{DP}_\pi \propto \mathbf{P}_\pi$）。正如我们将在后续章节中详细阐述的那样，分解$\mathbf{DP}_\pi$ 的选择具有两大优势：**（a）**它将提供一种实用的表示学习过程，只需对观测到的转移元组进行采样；**（b）**它还将几乎自动地提供伪计数的估计值，而伪计数将对探索有所帮助。

表示学习： 本节概述表示学习的核心思想。在继续之前，我们先引入一些额外的记号。令 $x_s \in \mathbb{R}^{|\mathcal{S}|}$为与状态 s对应的独热编码，并令$\mathbf{X} \in \mathbb{R}^{n\times|\mathcal{S}|}$ 为数据集$\mathcal{D}$ 中状态 s 对应的 x_s 的堆叠。同样，令 $y_s \in \mathbb{R}^{|\mathcal{S}|}$ 为与 s' 对应的独热编码，并令 $\mathbf{Y} \in \mathbb{R}^{n\times|\mathcal{S}|}$ 为由数据集 $\mathcal{D}$ 中的 s' 堆叠而成的矩阵。

利用数据集 $\mathcal{D}$，我们给出一个$\mathbf{DP}_\pi$的样本估计，并希望对其进行分解。虽然我们绝不会实际构造这个可能难以处理的$\mathbf{DP}_\pi$估计，但它有助于推导所提出的过程。

因此，令

$$\mathbf{A_n} \coloneqq \frac{1}{n}\mathbf{X}^\top\mathbf{Y} = \frac{1}{n}\sum_{i=1}^n x_i y_i^\top. \tag{1}$$

定理 2。 $\mathbf{A_n}$ 是$\mathbf{DP}_\pi$的无偏且一致估计量，即：

$$\forall i,j,\quad \mathbb{E}\left[\mathbf{A_n}_{[i,j]}\right] = \mathbf{DP}_{\pi[i,j]},\quad \mathbf{A_n}_{[i,j]} \xrightarrow{a.s.} \mathbf{DP}_{\pi[i,j]}.$$

稍微扩展一下记号，令 $\mathbf{U} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ 和$\mathbf{V} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ 分别为$\mathbf{A_n}$的左、右奇异向量，并令 $\mathbf{\Sigma} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ 为奇异值，使得

$$\mathbf{A_n} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top. \tag{2}$$

现在，在（2）两边左乘 $\mathbf{U}^\top$，再右乘 $\mathbf{V}$，并注意到$\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}$，

$$\mathbf{U}^\top\mathbf{A_n}\mathbf{V} = \mathbf{\Sigma}. \tag{3}$$

将（3）代入（1）展开，可得

$$\frac{1}{n}\sum_{i=1}^n \mathbf{U}^\top x_i y_i^\top \mathbf{V} = \mathbf{\Sigma}. \tag{4}$$

现在，令 $f_\mathbf{U}: \mathcal{S} \to \mathbb{R}^{|\mathcal{S}|}$ 和 $g_\mathbf{V}: \mathcal{S} \to \mathbb{R}^{|\mathcal{S}|}$ 为分别由“权重” $\mathbf{U}$ 和 $\mathbf{V}$ 参数化的线性函数。于是，（4）可表示为：

$$\frac{1}{n}\sum_{i=1}^n f_\mathbf{U}(x_i) g_\mathbf{V}(y_i)^\top = \mathbf{\Sigma}. \tag{5}$$

（5）中的形式很有用，因为它自然地引出了一个损失函数，用于寻找能够提供（$\mathbf{U}$ 和 $\mathbf{V}$ 分解的“参数”，即（2）。具体而言，如果我们需要 top-k 分解，令 $\widehat{\mathbf{U}} \in \mathbb{R}^{|\mathcal{S}|\times k}$ 和 $\widehat{\mathbf{V}} \in \mathbb{R}^{|\mathcal{S}|\times k}$ 分别为 top-k 左奇异向量和右奇异向量的估计值，并令 $\mathbf{\Sigma}$（$\widehat{\mathbf{U}}$ $\widehat{\mathbf{V}}$）$\in \mathbb{R}^{k\times k}$ 定义如下：

$$\mathbf{\Sigma}\left(\widehat{\mathbf{U}},\widehat{\mathbf{V}}\right) \coloneqq \frac{1}{n}\sum_{i=1}^n f_{\widehat{\mathbf{U}}}(x_i) g_{\widehat{\mathbf{V}}}(y_i)^\top,$$

其中 $f_{\widehat{\mathbf{U}}}: \mathcal{S} \to \mathbb{R}^k$和 $g_{\widehat{\mathbf{V}}}: \mathcal{S} \to \mathbb{R}^k$。为了执行 top-k 分解，现在可以搜索$\mathbf{U}$ 和 $\widehat{\mathbf{V}}$，使 $\mathbf{\Sigma}(\mathbf{U},\mathbf{V})$ 具有 $\mathbf{\Sigma}$ 的性质，即 $\mathbf{\Sigma}(\mathbf{U},\mathbf{V})$ 是一个对角矩阵，其对角线上具有较大的值。具体而言，我们考虑如下优化问题：给定 $\widehat{\mathbf{U}}^\top\widehat{\mathbf{U}} = \widehat{\mathbf{V}}^\top\widehat{\mathbf{V}} = \mathbf{I}$，最大化 $\mathbf{\Sigma}(\mathbf{U},\mathbf{V})$ 的对角元素，同时使非对角元素为 0，

$$\mathbf{U}^*, \mathbf{V}^* \coloneqq \arg\min_{\widehat{\mathbf{U}},\widehat{\mathbf{V}}} -\sum_{i=1}^k \mathbf{\Sigma}(\widehat{\mathbf{U}},\widehat{\mathbf{V}})_{[i,i]} \quad \text{s.t. } \mathbf{\Sigma}(\widehat{\mathbf{U}},\widehat{\mathbf{V}})_{[i,j]} = 0 \quad \forall i \neq j \tag{6}$$

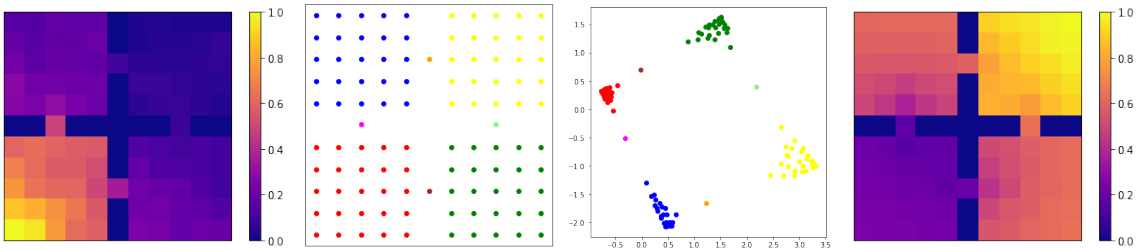

图 2。（**A**）使用随机策略收集的数据中的状态访问频率。（**B**）4rooms 中状态的参考颜色。（**C**）从 121 维表格表示中学习到的二维表示。（**D**）使用学习到的表示构建的奖励。图例：数值在 0 到 1 之间归一化，较深的颜色表示较低的数值。

所得函数 $f_{\mathbf{U}^*}$ 为状态提供了表示。例如，类似于图 1，$f_{\mathbf{U}^*}(x_s) = \mathbf{U}^{*\top} x_s$，它是 $\mathbf{U}^* \in \mathbb{R}^{|S|\times k}$中与状态 s 相对应的行，因为 x_s 是独热的。函数 $g_{\mathbf{V}^*}$ 仅用于搜索函数 $f_{\mathbf{U}^*}$。在图2(c) 中，我们给出了四房间领域的一个示例，可以观察到，学习到的表示会将具有相似转移分布的状态聚集在一起（即使$\mathbf{D}$ 远非均匀分布）。

探索： 现在我们来讨论如何利用从 $\mathbf{DP}_\pi$ 的 SVD 分解中获得的表示来促进探索。接下来，我们将展示，只需计算学得的状态表示的范数，就能即时估计某个状态相对于其他状态的访问频率伪计数。

令分解为 $\mathbf{DP}_\pi = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$，并且与前面一样，令表示函数 $f_\mathbf{U}$ 采用左奇异值 $\mathbf{U}$进行线性参数化。令 $\mathbf{\Lambda} := \mathbf{\Sigma}^2$，则表征的加权范数 $\|f_\mathbf{U}(x_s)\|_{\mathbf{\Lambda}^{-1}}$ 与 $d(s)$ 成反比，即 出现的频率 s 相对于其他状态而言，该状态在数据集中被观测到的频率。

定理 3。 如果$\mathbf{P}_\pi \mathbf{P}_\pi{}^\top$ 和$\mathbf{D}$ 可逆，那么对于 $\alpha_s := (\mathbf{P}_\pi \mathbf{P}_\pi{}^\top)^{-1}_{[s,s]}$，

$$\left\|f_\mathbf{U}(x_s)\right\|^2_{\mathbf{\Lambda}^{-1}} = \frac{\alpha_s}{d(s)^2}. \tag{7}$$

这里 α_s 是一个依赖于状态且独立于$d(s)$的常数。定理 3 尤其值得关注，因为 f_U 不仅能提供尊重下一状态转移相似性的表征，还能对表征进行定向，使其范数可用于表示伪计数。此外，估算奖励只需计算$\left\|f_\mathbf{U}(x_s)\right\|^2_{\mathbf{\Lambda}^{-1}}$，其中 $\mathbf{\Lambda}$是一个对角矩阵，可在 $O(k)$ 时间内求逆，其中 k 是表征的维度。因此，计算范数也是一个耗时 $O(k)$ 的操作。图2(D) 展示了一个示例，说明 $\|f_\mathbf{U}(x_s)\|_{\mathbf{\Lambda}^{-1}}$在四房间领域中所捕捉的内容。

注 1。 直观而言，表征的旋转或重新定向方式存在额外的自由度：表征之间的相对距离保持不变，但其范数会发生变化。在(7) 中利用了这种自由度，对表征重新定向，使得访问频率较高的状态往往具有范数较低的表征。这一点可通过图*2(C)* 和图*3(C)* 直观地看出。

备注 2。 或者，可以注意到 $\mathbf{DP}_\pi$ 与（归一化的）计数矩阵（*Bellemare* 等人，*2016*年）相关。因此，对其进行分解会得到同时具有D 所表示的访问频率和P 所表示的转移结构性质的表征π。

4. 函数近似

虽然（6）给出了问题的形式化表述，但并未提供一种可扩展的方法来执行优化。此外，我们希望将该过程扩展到能够与深度神经网络配合使用。下面我们将详细介绍一种实现这一目标的方法。

表示学习 利用（5）中的形式，我们使用非线性函数扩展这一思想 $f_{\theta_1}: \mathbb{R}^d \to \mathbb{R}^k$以及 $g_{\theta_2}: \mathbb{R}^d \to \mathbb{R}^k$ 它们由权重参数化$\theta := [\theta_1, \theta_2]$ 并以状态（不必是表格形式）作为输入，

$$\mathbf{\Sigma}(\theta) := \frac{1}{n}\sum_{i=1}^{n} f_{\theta_1}(s_i) g_{\theta_2}(s_i')^\top \in \mathbb{R}^{k\times k}. \tag{8}$$

具体而言，设

$$\mathcal{L}_{\text{diag}}\left(\mathbf{\Sigma}(\theta)\right) := \frac{1}{k}\sum_{i=1}^{k} \mathbf{\Sigma}(\theta)_{[i,i]}, \tag{9}$$

$$\mathcal{L}_{\text{off}}\left(\mathbf{\Sigma}(\theta)\right) := \frac{1}{k^2-k}\sum_{\substack{i,j=1\\ i\neq j}}^{k} \mathbf{\Sigma}(\theta)^2_{[i,j]}. \tag{10}$$

现在，类似于（6），我们通过结合（9）和（10），定义一个带有软约束的损失

$$\mathcal{L}(\mathbf{\Sigma}(\theta)) := -\mathcal{L}_{\text{diag}}(\mathbf{\Sigma}(\theta)) + \lambda_r \mathcal{L}_{\text{off}}(\mathbf{\Sigma}(\theta)), \tag{11}$$

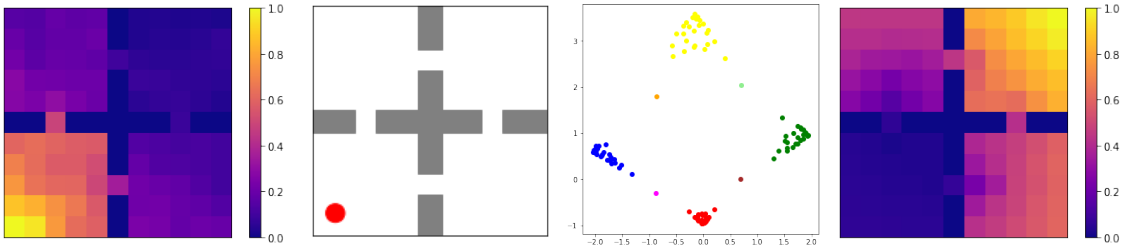

图 3. （A）使用随机策略收集的数据中的状态访问频率。（B）用作状态的领域图像示例。（C）与各状态对应的图像（30x30x3）学习到的二维表示。状态的参考颜色与图 2（B）中所用颜色相同。（D）使用学习到的表示构造的奖励。图例：数值归一化至 0 至 1，颜色越深表示数值越低。

其中 λ_r 是一个超参数。对于（6）中的表格设置，$\hat{\mathrm{U}}^\top\hat{\mathrm{U}} = \hat{\mathrm{V}}^\top\hat{\mathrm{V}} = \mathrm{I}$ 中的条件通过确保 U 的列（即每个特征维度）和 V 的列的范数为 1，防止对角线被无限制地最大化。在函数逼近设置中，这一条件并不显而易见。因此，为了防止（9）中的 $\boldsymbol{\Sigma}(\theta)_{[i,i]}$ 被无限制地最大化，我们对 f 和 g 的输出使用批量归一化（Ioffe 和 Szegedy，2015），以确保每个特征维度都保持归一化。在图 3(C) 中，我们展示了使用卷积神经网络、批量归一化和 $\lambda_r = 1$ 最小化（11）中的损失后学到的表征。

小批量优化：实践中，使用所有 $\boldsymbol{\Sigma}(\theta)$ 计算（8）中的 n 个样本并不实际。相反，最好能得到 $\boldsymbol{\Sigma}(\theta)$ 的小批量估计，

$$\widehat{\boldsymbol{\Sigma}}(\theta) := \frac{1}{b}\sum_{i=1}^{b} f_{\theta_1}(s_i) g_{\theta_2}(s'_i)^\top, \tag{12}$$

其中 $b \ll n$ 是小批量大小。不幸的是，可以证明，虽然 $\mathcal{L}_{\mathrm{diag}}(\boldsymbol{\Sigma}(\theta))$ 的梯度是无偏的，但一般而言，$\mathcal{L}_{\mathrm{off}}(\boldsymbol{\Sigma}(\theta))$ 的梯度会有偏。

定理4. $\mathcal{L}_{\mathrm{diag}}(\boldsymbol{\Sigma}(\theta))$ 的梯度是 $\mathcal{L}_{\mathrm{diag}}(\boldsymbol{\Sigma}(\theta))$的无偏估计量，而 $\mathcal{L}_{\mathrm{off}}(\boldsymbol{\Sigma}(\theta))$ 的梯度通常是 $\mathcal{L}_{\mathrm{off}}(\boldsymbol{\Sigma}(\theta))$ 的有偏估计，即，

$$\mathbb{E}\left[\partial_\theta \mathcal{L}_{\mathrm{diag}}\left(\widehat{\boldsymbol{\Sigma}}(\theta)\right)\right] = \partial_\theta \mathcal{L}_{\mathrm{diag}}(\boldsymbol{\Sigma}(\theta)),$$
$$\mathbb{E}\left[\partial_\theta \mathcal{L}_{\mathrm{off}}\left(\widehat{\boldsymbol{\Sigma}}(\theta)\right)\right] \neq \partial_\theta \mathcal{L}_{\mathrm{off}}(\boldsymbol{\Sigma}(\theta)),$$

其中，期望是对小批量的随机性求取的。

这个问题可归因于式（10）中，

$$\boldsymbol{\Sigma}(\theta)^2_{[i,j]} = \left(\frac{1}{n}\sum_{k=1}^{n} f_\theta(s_k)\, g_\theta(s'_k)^\top_{[i,j]}\right)^2,$$

平方项位于期望之外。这让人想起残差梯度优化均方贝尔曼误差时遇到的问题（Baird，1995）。因此，为了解决这个问题，我们也使用双重采样技巧来获得无偏梯度估计。具体来说，令

$$\widehat{\widehat{\boldsymbol{\Sigma}}}(\theta) := \left(\frac{1}{b_2}\sum_{i=1}^{b_2} f_{\theta_1}(s_i) g_{\theta_2}(s'_i)^\top\right), \tag{13}$$

使用 b_2 与（12）中所用数据批次不同的批次样本进行估计。然后利用（12）和（13），我们在（10）中使用$\widetilde{\mathcal{L}}_{\mathrm{off}}(\theta)$ 代替 $\mathcal{L}_{\mathrm{off}}(\boldsymbol{\Sigma}(\theta))$，其中

$$\widehat{\mathcal{L}}_{\mathrm{off}}(\theta) := \frac{1}{(k^2-k)}\sum_{\substack{i,j=1\\ i\neq j}}^{k} \widehat{\boldsymbol{\Sigma}}(\theta)_{[i,j]}\ \mathrm{sg}\left(\widehat{\widehat{\boldsymbol{\Sigma}}}(\theta)_{[i,j]}\right)$$
$$+ \frac{1}{(k^2-k)}\sum_{\substack{i,j=1\\ i\neq j}}^{k} \widehat{\widehat{\boldsymbol{\Sigma}}}(\theta)_{[i,j]}\ \mathrm{sg}\left(\widehat{\boldsymbol{\Sigma}}(\theta)_{[i,j]}\right) \tag{14}$$

其中，sg 表示停止梯度算子。

定理 5. 梯度 $\partial_\theta \widehat{\mathcal{L}}_{\mathrm{off}}(\theta)$ 是 $\partial_\theta \mathcal{L}_{\mathrm{off}}(\boldsymbol{\Sigma}(\theta))$的无偏估计量，即，

$$\mathbb{E}\left[\partial_\theta \widehat{\mathcal{L}}_{\mathrm{off}}(\theta)\right] = \partial_\theta \mathcal{L}_{\mathrm{off}}(\boldsymbol{\Sigma}(\theta)).$$

探索：在开发出用于获得表征的小批量优化流程后，我们现在将（7）扩展到函数逼近场景。具体而言，利用$\boldsymbol{\Lambda}(\theta) := \boldsymbol{\Sigma}^2(\theta)$，我们使用 $\|f_{\theta_1}(s)\|_{\boldsymbol{\Lambda}(\theta)^{-1}}$来构造伪计数的估计值。

然而，正如前文所述，在实践中精确求得 $\boldsymbol{\Sigma}(\theta)$ 可能很困难。因此，在实践中，(a) 我们会针对在多个小批量中观测到的不同 $\boldsymbol{\Sigma}(\theta)$，持续更新其$\widetilde{\boldsymbol{\Sigma}}(\theta)$均值估计 $\widehat{\boldsymbol{\Sigma}}(\theta)$。此外，(b) 由于式 (11) 中的损失会促使 $\widetilde{\boldsymbol{\Sigma}}(\theta)$ 成为对角矩阵，我们只持续更新其对角元素的估计$\widetilde{\boldsymbol{\Sigma}}(\theta)$，并将非对角$\widetilde{\boldsymbol{\Sigma}}(\theta)$元素设为零。现在，利用 $\boldsymbol{\Lambda}(\theta) := \boldsymbol{\Sigma}^2(\theta)$，我们构造

使用 s 得到的 $\|f_{\theta_1}(s)\|_{\tilde{\mathbf{\Lambda}}(\theta)^{-1}}$的伪计数估计。在图 3 (D) 中，我们展示了使用此过程估计伪计数的有效性。

5. 部分可观测马尔可夫决策过程与预测状态表示

所提出方法的一个重要优势是，我们还可以考虑其他矩阵（例如$\mathbf{P} \in \mathbb{R}^{|S||A|\times|S|}$ 而不是$\mathbf{P}_\pi \in \mathbb{R}^{|S|\times|S|}$）。在本节中，我们更进一步，考虑适用于部分可观测场景的系统动力学矩阵 。

简要回顾：对于一个部分可观测系统，令 $\mathcal{O}$ 表示可能的观测集合，并令 $O_t \in \mathcal{O}$ 表示时间 t 时观测的随机变量。令 $\mathcal{H}$ 表示观测-动作序列的集合，使得 $\mathcal{H} = \bigcup_{t\in\{0,1,\ldots,T\}} \mathcal{H}_t$，其中 T 是时域长度，$\mathcal{H}_0 = \mathcal{O}$，且 $\forall t > 0$、$\mathcal{H}_t = \mathcal{H}_{t-1} \times \mathcal{A} \times \mathcal{O}$。测试 $\tau \in \mathcal{H}$ 是任意的动作与观测序列（a_1、o_1、....、a_j、o_j）。给定历史 $h = (o_0, a_0, ..., a_{i-1}, o_i) \in \mathcal{H}$，令 $p(\tau|h) := \Pr(O_{i+1} = o_1, ..., O_{i+j} = o_j | h, A_i = a_1, ..., A_{i+j-1} = a_j)$ 表示测试 τ 的历史条件预测。直观地说，$p(\tau|h)$ 编码了给定观测历史时未来结果的概率。系统动力学矩阵 $\mathbf{W} \in \mathbb{R}^{|\mathcal{H}|\times|\mathcal{H}|}$ 定义为其行对应历史、列对应测试，使得 $\mathbf{W}[i,j] := p(\tau_j|h_i)$。图 4 展示了一个低秩系统动力学矩阵 $\mathbf{W} = \mathbf{U}\mathbf{V}'$，其中 $\mathbf{U} \in \mathbb{R}^{|\mathcal{H}|\times k}$，$\mathbf{V}' \in \mathbb{R}^{k\times|\mathcal{H}|}$。此外，令 $(z_i)_{i=1}^k$ 表示系统的潜变量。

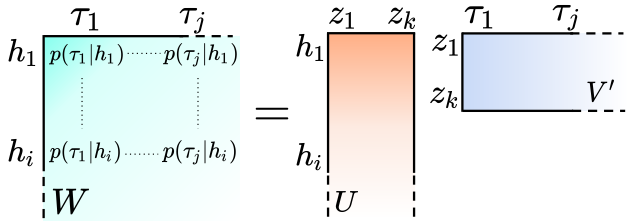

图 4. 低秩系统动力学矩阵（改编自 Singh 等人 （2012）的研究）。直观地说，行 $\mathbf{U}[i]$ 为历史 h_i 提供了良好的表示，因为它包含了在给定 h_i的情况下推断所有未来（测试）的结果概率所需的全部信息。

从不同角度看待 z_i，可以带来不同的见解。如果一个部分可观测马尔可夫决策过程有 k 未观测状态，那么 z 可以对应系统的隐状态。相应地，行 $\mathbf{U}[i]$ 表示给定截至目前的历史 h，这些隐状态上的信念分布；而 h_i 列 $\mathbf{V}'$则包含智能体处于某个给定隐状态时观测到未来/测试的概率。相比之下，预测状态表示将z 视为核心测试 (Littman 和 Sutton，2001；Singh 等人，2003；2012；James 和 Singh，2004；Rudary 和 Singh，2003），这样所有其余测试的概率都可以表示为这些核心测试发生概率的线性组合。因此，$\mathbf{U}[i]$ 编码了给定历史 h_i 时这些 k 核心测试发生的概率，而 $\mathbf{V}'$ 的各列则包含线性组合的系数。从数学上看，与标准的部分可观测马尔可夫决策过程视角相比，这一细微差异使 $\mathbf{V}'$ 的取值不必受概率分布约束，也不必为正值，因此预测状态表示能够对范围更广的动力系统进行建模（Singh 等人，2012）。变换预测状态表示（TPSR）进一步拓展了这一视角，使 $\mathbf{U}[i]$ 对应于对大量测试概率进行线性组合后得到的少量结果。从数学上看，关键区别在于，这使 $\mathbf{U}[i]$不必是概率分布，也可以取负值。这一视角尤其有用，因为现在对 $\mathbf{W}$ 进行奇异值分解所得的左奇异向量，能够直接提供从历史中推断未来结果所需的充分信息（Rosencrantz 等人，2004）。

5.1. De-Rex：基于分解的表示与探索

我们基于 TPSR 的思想进行研究。TPSR 通常仅限于小规模设置（Rosencrantz 等人，2004；Boots 等人，2011；Boots 和 Gordon，2011），我们利用第 4 节提出的思想，借助函数近似器将其扩展到大规模设置。也就是说，在（8）中，我们使用 $f_{\theta_1}: \mathcal{H} \to \mathbb{R}^k$和 $g_{\theta_2}: \mathcal{H} \to \mathbb{R}^k$，它们分别以历史和测试作为输入，而不是分别以状态和下一状态作为输入。

具体而言，设数据集 $\mathcal{D}$ 由 n 条轨迹组成。对于 $\mathcal{D}$ 中的第 i^{th} 条轨迹（$o_0, a_0, ...a_{T-1}, o_T$），我们定义 $h_{ij} := (o_0, a_0, ..., o_j)$ 和 $\tau_{ij} := (a_j, o_{j+1}, ..., a_{T-1}, o_T)$。我们扩展（8），使得

$$\mathbf{\Sigma}(\theta) := \frac{1}{nT}\sum_{i=1}^{n}\sum_{j=0}^{T-1} f_{\theta_1}(h_{ij}) g_{\theta_2}(\tau_{ij})^\top \in \mathbb{R}^{k\times k}. \quad (15)$$

现在，使用来自(14)的双重采样技巧，并使用类似于(11)的损失，我们可以将 $\mathbf{\Sigma}(\theta)$ 在(15)中对角化。我们在附录 C中对此进行了更多讨论。

4 个房间（部分可观测马尔可夫决策过程）： 与第 3 节和第 4节类似，我们还将可视化学习到的 历史表征 以及对 历史伪计数的估计结果。为此，我们构建了一个领域，其中时刻 O_t 的观测 t对应于尺寸为 （30，30，3）的图像，该图像呈现该领域的俯视图（类似于图 3 （B））， 但以概率 p 智能体的标记（红点）会完全隐藏。因此，当标记隐藏时，观测 O_t 无法提供有关智能体位置的任何信息，而位置需要利用历史进行推断。有关该部分可观测马尔可夫决策过程领域的更多详细信息

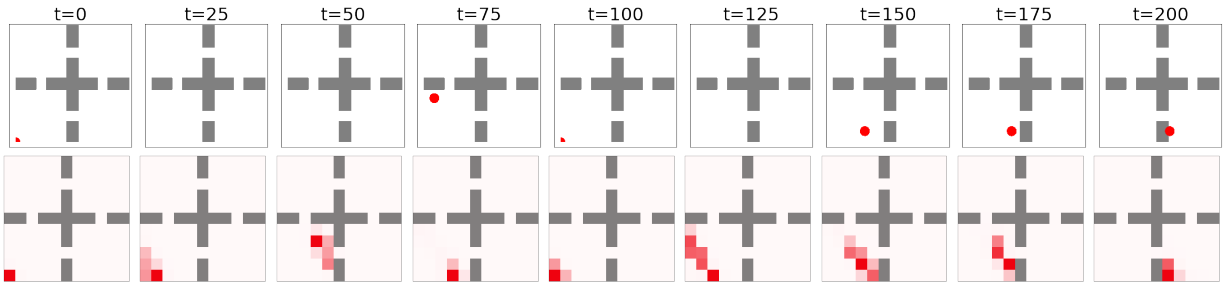

图 *5*。 与图 2 和 3类似，我们在图 9 （附录 D）中提供了绘图，以评估学习到的分布的质量。然而，对于高维中的表征而言，这并不能提供特别有用的信息。因此，作为替代方案，我们学习了一个分类模型（使用单层神经网络），旨在利用给定（部分）历史的表示，对该（部分）历史的观测序列末尾的真实潜在状态进行分类。这为我们提供了一个代理指标，用于衡量学习到的表示在多大程度上可用于解码关于潜在真实状态的信念。顶行展示了在时间步 $t = \{0, 25, 50, 75, 100, 125, 150, 175, 200\}$ 提供给智能体的观测快照（请注意，智能体在某些帧中缺失）。底部各行表示从截至时间步 t的历史表示中解码出的信念。详见正文。 **图例：** 对于底行，越深的红色表示越高程度的信念。对于顶行，红色表示智能体的位置。

请参见附录 D，其中我们还针对 $p = \{0.0, 0.3, 0.6, 0.9\}$的不同取值提供了消融研究。在此，我们展示 $p = 0.3$的结果。

任意历史的充分统计量是该历史结束时的真实状态。可以从图 9和 10 （附录D）中看出，对于结束于相似真实状态（智能体无法得知）的历史，所提出的 De-Rex 方法能够为这些历史提供相似的表示。此外，即使真实状态未知，所提出的方法也能够为潜在的真实状态提供伪计数。

在图5中，我们提供了一种评估所学历史表示质量的替代方法。可以看出，De-Rex 提供的表示能够有效解码潜在的真实状态。

6. 相关工作

近期Tang 等人（2022年）在理想化假设下说明了自预测表示（Guo 等人，2020年；2022年；Schwarzer 等人，2020年）与谱分解之间的关系。另一种做法是，Ren 等人（2022年）考虑显式分解转移核；然而，他们的方法使用了估计模型，而且其损失函数需要访问状态分布，而状态分布通常无法直接获取。此外，Touati 和 Ollivier（2021）；Lan 等人（2022）说明了长期占用矩阵的分解如何有助于表示学习，但他们没有解决探索问题。其中一些方法采用对比学习损失，这可能与本文提出的非对比损失互为补充（Garrido 等人，2022）。转移矩阵的谱性质也被用于学习有助于探索的选项（Machado 等人，2017a；b；Liu 等人，2017年）。此外，相关量的范数（称为后继表示）已被证明可用作伪计数（Machado 等人，2020年）。

我们在附录 A中详细讨论其他相关工作。或许与我们最相关的是 Wu 等人（2018年）关于如何分解 $\mathbf{P}_\pi$的拉普拉斯算子以获得相应特征函数的研究。与我们的方法类似，他们的方法也可以利用深度网络，且只需访问转移样本。然而，他们的方法基于使用对称的 拉普拉斯算子矩阵；根据领域的潜在结构（例如，当 $\mathbf{P}_\pi$ 不对称时），这一做法可能并不理想。此外，他们的方法也无法直接用于估计探索所需的伪计数。在第 7 节中，我们将通过实验与该方法进行比较。

7. 实证结果

在本节中，我们旨在通过实证研究所提出的 De-Rex 算法在大规模领域中的性质。具体而言，我们希望回答与表示学习和探索相关的两个主要问题。

问题 ***1***：***De-Rex*** 能否从丰富的观测中学习有用的表征？

为了探究这个问题，我们使用 DMLab 30（Beattie 等人，2016年）环境（Beattie 等人，2016年）。DMLab 在这里尤其适用，因为它是一个程序生成、部分可观测的 3D 世界，其中物体形状、颜色和位置等属性每个回合都可能发生变化。环境包含 30 种多样的导航和解谜任务，状态观测从语言指令到第一人称视角的像素图像不等，

以及奖励。动作集合由离散动作和连续动作组成，分别对应于改变智能体位置的移动、调整视觉角度，以及使用环境中各种可用工具的控制操作。

为了增加难度，我们考虑多任务设置：单个智能体同时与所有任务交互，并且必须学习能够涵盖所有这些任务的共同表征。在我们的设置中，没有向智能体提供明确的任务标识，因此智能体必须能够根据环境推断任务。我们考虑以下算法进行比较。有关这些算法的实现和超参数的详细讨论，请参见附录C。

强化学习基线：作为基础强化学习算法，我们使用V-MPO（Song 等人，2019年），这是一种策略优化方法，在涉及离散和连续任务的多个强化学习基准上取得了最先进的性能。为避免混杂因素，该基线智能体不包含任何明确的表示学习损失。

拉普拉斯分解：我们将该方法与Wu 等人（2018年）的工作进行比较。该方法通过分解与转移核相关联的拉普拉斯算子，利用谱图绘制来学习状态表示。作为一种基于分解的基线方法，我们将他们提出的损失作为 VMPO 智能体的辅助任务。

De-Rex 表示：与基于分解的基线方法类似，我们将基于（11）和（15）提出的辅助表示学习损失与VMPO 智能体结合使用。在这些实验中，我们仅关注理解所学习表示的实用性，不使用任何额外的探索奖励。

图中展示了针对每项任务对算法进行的实证比较 6。在附录E中，我们还提供了每项任务的汇总学习曲线和单独学习曲线。可以观察到，在大规模、丰富观测场景中，不显式使用任何表示学习损失的基线VMPO 难以取得良好效果。相比之下，使用拉普拉斯算子分解来学习状态表示的基线方法，性能略有提升。然而，所提出的 De-Rex 表示学习方法旨在显式地

处理部分可观测性带来了最大的性能提升。我们发现，在这些任务中，添加 De-Rex 探索奖励并未 改善实验结果，因为在稠密奖励设置下，额外奖励可能弊大于利，反而会分散注意力。

***Q2. De-Rex* 能否在稀疏奖励的复杂环境中进行探索？**

解决大规模领域需要同时具备两项能力：学习表征和高效探索。为了研究这种能力，我们使用 DMLab Hard-8 套件（Gulcehre 等人，2019年）

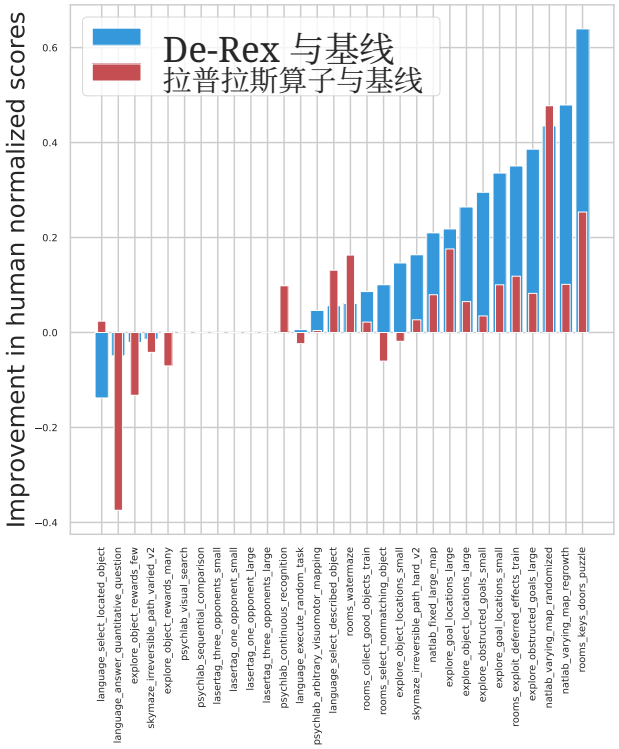

图 *6.* 逐游戏比较 (a) De-Rex 表示与 VMPO 基线，以及 (b) 拉普拉斯表征与 VMPO 基线。我们比较最终的人类归一化分数（即总计 $2 \cdot 10^{10}$ 训练步数最后 2% 的平均值）。每个分数均为三个随机种子的平均值。

该套件由8 个 DMLab 环境中的困难探索任务组成（Beattie 等人，2016年），这些任务具有稀疏奖励，要求智能体按照正确顺序解决一系列具有挑战性的子任务，才能获得最终奖励。与 DMLab-30 类似，我们通过采用多任务设置使该环境更具挑战性。对于所有智能体，我们通过在时间步 t 的奖励上添加依赖于算法的奖励来鼓励探索，$R'_t := R_t + \lambda_b$ 奖励$_t$，其中 λ_b 是一个超参数。除了与前文介绍的方法进行比较（这些方法不采用任何探索策略，并且其奖励$_t = 0$ 处处保持不变）之外，我们还比较以下方法。

随机网络蒸馏： RND（Burda 等人，2018年）已成为一种强大的基线探索方法，其做法是训练一个观测编码器，使其预测另一个独立的、固定的、随机初始化的编码器对同一观测的输出。预测损失随后用于计算奖励$_t$以进行探索。

De-Rex 探索： 所提出的方法的一个重要特点是，可以轻松获得状态访问伪计数的估计值。除了使用表示

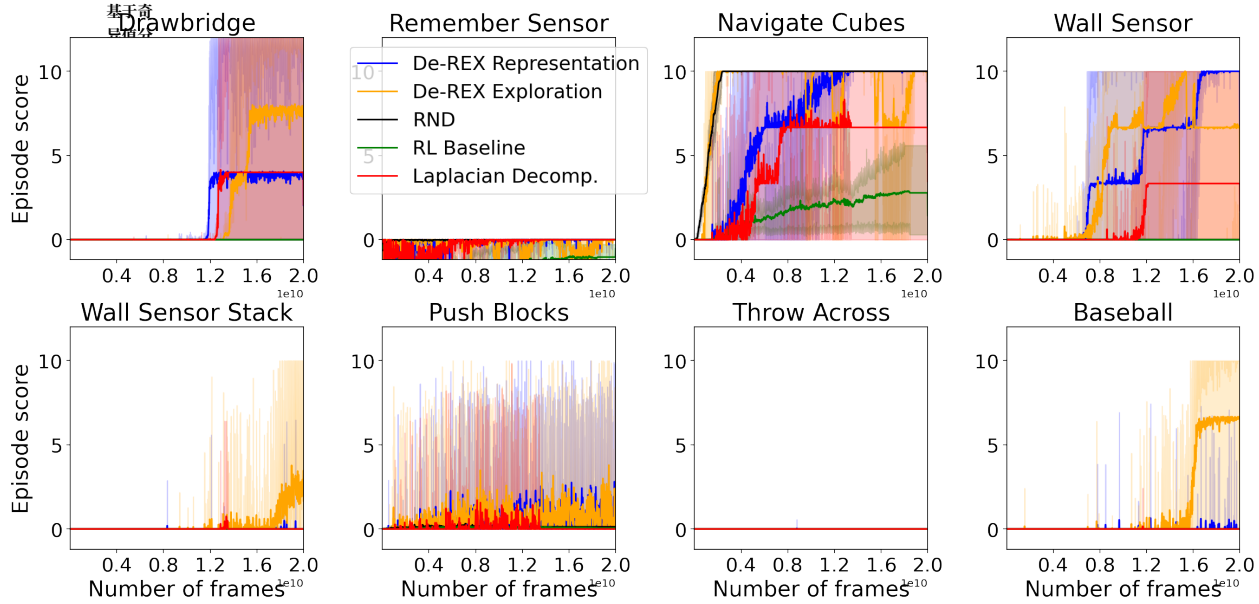

图 7。 多任务设置下 DM-Hard-8 套件中各任务的学习曲线（Gulcehre 等人，2019年）。阴影区域表示三个随机种子下的最小值和最大值。这些环境中的高方差是由于奖励极其稀疏（例如，只有在完成一系列复杂任务后，智能体才能获得任何奖励）。因此，如果智能体能够充分探索，就能解决任务；否则就会彻底失败。

损失外，我们还在时间步 t使用状态/历史表示的范数来设置奖励$_t$ 。

这些方法的实证比较结果如图 7 所示。可以观察到，基础 VMPO 方法无法解决任何任务，因为它没有任何明确鼓励探索的方式。类似地，RND 和拉普拉斯分解（Burda 等人，2018年）方法也只成功解决了 一个 DMHard-8 任务。这可能是因为 RND 中的探索奖励不包含任何表示学习组件，而该组件在丰富观测设置中至关重要。相比之下，拉普拉斯分解方法包含表征学习损失，但没有明确的探索鼓励机制。

De-Rex 为表示学习和估计用于鼓励探索的奖励提供了一套统一的流程。仅使用表示损失时，*De-Rex* 表示仍能解决近两项任务。同时使用表示损失和探索奖励时，*De-Rex* 探索能够解决近四项任务。附录E包含了使用不同超参数组合对 DM-Hard-8 任务进行的更多消融研究。

8. 讨论与结论

我们提出了一种基于奇异值分解的表示与探索过程。我们的方法有三个关键优势。(a) 简洁性：所提出的表示学习损失完全无模型，并且只需计算表示的范数，即可轻松构造探索奖励。(b) 理论与实证结果：所提出的方法以奇异值分解的基本框架为基础，并展示了奇异值分解不仅可用于表示学习，也可用于在具有挑战性的环境中进行探索。(c) 可扩展性：我们的方法不仅可应用于其他感兴趣的矩阵，还可将所提出的过程作为辅助损失，与未来任何基线（深度）强化学习算法结合使用。

局限性：虽然我们已初步对表格设定进行了分析，但我们认为，未来一项重要的研究方向是通过理论分析进一步理解函数逼近设定。此外，研究如何在表示学习过程中显式地考虑奖励可能会有所助益（将所提出的方法用作辅助损失，会隐式地依赖奖励，因为策略/价值函数的损失也会影响表示）。

9. 致谢

这项工作是在 YC 于 DeepMind 实习期间完成的。我们感谢 Mohammad Gheshlaghi Azar、Hado van Has- selt、Tom Schaul、Mark Rowland、Bernardo Avila Pires、BilalPiot、DanieleCalandriello、Michal Valko、David Abel、Marlos Machado、Doina Precup 和 Michael Bowling，感谢他们参与了多次富有启发性的讨论。

参考文献

A.Abdolmaleki、J. T. Springenberg、Y. Tassa、R. Munos、N. Heess 和 M. Riedmiller。最大后验策略优化。*arXiv* 预印本 *arXiv:1806.06920*，2018年。

A. Agarwal、S. Kakade、A. Krishnamurthy 和 W. Sun。Flambe：低秩马尔可夫决策过程的结构复杂度与表示学习。神经信息处理系统进展，33：20095–20107，2020年。

C. Allen、N. Parikh、O.Gottesman 和 G. Konidaris。学习用于深度强化学习的马尔可夫状态抽象。神经信息处理系统进展，34：8229–8241，2021。

M. G. Azar、B. Piot、B. A. Pires、J.-B. Grill、F. Altché 和 R. Munos。世界发现模型。*arXiv* 预印本 *arXiv:1902.07685*，2019年。

K. Azizzadenesheli、A. Lazaric 和 A. Anandkumar。使用谱方法对部分可观测马尔可夫决策过程进行强化学习。载于学习理论会议，第 193–256 页。PMLR，2016年。

A. P. Badia、P. Sprechmann、A. Vitvitskyi、D. Guo、B. Piot、S. Kapturowski、O. Tieleman、M. Arjovsky、A. Pritzel、A. Bolt 等。永不放弃：学习定向探索策略。*arXiv* 预印本 *arXiv:2002.06038*，2020年。

L. Baird。残差算法：结合函数逼近的强化学习。载于*1995*年机器学习论文集，第 30–37 页。Elsevier，1995年。

C. Beattie、J. Z. Leibo、D. Teplyashin、T. Ward、M. Wainwright、H. Küttler、A. Lefrancq、S. Green、V. Valdés、A. Sadik 等。DeepMind Lab。*arXiv* 预印本 *arXiv:1612.03801*，2016年。

B. Behzadian、S. Gharatappeh 和 M. Petrik。用于线性价值函数逼近的快速特征选择。载于自动规划与调度国际会议论文集，第29卷，第601–609页，2019年。

M. Bellemare、S. Srinivasan、G. Ostrovski、T. Schaul、D. Saxton 和 R. Munos。统一基于计数的探索与内在动机。神经信息处理系统进展，第29卷，2016年。

M. Bellemare、W. Dabney、R. Dadashi、A. Ali Taiga、P. S. Castro、N. Le Roux、D. Schuurmans、T. Lattimore 和 C. Lyle。强化学习最优表征的几何视角。神经信息处理系统进展，32，2019年。

B. B根 an d G. Gordon. An onli谱 l learning al 针对部分可观测非线性动力系统的算法。载于*AAAI*人工智能会议论文集，第25卷，第293–300页，2011年。B. Boots、S. M. Siddiqi 和 G. J. Gordon。利用预测状态表征闭合学习与规划的循环。《国际机器人研究杂志》，30(7):954–966，2011年。Y. Burda、H. Edwards、A. Storkey 和 O. Klimov。通过随机网络蒸馏进行探索。*arXiv*预印本 *arXiv:1810.12894*，2018年。

T. Chen、S. Kornblith、M. Norouzi 和 G. Hinton。用于视觉表征对比学习的简单框架。发表于国际机器学习会议，第 1597–1607 页。PMLR，2020年。

W. Dabney、A. Barreto、M. Rowland、R. Dadashi、J. Quan、M. G. Bellemare 和 D. Silver。价值改进路径：迈向更优的强化学习表征。发表于*AAAI* 人工智能会议论文集，第 35 卷，第 7160–7168 页，2021年。

P. De Handschutter、N. Gillis 和 X. Siebert。深度矩阵分解综述。《计算机科学评论》，42：100423，2021年。

Z. Deng、J. Shi 和 J. Zhu。NeuralEF：利用深度神经网络解构核函数。*arXiv* 预印本 *arXiv:2205.00165*，2022。

A. Erraqabi、M. C. Machado、M. Zhao、S. Sukhbaatar、A. Lazaric、L. Denoyer 和 Y. Bengio。时间抽象增强的时间对比学习：强化学习中的拉普拉斯算子的替代方法。*arXiv* 预印本 *arXiv:2203.11369*，2022。

L. Espeholt、H. Soyer、R. Munos、K. Simonyan、V. Mnih、T. Ward、Y. Doron、V. Firoiu、T. Harley、I. Dunning 等。Impala：采用重要性加权的 actor-learner 架构实现可扩展的分布式深度强化学习。发表于国际机器学习会议，第 1407–1416 页。PMLR，2018年。

Q. Garrido、Y. Chen、A. Bardes、L. Najman 和 Y. LeCun。对比与非对比自监督学习之间的对偶性。*arXiv* 预印本 *arXiv:2206.02574*，2022。

I. Gemp、B. McWilliams、C. Vernade 和 T. Graepel。EigenGame：作为均衡的主成分分析。*arXiv*预印本 *arXiv:2010.00554*，2020年。

D. Ghosh 和 M. G. Bellemare。用于稳定的离策略强化学习的表征。载于国际机器学习会议，第 3556–3565 页。PMLR，2020年。

K. Gregor、D. Jimenez Rezende、F. Besse、Y. Wu、H.Merzic 和 A. van den Oord。使用生成式环境模型塑造强化学习的信念状态。神经信息处理系统进展，32，2019年。

Y. Grinberg、H. Aboutalebi、M. Lyman-Abramovitch、B. Balle 和 D. Precup。从非均匀采样中学习预测状态表示。载于*AAAI*人工智能会议论文集，第 32 卷，2018年。

C. Gulcehre、T. Le Paine、B. Shahriari、M. Denil、M. Hoffman、H. Soyer、R. Tanburn、S. Kapturowski、N. Rabinowitz、D. Williams 等人。高效利用演示来解决棘手的探索问题。发表于 国际学习表征会议，2019年。

Z. D. Guo、S. Doroudi 和 E. Brunskill。用于情节式部分可观测马尔可夫决策过程的 PAC 强化学习算法。载于人工智能与统计，第 510–518 页。PMLR，2016年。

Z. D. Guo、B. A. Pires、B. Piot、J.-B. Grill、F. Altché、R. Munos 和 M. G. Azar。《用于多任务强化学习的自举潜在预测表征》。载于国际机器学习会议，第 3875–3886 页。PMLR，2020年。

Z. D. Guo、S. Thakoor、M. Pîslar、B. A. Pires、F. Altché、C. Tallec、A. Saade、D. Calandriello、J.-B. Grill、Y. Tang 等。《BYOL-Explore：通过自举预测进行探索》。*arXiv* 预印本 *arXiv:2206.08332*，2022。

P. D. Handschutter、N. Gillis 和 X. Siebert。深度矩阵分解综述。计算机科学评论，42:100423，2021年11月。doi: 10.1016/j.cosrev.2021.100423。网址 https://doi.org/10.1016%2Fj.cosrev.2021.100423。

M. Hessel、H. Soyer、L. Espeholt、W. Czarnecki、S. Schmitt 和 H. van Hasselt。采用 PopArt 的多任务深度强化学习。载于*AAAI*人工智能会议论文集，第 33 卷，第 3796–3803 页，2019年。

S. Hochreiter和J. Schmidhuber. 长短期记忆。神经计算, 9(8):1735–1780, 1997。

R. Houthooft、X. Chen、Y. Duan、J. Schulman、F. De Turck 和 P. Abbeel。Vime：最大化变分信息的探索。神经信息处理系统进展，29，2016年。

D. Hsu、S. M. Kakade 和 T. Zhang。用于学习隐马尔可夫模型的谱算法。计算机与系统科学杂志，78(5):1460–1480，2012年。

S. Ioffe 和 C. Szegedy。批归一化：通过减少内部协变量偏移加速深度网络训练。

载于国际机器学习会议，第 448–456 页。PMLR，2015 年。M. Jaderberg、V. Mnih、W. M. Czarnecki、T. Schaul、J. Z. Leibo、D. Silver 和 K. Kavukcuoglu。基于无监督辅助任务的强化学习。*arXiv* 预印本 *arXiv:1611.05397*，2016年。M. R. James 和 S. Singh。学习和发现带重置功能的动力系统中的预测状态表示。载于第二十一届国际机器学习会议论文集，第 53 页，2004 年。

J. Kwon, Y. Efroni, C. Caramanis 和 S. Mannor。《潜在马尔可夫决策过程的强化学习：遗憾保证与下界》。神经信息处理系统进展，34：24523–24534，2021年。

C. L. Lan, J. Greaves, J. Farebrother, M. Rowland, F. Pedregosa, R. Agarwal 和 M. G. Bellemare。《一种用于学习主子空间的新型随机梯度下降算法》。*arXiv* 预印本 *arXiv:2212.04025*，2022。

M. Littman 和 R. S. Sutton。《状态预测表征》。神经信息处理系统进展，14，2001年。

M. Liu, M. C. Machado, G. Tesauro 和 M. Camp- bell。《特征选项评论家框架》。*arXiv* 预印本 *arXiv:1712.04065*，2017年。

Q. Liu、A. Chung、C. Szepesvári 和 C. Jin。什么时候部分可观测强化学习并不可怕？*arXiv*预印本 *arXiv:2204.08967*，2022。

C. Lyle、M. Rowland、G. Ostrovski 和 W. 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。

C. Lyle、M. Rowland 和 W. Dabney。理解并防止强化学习中的容量损失。*arXiv* 预印本 *arXiv:2204.09560*，2022a。

C. Lyle、M. Rowland、W. Dabney、M. Kwiatkowska 和 Y. Gal。强化学习中的学习动力学与泛化。*arXiv* 预印本 *arXiv:2206.02126*，2022b。

M. C. Machado、M. G. Bellemare 和 M. Bowling。《强化学习中选项发现的拉普拉斯框架》。载于国际机器学习会议，第 2295–2304 页。PMLR，2017a。

M. C. Machado、C. Rosenbaum、X. Guo、M. Liu、G. Tesauro 和 M. Campbell。通过深度后继表示发现特征选项。*arXiv* 预印本 *arXiv:1710.11089*，2017b。

M. C. Machado、M. G. Bellemare 和 M. Bowling。利用后继表征进行基于计数的探索。收录于人工智能 *AAAI* 会议论文集，第 34 卷，第 5125–5133 页，2020年。

S. Mahadevan。原型价值函数：发展性强化学习。载于第*22*届国际机器学习会议论文集，第553–560页，2005年。

S. Mahadevan 和 M. Maggioni。原型价值函数：用于在马尔可夫决策过程中学习表示与控制的拉普拉斯框架。机器学习研究期刊，8(10)，2007年。

P. Mirowski、R. Pascanu、F. Viola、H. Soyer、A. J. Ballard、A. Banino、M. Denil、R. Goroshin、L. Sifre、K. Kavukcuoglu 等。学习在复杂环境中导航。*arXiv*预印本 *arXiv:1611.03673*，2016年。

D. Misra、M. Henaff、A. Krishnamurthy 和 J. Langford。运动学状态抽象与可证明高效的丰富观测强化学习。载于国际机器学习会议，第6961–6971页。PMLR，2020年。

O. Nachum、S.Gu、H. Lee 和 S. Levine。面向分层强化学习的近最优表征学习。*arXiv* 预印本 *arXiv:1810.01257*，2018年。

G. Ostrovski、M. G. Bellemare、A. Oord 和 R. Munos。利用神经密度模型进行基于计数的探索。发表于国际机器学习会议，第 2721–2730 页。PMLR，2017年。

T. L. Paine、C. Gulcehre、B. Shahriari、M. Denil、M. Hoff- man、H. Soyer、R. Tanburn、S. Kapturowski、N. Rabi- nowitz、D. Williams 等。有效利用示范来解决具有挑战性的探索问题。*arXiv* 预印本 *arXiv:1909.01387*，2019年。

R. Parr、C. Painter-Wakefield、L. Li 和 M. Litt man。分析用于价值函数近似的特征生成。在第*24*届国际机器学习会议论文集中，第737–744页，2007年。

R. Parr、L. Li、G. Taylor、C. Painter-Wakefield 和 M. L. Littman。对强化学习中的线性模型、线性价值函数逼近和特征选择的分析。载于第*25*届国际机器学习会议论文集，第752–759页，2008年。

D. Pathak、P. Agrawal、A. A. Efros 和 T. Darrell。通过自监督预测进行好奇心驱动的探索。载于国际机器学习会议，第 2778–2787 页。PMLR，2017年。

M. Petrik。对马尔可夫决策过程中价值函数近似的拉普拉斯方法的分析。载于国际人工智能联合会议，第2574–2579 页，2007年。T. Ren、T. Zhang、L. Lee、J. E. Gonzalez、D. Schuurmans 和 B. Dai。用于强化学习的谱分解表示。*arXiv* 预印本 *arXiv:2208.09515*，2022。M. Rosencrantz、G. Gordon 和 S. Thrun。学习低维预测性表示。载于第二十一届国际机器学习会议论文集，第 88 页，2004年。

N. Roy、G. Gordon 和 S. Thrun。通过信念压缩寻找近似部分可观测马尔可夫决策过程解。人工智能研究期刊，23：1–40，2005年。

M. Rudary 和 S. Singh。一种非线性预测状态表征。神经信息处理系统进展，16，2003。

J. Schmidhuber。模型构建神经控制器中实现好奇心与无聊的一种可能性。载于自适应行为模拟国际会议论文集：从动物到人工动物，第 222–227 页，1991年。

B. Schölkopf、A. Smola 和 K.-R. Müller。核主成分分析。载于国际人工神经网络会议，第 583–588 页。Springer，1997 年。

J. Schulman、S. Levine、P. Abbeel、M. Jordan 和 P. Moritz。信赖域策略优化。载于国际机器学习会议，第 1889–1897 页。PMLR，2015 年。

J. Schulman、F. Wolski、P. Dhariwal、A. Radford 和 O. Klimov。近端策略优化算法。*arXiv*预印本 *arXiv:1707.06347*，2017年。

M. Schwarzer、A. Anand、R. Goel、R. D. Hjelm、A. Courville 和 P. Bachman。利用自预测表示进行数据高效的强化学习。*arXiv* 预印本 *arXiv:2007.05929*，2020年。

S. Singh、M. James 和 M. Rudary。预测状态表示：一种用于建模动力系统的新理论。*arXiv* 预印本 *arXiv:1207.4167*，2012年。

S. P. Singh、M. L. Littman、N. K. Jong、D. Pardoe 和 P. Stone。学习预测状态表示。载于第 *20* 届国际机器学习会议（*ICML-03*）论文集，第712–719页，2003年。

H. F. Song、A. Abdolmaleki、J. T. Springenberg、A. Clark、H. Soyer、J. W. Rae、S. Noury、A. Ahuja、S. Liu、D. Tirumala 等。V-MPO：针对离散控制与连续控制的在策略最大后验

策略优化。*arXiv* 预印本 *arXiv:1909.12238*，2019。Z. Song、R. E. Parr、X. Liao 和 L. Carin。面向强化学习的线性特征编码。神经信息处理系统进展，29，2016。A. Srinivas、M. Laskin 和 P. Abbeel。Curl：面向强化学习的对比式无监督表征。*arXiv* 预印本 *arXiv:2004.04136*，2020。R. S. Sutton 和 A. G. Barto。《强化学习：导论》。麻省理工学院出版社，2018。H. Tang、R. Houthooft、D. Foote、A. Stooke、O. Xi Chen、Y. Duan、J. Schulman、F. DeTurck 和 P. Abbeel。# 探索：基于计数的深度强化学习探索研究。神经信息处理系统进展，30，2017。Y. Tang、Z. D. Guo、P. H. Richemond、B. Á. Pires、Y. Chan- dak、R. Munos、M. Rowland、M. G. Azar、C. L. Lan、C. Lyle 等。理解强化学习中的自预测学习。*arXiv* 预印本 *arXiv:2212.03319*，2022。A. Touati 和 Y. Ollivier。学习单一表征以优化所有奖励。神经信息处理系统进展，34：13–23，2021。C. F. Van Loan。广义化奇异值分解。《*SIAM* 数值分析杂志》，13(1)：76–83，1976。K. Wang、K. Zhou、Q. Zhang、J. Shao、B. Hooi 和 J. Feng。通过广义图绘制改进强化学习中的拉普拉斯表征。在国际机器学习会议上，页码 11003–11012。PMLR，2021。L. Wang、Q. Cai、Z. Yang 和 Z. Wang。嵌入以控制部分可观测系统：具有可证明样本效率的表征学习。*arXiv* 预印本 *arXiv:2205.13476*，2022。L. Weng。深度强化学习中的探索策略。在线：*https://lilianweng. github. io/lil-log/2020/06/07/exploration-strategies-in-deep-reinforcement-learning. html*，2020。Y. Wu、G. Tucker 和 O. Nachum。强化学习中的拉普拉斯算子：通过高效近似学习表征。*arXiv* 预印本 *arXiv:1810.04586*，2018。J. Zbontar、L. Jing、I. Misra、Y. LeCun 和 S. Deny。Barlow Twins：通过冗余消除进行自监督学习。在国际机器学习会议上，页码 12310–12320。PMLR，2021。W. Zhan、M. Uehara、W. Sun 和 J. D. Lee。用于预测状态表示的 PAC 强化学习。*arXiv* 预印本 *arXiv:2207.05738*，2022。

使用奇异值分解进行深度强化学习的表征与探索（补充材料）

A. 扩展相关工作

辅助损失与探索奖励：一些先前的研究分析了时序差分学习中状态特征的性质（Parr 等，2007年；2008年；Song 等，2016年），研究了不同的最优性概念（Nachum 等，2018年；Bellemare 等，2019年；Allen 等，2021），以及辅助任务对表示学习的影响（Dabney 等，2021；Lyle 等，2021；2022a；b）。具体而言，与 D-REX 类似，已有研究表明，辅助损失能够通过学习更好的表示，更快地获得更高的期望回报（Jaderberg 等，2016；Mirowski 等，2016；Hessel 等，2019；Gregor 等，2019；Guo 等，2020；Erraqabi 等，2022）。然而，与这些方法不同，D-REX 既不需要负采样，也不需要用于表示学习的目标网络，并且不要求在原始观测或表示层面进行重构。

辅助损失也可以提供内在动机信号以驱动探索（Houthooft 等人，2016年；Pathak 等人，2017年；Burda 等人，2018年；Azar 等人，2019年；Badia 等人，2020年；Guo 等，2022）。这类内在奖励也与状态访问的伪计数有关（Bellemare 等人，2016；Ostrovski 等人，2017年；Tang 等人，2017年）。然而，与这些方法不同，D-REX 无需任何复杂的密度估计过程，可以直接利用学习到的表示的范数来估计伪计数。关于各种探索策略的详尽综述，请读者参阅 Weng（2020年）的研究。

基于分解的方法 继 Mahadevan（2005年）；Mahadevan 和 Maggioni（2007年）的开创性工作之后，多位研究者探索了谱表示和分解在马尔可夫决策过程中的决策应用（Petrik，2007年；Wang 等人，2021；Behzadian 等人，2019年），以及在部分可观测马尔可夫决策过程中的应用（Hsu 等人，2012年；Roy 等人，2005年；Azizzadenesheli 等人，2016年；Boots 等人，2011年；Grinberg 等人，2018年）。研究者还利用动力学的各种秩条件，针对马尔可夫决策过程（Misra 等人，2020年；Agarwal 等人，2020年）和部分可观测马尔可夫决策过程（Wang 等人，2022；Kwon 等人，2021；Guo 等人，2016年；Liu 等人，2022；Zhan 等人，2022）提出了 PAC 可学习性保证和遗憾保证。与这些成果互为补充的是，我们的方法旨在解决大规模环境中的问题，这类环境在表示学习和探索方面都面临挑战。

B. 证明

定理 1. 如果$\mathbf{P}_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ 的特征值均为实数且互不相同，[2] 那么对于任意 k，$\mathbf{P}_\pi$ 和$(\mathbf{I} - \gamma\mathbf{P}_\pi)^{-1}$ 的前 k 个特征向量相同。

证明。 我们首先证明，$\mathbf{P}_\pi$ 的任意特征向量也是$(\mathbf{I} - \gamma\mathbf{P}_\pi)^{-1}$的特征向量。随后，我们将证明，根据这些特征向量对应特征值进行的排序也相同。

令 $v \in \mathbb{R}^{|S|\times 1}$ 为$\mathbf{P}_\pi$ 的特征向量， λ 为其对应特征值，即，

$$\mathbf{P}_\pi v = \lambda v.$$

现在可以观察到，$\mathbf{P}_\pi$ 和 $(\mathbf{I} - \gamma\mathbf{P}_\pi)^{-1}$ 的特征向量相同，

$$(\mathbf{I}-\gamma\mathbf{P}_\pi)^{-1}v = \left(\sum_{t=0}^{\infty}\gamma^t{\mathbf{P}_\pi}^t\right)v = \sum_{t=0}^{\infty}(\gamma\lambda)^t\, v.$$

进一步地，作为 $0 \le \gamma < 1$ 和 $|\lambda| \le 1$,

$$(\mathbf{I} - \gamma\mathbf{P}_\pi)^{-1}v = \frac{1}{1-\gamma\lambda}v.$$

令 $\lambda' \coloneqq 1/(1-\gamma\lambda)$ 为矩阵 $(\mathbf{I} - \gamma\mathbf{P}_\pi)^{-1}$ 中与 v 相关的特征值。由于 $|\lambda| \le 1$、 λ 和 λ' 单调相关，即 λ 的值越大， λ' 的值也越大，因此$\mathbf{P}_\pi$ 与 $(\mathbf{I} - \gamma\mathbf{P}_\pi)^{-1}$ 的前 k 个特征向量排序相同。

[2]为简化“前 k 个”特征值的选择，我们假设特征值均为实数且互不相同。

□

定理 2. $\mathbf{A_n}$ 是$\mathbf{DP}_\pi$的无偏相合估计量，即，

$$\forall i,j,\quad \mathbb{E}\left[\mathbf{A_n}_{[i,j]}\right]=\mathbf{DP}_{\pi[i,j]},\qquad\qquad \mathbf{A_n}_{[i,j]}\xrightarrow{a.s.}\mathbf{DP}_{\pi[i,j]}.$$

证明。**无偏：**

$$\mathbb{E}\left[\mathbf{A_n}\right]=\mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top\mathbf{Y}\right]=\mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top\mathbb{E}[\mathbf{Y}|\mathbf{X}]\right].$$

在表格设置中，由于 $\mathbb{E}[\mathbf{Y}|\mathbf{X}]$ 等于使用 π 从状态 X进行交互时下一状态 Y的分布，因此 $\mathbb{E}[\mathbf{Y}|\mathbf{X}]=\mathbf{P}_\pi$,

$$\mathbb{E}\left[\mathbf{A_n}\right]\overset{(a)}{=}\mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top\mathbf{X}\mathbf{P}_\pi\right]=\mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right]\mathbf{P}_\pi. \tag{16}$$

现在注意到，由于采用独热编码，$\frac{1}{n}\mathbf{X}^\top\mathbf{X}$ 是一个对角矩阵，其中 i^{th} 对角线元素由状态 i的归一化访问计数组成。因此,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right]_{[i,i]}&=\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n x_j x_j^\top\right]_{[i,i]}\\&=\mathbb{E}\left[\frac{1}{n}\sum_{j=1}^n 1_{\{s_j=s_i\wedge s_j=s_i\}}\right]\\&=\mathbb{E}\left[1_{\{S=s_i\}}\right]\\&=d(s_i).\end{aligned}$$

因此，$\mathbb{E}\left[\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right]=\mathbf{D}$，并使用（16），

$$\mathbb{E}\left[\mathbf{A_n}\right]=\mathbf{DP}_\pi.$$

一致性：

$$\begin{aligned}\mathbf{DP}_{\pi[i,j]}&=\mathbf{D}_{[i,i]}\mathbf{P}_{\pi[i,j]}\\&=d(s_i)\Pr(S'=s_j|S=s_i;\pi)\\&=\Pr(S'=s_j,S=s_i;\pi)\end{aligned}$$

$$\mathbf{A_n}_{[i,j]}=\frac{1}{n}\mathbf{X}^\top\mathbf{Y}_{[i,j]}=\frac{1}{n}\sum_{k=1}^n(x_k y_k^\top)_{[i,j]}$$

由于 x_k 和 y_k 是状态 S_k 和 S'_k的独热编码向量,

$$\mathbf{A_n}_{[i,j]}=\frac{1}{n}\sum_{k=1}1_{\{S_k=s_i\wedge S'_k=s_j\}}$$

令 $X_k=1_{\{S_k=s_i\wedge S'_k=s_j\}}$，由强大数定律可知,

$$\frac{1}{n}\sum_{k=0}^\infty X_k\xrightarrow{a.s.}\mathbb{E}[1_{\{S_k=s_i\wedge S'_k=s_j\}}]=\Pr(S'=s_j,S=s_i;\pi)$$

因此,

$$\begin{aligned}\forall i,j,\quad \mathbf{A_n}_{[i,j]}&\xrightarrow{a.s}\Pr(S'=s_j,S=s_i;\pi)\\&=\mathbf{DP}_{\pi[i,j]}.\end{aligned}$$

□

定理 3. 如果$\mathbf{P}_\pi {\mathbf{P}_\pi}^\top$ 且 $\mathbf{D}$ 可逆，则对于 $\alpha_s := (\mathbf{P}_\pi {\mathbf{P}_\pi}^\top)^{-1}_{[s,s]}$,

$$\left\|f_{\mathbf{U}}(x_s)\right\|^2_{\mathbf{\Lambda^{-1}}} = \frac{\alpha_s}{d(s)^2}.$$

证明。首先注意到,

$$\begin{aligned}
\mathbf{D}\mathbf{P}_\pi &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \\
\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top \\
&= \mathbf{U}\mathbf{\Sigma^2}\mathbf{U}^\top, && \because \mathbf{V}^\top\mathbf{V} = \mathbf{I} \\
&= \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top && \text{其中，} \mathbf{\Lambda} := \mathbf{\Sigma^2}.
\end{aligned}$$

回顾一下，对于任意满列秩矩阵 M，其伪逆为$\mathbf{M}^\dagger := (\mathbf{M}^\top\mathbf{M})^{-1}\mathbf{M}^\top$，使得$\mathbf{M}^\dagger\mathbf{M} = \mathbf{I}$。由于$(\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top)^\top = \mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top$,

$$\begin{aligned}
(\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top)^\dagger &= (\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top)^{-1}\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top \\
&= (\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top)^{-1}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \\
&\overset{(a)}{=} (\mathbf{U}\mathbf{\Lambda^2}\mathbf{U}^\top)^{-1}\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \\
&\overset{(b)}{=} \mathbf{U}\mathbf{\Lambda^{-2}}\mathbf{U}^\top\mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \\
&= \mathbf{U}\mathbf{\Lambda^{-1}}\mathbf{U}^\top, \qquad (17)
\end{aligned}$$

其中，(a) 是由$\mathbf{U}^\top\mathbf{U} = \mathbf{I}$这一事实得出的，而 (b) 是由$\mathbf{U^{-1}} = \mathbf{U}$这一事实得出的 $^\top$。

或者，在假设$\mathbf{P}_\pi\mathbf{P}_\pi^\top$和$\mathbf{D}$ 可逆的情况下,

$$\begin{aligned}
(\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top)^\dagger &= \left((\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top)^\top(\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top)\right)^{-1}(\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}^\top)^\top \\
&= \left(\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D^2}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D}\right)^{-1}\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D} \\
&= \mathbf{D^{-1}}(\mathbf{P}_\pi\mathbf{P}_\pi^\top)^{-1}\mathbf{D^{-2}}(\mathbf{P}_\pi\mathbf{P}_\pi^\top)^{-1}\mathbf{D^{-1}}\mathbf{D}\mathbf{P}_\pi\mathbf{P}_\pi^\top\mathbf{D} \\
&= \mathbf{D^{-1}}(\mathbf{P}_\pi\mathbf{P}_\pi^\top)^{-1}\mathbf{D^{-1}}. \qquad (18)
\end{aligned}$$

结合式（17）和式（18），

$$\mathbf{U}\mathbf{\Lambda^{-1}}\mathbf{U}^\top = \mathbf{D^{-1}}(\mathbf{P}_\pi\mathbf{P}_\pi^\top)^{-1}\mathbf{D^{-1}}.$$

在表格设置中，$x_s \in \mathbb{R}^{|S|\times 1}$ 对应于状态 s的独热特征编码，而线性函数$f_U(x_s) = \mathbf{U}^\top x_s$

$$\begin{aligned}
\langle f_U(x_s), f_U(x_s)\rangle_{\mathbf{\Lambda^{-1}}} &= x_s^\top\mathbf{U}\Lambda^{-1}\mathbf{U}^\top x_s \\
&= x_s^\top\mathbf{D}^{-1}(\mathbf{P}_\pi{\mathbf{P}_\pi}^\top)^{-1}\mathbf{D}^{-1}x_s \\
&\overset{(a)}{=} \frac{\alpha_s}{d(s)^2},
\end{aligned}$$

其中，$\alpha_s = (\mathbf{P}_\pi{\mathbf{P}_\pi}^\top)^{-1}_{[s,s]}$，步骤 (a) 中的化简是因为 x_s 是 one-hot 向量，且 $\mathbf{D}$ 是一个对角矩阵。 □

定理4。 $\mathcal{L}_{\text{diag}}$（$\mathbf{\Sigma}$（$\theta$））的梯度是 $\mathcal{L}_{\text{diag}}$（$\mathbf{\Sigma}$（$\theta$））的无偏估计量，而 $\mathcal{L}_{\text{off}}$（$\mathbf{\Sigma}$（θ））的梯度通常是 $\mathcal{L}_{\text{off}}$（$\mathbf{\Sigma}$（θ））的有偏估计，即，

$$\begin{aligned}
\mathbb{E}\left[\partial_\theta\mathcal{L}_{\text{diag}}\left(\widehat{\mathbf{\Sigma}}(\theta)\right)\right] &= \partial_\theta\mathcal{L}_{\text{diag}}(\mathbf{\Sigma}(\theta)), \\
\mathbb{E}\left[\partial_\theta\mathcal{L}_{\text{off}}\left(\widehat{\mathbf{\Sigma}}(\theta)\right)\right] &\neq \partial_\theta\mathcal{L}_{\text{off}}(\mathbf{\Sigma}(\theta)),
\end{aligned}$$

其中，期望是针对小批量的随机性计算的。

证明。 我们将（8）改写为

$$\boldsymbol{\Sigma}(\theta) \coloneqq \mathbb{E}\left[f_{\theta_1}(S) g_{\theta_2}(S')^\top\right],$$

其中，期望是针对 n 中的 $\mathcal{D}$采样转移元组所指定的分布计算的，即， $\boldsymbol{\Sigma}(\theta)$是 $\mathcal{D}$中 n 个样本的 $f_{\theta_1}(s_i) g_{\theta_2}(s_i')^\top$ 样本均值。

来自（10）和（12），

$$\begin{aligned}
\mathbb{E}\left[\partial_\theta \mathcal{L}_{\text{diag}}\left(\widehat{\boldsymbol{\Sigma}}(\theta)\right)\right] &= \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} \partial_\theta \widehat{\boldsymbol{\Sigma}}(\theta)_{[i,i]}\right] \\
&\overset{(a)}{=} \frac{1}{k}\sum_{i=1}^{k} \partial_\theta \mathbb{E}\left[\widehat{\boldsymbol{\Sigma}}(\theta)_{[i,i]}\right] \\
&= \frac{1}{k}\sum_{i=1}^{k} \partial_\theta \boldsymbol{\Sigma}(\theta)_{[i,i]} \\
&= \partial_\theta \mathcal{L}_{\text{diag}}(\boldsymbol{\Sigma}(\theta)),
\end{aligned}$$

其中，(a)成立是因为我们考察的是采样分布固定的情况。

下面我们用一个反例说明，基于非对角项的损失的小批量版本的梯度通常并不能提供无偏估计量。对于这个反例，令小批量大小为 $b = 1$。因此，

$$\begin{aligned}
\mathbb{E}\left[\partial_\theta \mathcal{L}_{\text{off}}\left(\widehat{\boldsymbol{\Sigma}}(\theta)\right)\right] &= \mathbb{E}\left[\frac{1}{k^2-k}\sum_{\substack{i,j=1 \\ i\neq j}}^{k} \partial_\theta \widehat{\boldsymbol{\Sigma}}(\theta)^2_{[i,j]}\right], \\
&= \frac{1}{k^2-k}\sum_{\substack{i,j=1 \\ i\neq j}}^{k} \partial_\theta \mathbb{E}\left[\widehat{\boldsymbol{\Sigma}}(\theta)^2_{[i,j]}\right], \\
&\overset{(b)}{=} \frac{1}{k^2-k}\sum_{\substack{i,j=1 \\ i\neq j}}^{k} \partial_\theta \mathbb{E}\left[f_{\theta_1}(S) g_{\theta_2}(S')^\top_{[i,j]}{}^2\right], \\
&\neq \frac{1}{k^2-k}\sum_{\substack{i,j=1 \\ i\neq j}}^{k} \partial_\theta \mathbb{E}\left[f_{\theta_1}(S) g_{\theta_2}(S')^\top_{[i,j]}\right]^2, \\
&= \partial_\theta \mathcal{L}_{\text{off}}\left(\boldsymbol{\Sigma}(\theta)\right),
\end{aligned}$$

其中，步骤 (b) 成立是因为所考虑的小批量大小为 1，而标记为 红色 的步骤 不成立，因为对于随机变量 X，一般而言 $\mathbb{E}[X^2] \neq \mathbb{E}[X]^2$。

□

定理 5。 梯度 $\partial_\theta \mathcal{L}_{\text{off}}(\theta)$ 是 $\partial_\theta \mathcal{L}_{\text{off}}(\boldsymbol{\Sigma}(\theta))$的无偏估计量，即

$$\mathbb{E}\left[\partial_\theta \widehat{\mathcal{L}}_{\text{off}}(\theta)\right] = \partial_\theta \mathcal{L}_{\text{off}}(\boldsymbol{\Sigma}(\theta)).$$

证明。

$$\widehat{\mathcal{L}}_{\text{off}}(\theta) \coloneqq \underbrace{\frac{1}{(k^2-k)}\sum_{\substack{i,j=1 \\ i\neq j}}^{k} \widehat{\boldsymbol{\Sigma}}(\theta)_{[i,j]}\ \operatorname{sg}\left(\widehat{\widehat{\boldsymbol{\Sigma}}}(\theta)_{[i,j]}\right)}_{(I)} + \underbrace{\frac{1}{(k^2-k)}\sum_{\substack{i,j=1 \\ i\neq j}}^{k} \widehat{\widehat{\boldsymbol{\Sigma}}}(\theta)_{[i,j]}\ \operatorname{sg}\left(\widehat{\boldsymbol{\Sigma}}(\theta)_{[i,j]}\right)}_{(II)},$$

$$
\begin{aligned}
\mathbb{E}\left[\partial_\theta(I)\right] &:= \frac{1}{(k^2-k)} \sum_{\substack{i,j=1 \\ i\neq j}}^{k} \mathbb{E}\left[\partial_\theta \widehat{\mathbf{\Sigma}}(\theta)_{[i,j]} \ \operatorname{sg}\left(\widehat{\widehat{\mathbf{\Sigma}}}(\theta)_{[i,j]}\right)\right], \\
&\overset{(a)}{=} \frac{1}{(k^2-k)} \sum_{\substack{i,j=1 \\ i\neq j}}^{k} \mathbb{E}\left[\partial_\theta \widehat{\mathbf{\Sigma}}(\theta)_{[i,j]}\right] \mathbb{E}\left[\widehat{\widehat{\mathbf{\Sigma}}}(\theta)_{[i,j]}\right], \\
&= \frac{1}{(k^2-k)} \sum_{\substack{i,j=1 \\ i\neq j}}^{k} \partial_\theta \mathbb{E}\left[\widehat{\mathbf{\Sigma}}(\theta)_{[i,j]}\right] \mathbb{E}\left[\widehat{\widehat{\mathbf{\Sigma}}}(\theta)_{[i,j]}\right], \\
&= \frac{1}{(k^2-k)} \sum_{\substack{i,j=1 \\ i\neq j}}^{k} \partial_\theta \mathbf{\Sigma}(\theta)_{[i,j]} \mathbf{\Sigma}(\theta)_{[i,j]}, \\
&= \frac{1}{2}\partial_\theta \left(\frac{1}{(k^2-k)} \sum_{\substack{i,j=1 \\ i\neq j}}^{k} \mathbf{\Sigma}(\theta)^2_{[i,j]} \right), \\
&= \frac{1}{2}\partial_\theta \mathcal{L}_{\text{off}}\left(\mathbf{\Sigma}(\theta)\right),
\end{aligned}
$$

其中$\widehat{\mathbf{\Sigma}}(\theta)$步骤使用不同的数据批次独立计算得出的。同样，可以观察到 $\mathbb{E}\left[\partial_\theta(I)\right] = \mathbb{E}\left[\partial_\theta(II)\right]$。因此，

$$
\begin{aligned}
\mathbb{E}\left[\partial_\theta \widehat{\mathcal{L}}_{\text{off}}(\theta)\right] &= \mathbb{E}\left[\partial_\theta(I)\right] + \mathbb{E}\left[\partial_\theta(II)\right] \\
&= \partial_\theta \mathcal{L}_{\text{off}}(\mathbf{\Sigma}(\theta)).
\end{aligned}
$$

□

C. 实验细节

我们进一步介绍论文中深度强化学习实验的算法细节和实现细节。

DM-Lab-30 环境。DM-Lab-30 是一组视觉复杂的迷宫导航任务（Beattie 等人，2016年）。智能体接收的唯一输入是环境第一人称视角的图像、自然语言指令（在某些任务中）以及奖励。智能体可以提供多个同时动作，以控制移动（前进/后退、向左/向右侧移、蹲伏、跳跃）、视线方向（向上/向下、向左/向右）以及标记操作（在包含对手机器人的激光枪战关卡中）。DM-Lab-30 已被证明对无模型强化学习算法具有挑战性（Beattie 等人，2016年；Guo 等人，2020年）。

DM-Hard-8 环境：DM-Hard-8 是一组视觉上复杂且相似的迷宫导航任务，具有稀疏奖励，常用作探索测试平台（Guo 等人，2022）。该测试平台包含 8 个高难度探索任务，旨在突出在具有部分可观测性、连续控制和高度多变初始条件的程序生成三维世界中，从稀疏奖励中学习所面临的挑战。每项任务都要求智能体通过与环境中的特定物体互动，到达环境中的一个苹果。只有到达苹果时，智能体才能获得奖励。由于环境是程序生成的，因此物体形状、颜色和位置等属性在每个回合中都各不相同。智能体只能看到其所在位置的第一人称视角。

智能体架构。 由于环境是部分可观测的，我们采用循环网络架构，遵循此前的相关工作，例如（Beattie 等人，2016年；Song 等人，2019年；Guo 等人，2020年）。具体而言，智能体在 $\mathcal{F}_1: \ \mathcal{O} \to \mathbb{R}^k$ 内部使用卷积神经网络，将输入 $o_t \in \mathcal{O}$ 中的图像处理为潜在嵌入。观测包含：（i）智能体的手当前施加的力，占总握力的比例，是一个介于 0 和

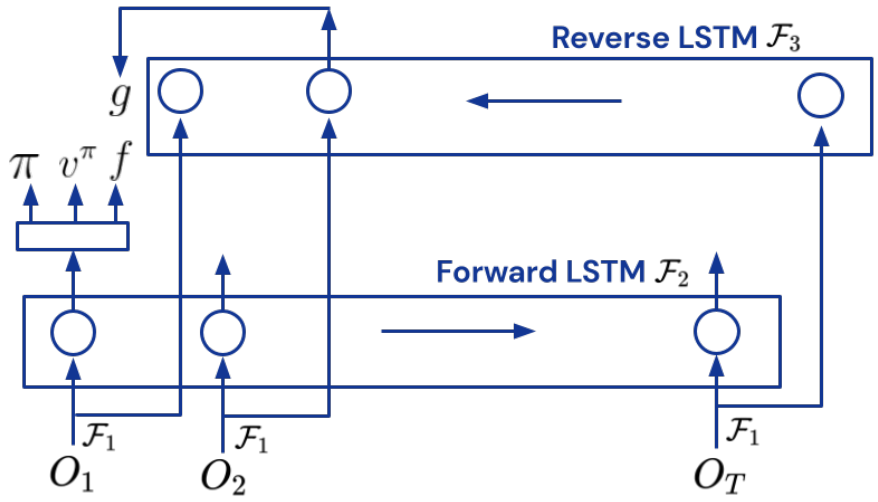

图8。回想一下，观测 O_t 包含先前的动作 A_{t-1}。设 $\mathcal{F}_1(O_i)$ 为时间步 i 的观测嵌入。设 $b_t = \mathcal{F}_2(\mathcal{F}_1(O_0)\ \mathcal{F}_1(O_1)\ ...,\mathcal{F}_1(O_t))$ 为通过正向长短期记忆网络得到的信念状态。设 $\mathcal{F}_3$ 为反向长短期记忆网络，使得 $b'_t = \mathcal{F}_3(\mathcal{F}_1(O_T)\ \mathcal{F}_1(O_{T-1})\ ...,\mathcal{F}_1(O_t))$。那么，式 (15) 中的 $f_{\theta_1}(h_t) \in \mathbb{R}^k$ 定义为 $f_{\theta_1}(b_t)$，而 $g_{\theta_2}(\tau_t)$ 定义为 $g_{\theta_2}(b'_{t+1})$。

1 之间的单个标量；（ii）一个布尔值，表示手当前是否正握着物体；（iii）智能体的手与主体之间的距离；（iv）智能体采取的上一个动作；这些量分别通过线性投影映射到 20 维，再经过 ReLU 处理。（v）智能体获得的上一个奖励，经过带符号双曲变换处理；（vi）针对当前任务的文本指令，其中每个词嵌入到 20 维，并由长短期记忆网络（Hochreiter 和 Schmidhuber，1997）依序处理，得到一个大小为 64 的嵌入。所有这些量以及 $\mathcal{F}_1$的输出会拼接在一起，并通过一个线性层映射为大小为 512 的嵌入。一个嵌入大小为 512 的长短期记忆网络 $\mathcal{F}_2$会依序处理这些观测表示，从而形成循环历史表示。长短期记忆网络的核心输出是信念 $b_t = \mathcal{F}_2（h_t）\in \mathbb{R}^k$。使用正文中的符号，我们可以将信念理解为历史 $h_t =$ （o_0， $a_0 ... o_t$）的函数，这得益于智能体架构的循环特性。

最后，我们使用多层感知机头，根据 $v_\theta(b_t)$ 信念计算价值函数，并计算策略 $\pi_\theta(a_t|b_t)$ 。此外，所有算法均基于 VMPO 构建，并采用与 RND 类似的奖励归一化方案（Burda 等人，2018年），我们使用其标准差的指数移动平均估计值对原始奖励进行归一化。具体步骤见 Guo 等人 （2022）的附录 A.3。对于 DM-Hard8 环境，在将奖励与奖励分别结合之前，我们先对奖励和奖励进行归一化，使用 λ_b。也就是说，设 R_t 和奖励$_t$ 分别为归一化后的奖励和奖励，则时间步 t 的有效奖励为$R_t + \lambda_b$ 奖励$_t$，其中 λ_b 是一个超参数。

当算法使用额外的表示损失 Rep$_{\text{loss}}$ 时，它会按以下方式与强化学习$_{\text{loss}}$ 结合：强化学习$_{\text{loss}} + w_{\text{loss}}Rep_{\text{loss}}$，其中 w_{loss} 是一个超参数，而强化学习$_{\text{loss}}$ 是基线 VMPO 为学习策略和评论家网络提供的损失。

VMPO 基线。 VMPO 是一种无模型强化学习算法（Song 等人，2019年），在多个具有挑战性的强化学习基准上取得了最先进的性能。VMPO 的核心是一种策略优化算法。值得注意的是，VMPO 更新的构建基于策略优化文献中的多项既有成果，例如信赖域优化（Schulman 等人，2015；2017）、基于回归的更新，以及基于原始-对偶优化调整信赖域（Abdolmaleki 等人，2018年），以及优势函数的多步估计（Espeholt 等人，2018年）。关于该算法及其超参数的完整说明，请参阅（Song 等人，2019年）。原始 VMPO 算法在 DM-Lab-30 实验中采用像素控制作为辅助目标（Beattie 等人，2016年）；为避免混杂效应，我们在实验中不使用这类辅助损失。此外，算法没有探索奖励。也就是说，$\lambda_b = w_{\text{loss}} = 0$。

De-Rex。 这是所提出的方法，它提供了一种统一的方式来学习表示，同时获取用于探索的伪计数。 *De-Rex* 表示 仅将表示学习损失用作辅助任务，不包含探索奖励，即 $\lambda_b = 0$。表示损失 $\text{Rep}_{\text{loss}} = \mathcal{L}(\boldsymbol{\Sigma}(\theta))$ 基于（11），并与 VMPO 基线结合。为实现（15），我们以正向和反向长短期记忆网络作为函数 f_{θ_1} 和 g_{θ_2}，如图 8所示。我们发现，将表示损失定义为针对信念的独立投影/变换所施加的辅助损失 b_t，比直接将损失应用于 b_t更有益；后者由策略和价值函数直接使用。

相比之下，*De-Rex* 探索同时采用表示学习损失和探索奖励。这里的奖励$_t$对应于表示的范数 $f_{\theta_1}(h_t) \in \mathbb{R}^k$ 如第5节所述。在表1和2 中，我们提供了有关超参数的更多细节。

超参数	值
λ_r	10.0（在较小规模的实验中调优）
λ_b	（不适用）
w_{loss}	0.01（超参数搜索 $\{1, 0.1, 0.01, 0.001\}$）
k	512（超参数搜索 $\{1024, 512, 64\}$）
批大小	256

表*1.* DMLab30 的 De-Rex 表示超参数

超参数	值
λ_r	10.0
λ_b	0.001（超参数搜索 $\{1, 0.1, 0.01, 0.001\}$）
w_{loss}	0.01（超参数搜索 $\{1, 0.1, 0.01, 0.001\}$）
k	64（超参数搜索 $\{512, 256, 64, 8\}$）
批大小	256

表 *2.* DMHard-8 的 De-Rex 探索超参数

拉普拉斯图绘制 (Wu 等人, 2018年)： 这是一种基线分解方法，用于分解与转移矩阵相关联的拉普拉斯算子。设 u 和 v 为状态与下一状态对，那么他们提出的损失为

$$\mathcal{L}_{\text{diag}} \coloneqq \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k}(f_{\theta_1}(u_i)_j - f_{\theta_1}(v_i)_j)^2$$

$$\mathcal{L}_{\text{off}} \coloneqq \frac{1}{nk^2}\sum_{i=1}^{n}\sum_{j=1}^{k}\sum_{l=1}^{k}(f_{\theta_1}(u_i)_j f_{\theta_1}(u_i)_l - \delta_{j,l})(f_{\theta_1}(v_i)_j f_{\theta_1}(v_i)_l - \delta_{j,l})$$

$$\text{Rep}_{\text{loss}} \coloneqq \mathcal{L}_{\text{diag}} + \lambda_r \mathcal{L}_{\text{off}}$$

原始公式针对马尔可夫决策过程定义，其中 u和 v 分别对应状态和下一状态。然而，由于我们处理的是部分可观测马尔可夫决策过程，因此令状态为截至目前的历史，于是设 $u = h$ 和 $v = h'$。函数 f_{θ_1} 的架构与 De-Rex 所用的相同。这种拉普拉斯基线不提供任何探索手段，因此始终使用 $\lambda_b = 0$。对于 λ_r，我们在 $[1, 10]$ 之间进行了搜索，发现 10 的效果最好。其他超参数与 De-Rex 类似。

随机网络蒸馏（RND）。RND（Burda 等人，2018年）提供了一种探索策略，根据当前网络 $\xi_\theta(o_t) \in \mathbb{R}^k$ 与随机初始化网络$\xi_{\theta_0}(o_t) \in \mathbb{R}^k$之间的预测生成探索奖励。请注意， θ_0 是随机初始化的，不会随着时间推移而训练。当前网络 θ 经过训练以最小化平方误差 $l_\theta(o_t) = \|\xi_\theta(o_t) - \xi_{\theta_0}(o_t)\|_2^2$。同时， $l_\theta(o_t)$也用作探索奖励。

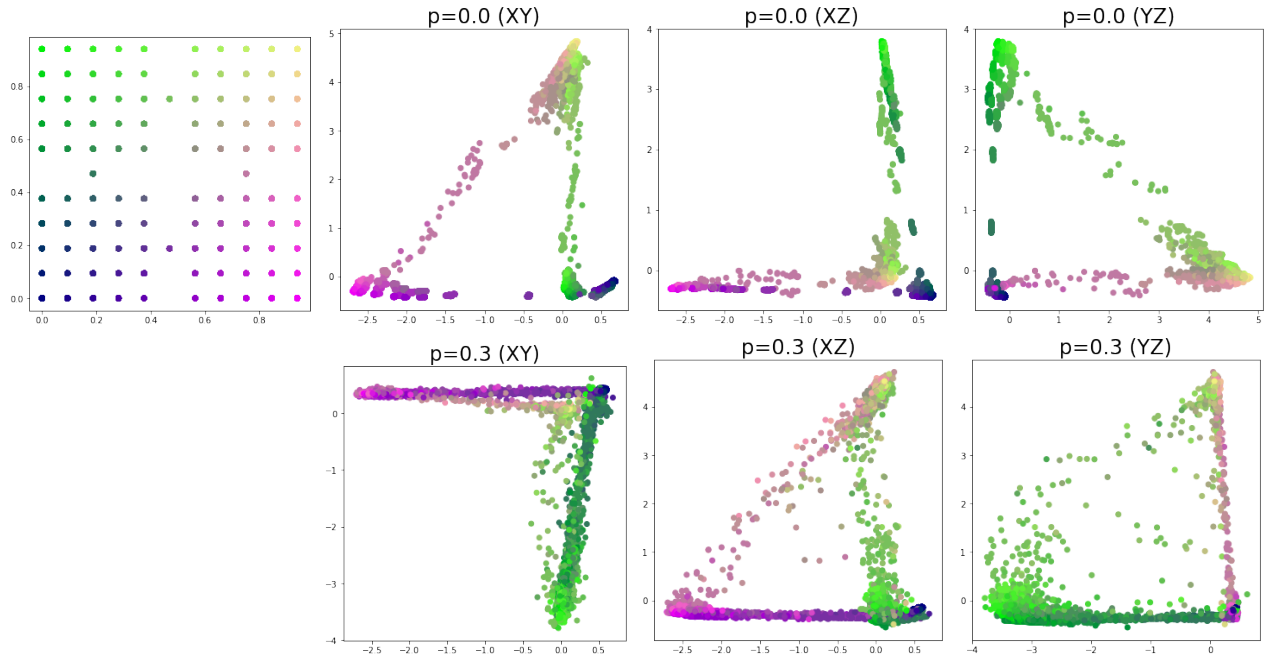

图*9*。此图与图2（B）和图2（C）类似，并将可视化技术扩展到部分可观测马尔可夫决策过程（POMDP）情境。我们绘制了在**左上**用于表示潜在真实状态的真实状态颜色 s下收集的数据中所有可能（部分）历史的表示。**上排**展示了四房间（POMDP）情境下历史的学习到的三维表示，其中$p = 0.0$。为查看三维表示，我们绘制坐标轴的切片（XY、XZ、YZ）。**下排**展示了四房间（POMDP）情境下历史的学习到的三维表示，其中 $p = 0.3$。为查看三维表示，我们绘制坐标轴的切片（XY、XZ、YZ）。**图例：**若（部分）历史在其观测序列结束时的真实潜在状态为 s ，则其表示与 s的参考颜色相同。因此，由于可能有多个历史以同一给定状态结束，对于真实状态的每一种颜色，都有多个颜色相同的学习到的表示。

D. 四房间部分可观测马尔可夫决策过程领域的消融研究

在图 2和 3 中，我们可视化了完全可观测设置下学习到的表征和获得的奖励。本节旨在通过对部分可观测设置进行消融实验，进一步扩展这项定性评估。为了专注于表示学习组件和构造的奖励，我们考虑了四房间领域的部分可观测扩展。这为测试所提出的 De-Rex方法的有效性和局限性提供了一个受控环境。

4个房间（部分可观测马尔可夫决策过程）：在此领域中，时刻的观测 O_t t对应于尺寸为（30，30，3）的图像，该图像展示了领域的俯视图（类似于图3 （B）），但以概率 p 智能体的标记（红点）会完全隐藏。因此，当标记被隐藏时，观测 O_t 无法提供有关智能体位置的任何信息，需要利用历史推断其位置。自然地， p 的值越高，部分可观测性程度越高，而 $p = 0$ 表示完全可观测（所提出的De-Rex方法不假设完全可观测性，仍使用依赖历史的表示）。动作集合对应于四个移动方向，时域长度为250。

因此，该领域提供了一个受控实验环境，以便理解该方法在不同程度的部分可观测性下的表现。接下来，我们将对本方法在不同 p 取值下学习到的表征和奖励进行定性分析。数据是使用均匀随机策略收集的。

表示： 在完全可观测情况下，所提出的方法使用了 2 的表示维度（图 3）。对于部分可观测设置，我们发现表示维度 3 和 4 分别对 $p < 0.5$ 和 $p > 0.5$很有用。在图 9中，我们展示了使用 De-Rex 方法学习表征的有效性。我们 f 和 g 函数的架构遵循图 8所示的架构，其中观测分别通过一个

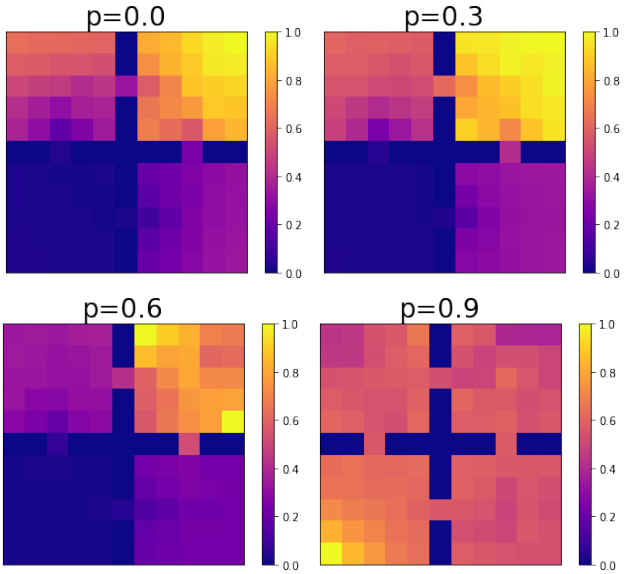

图*10*。 该图与图 2(D) 和图 3(D) 类似，并将可视化技术扩展到了部分可观测马尔可夫决策过程的情境。真实状态访问情况与图 2(A) 中相同。 **左上** 当 $p = 0.0$ 时，使用学习到的历史表示构造的奖励。 **右上** 当 $p = 0.3$ 时，使用学习到的历史表示构造的奖励。 **左下** 当 $p = 0.6$时，使用学习到的历史表示构造的奖励。 **右下** 当 $p = 0.9$ 时，使用学习到的历史表示构造的奖励。 **图例：** 颜色越亮表示奖励值越高。每个真实状态的颜色 s 根据所有（部分）历史的奖励平均值计算，这些历史 s 在其观测序列结束时具有该真实潜在状态。

卷积网络，然后分别传递给正向和反向长短期记忆网络，以编码历史和未来。随后，将这些长短期记忆网络的输出输入单层多层感知机，形成 f 和 g。

任意历史的充分统计量都是该历史结束时的真实状态。可以观察到，对于结束于相似真实状态（智能体无法获知）的历史，所提出的 De-Rex方法仍能够为这些历史提供相似的表示。

奖励： 在图10 中，我们绘制了使用 De-Rex程序获得的奖励。可以观察到，当部分可观测性并非极端严重时，即使真实状态未知，所提出的方法仍能够为底层真实状态提供伪计数。这可以归因于 De-Rex 提供了一种联合机制，既用于学习能够捕获底层充分统计量的表示，也用于利用这些表示估计伪计数。

信念分布： 虽然图 9 提供了一种评估学习到的分布质量的方法，但它可用于可视化高维空间中的表示。因此，在图11、12、13和14 中，我们提供了另一种评估学习到的表示质量的方法。

对于这些图，我们学习了一个分类模型（使用单层神经网络），旨在利用给定（部分）历史的表示，对其观测序列结束时的真实潜在状态进行分类。这为我们提供了一个代理指标，用于衡量学习到的表示在解码关于潜在真实状态的信念方面的效果。

正如预期，当 $p = 0.0$ 不存在部分可观测性时，智能体会学会利用（忽略）历史来获取表征，从而能够准确解码关于智能体真实位置的信念。随着部分可观测性程度增加，可以观察到，由于对智能体位置的确定性降低，信念分布变得更加分散。尽管如此，De-Rex 仍能提供可用于解码邻近潜在状态的表征。

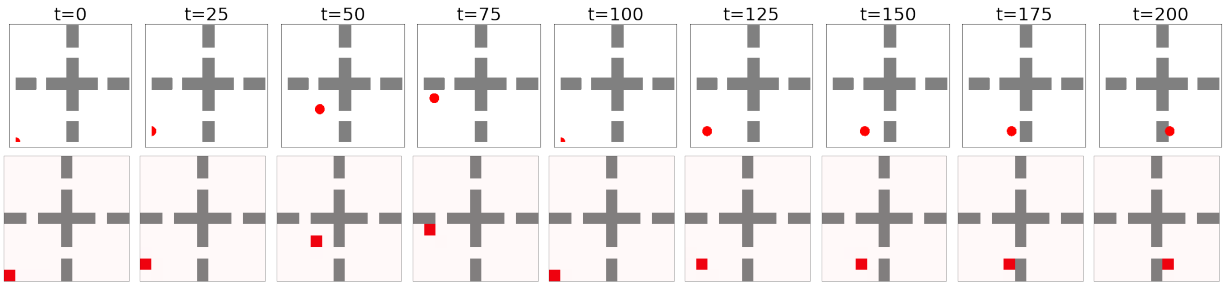

图*11.* 在部分可观测马尔可夫决策过程设置下，给定试验中的信念（带有 $p = 0.0$）。

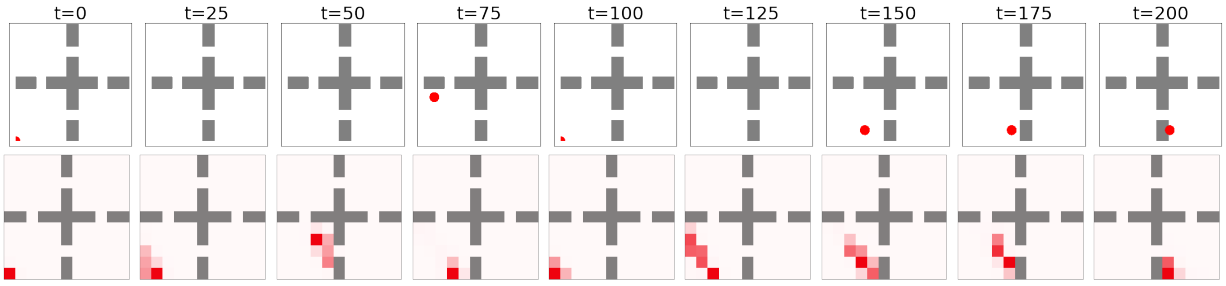

图*12.* 在部分可观测马尔可夫决策过程设置下，给定试验中的信念（带有 $p = 0.3$）。

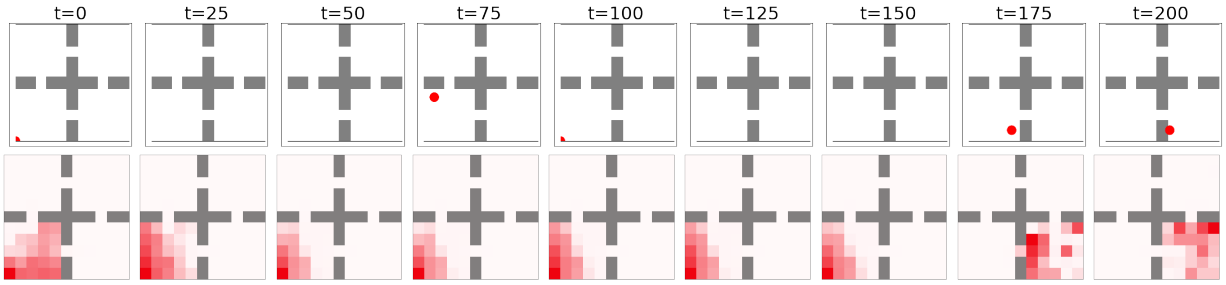

图*13.* 在部分可观测马尔可夫决策过程设置下，给定试验中的信念（带有 $p = 0.6$）。

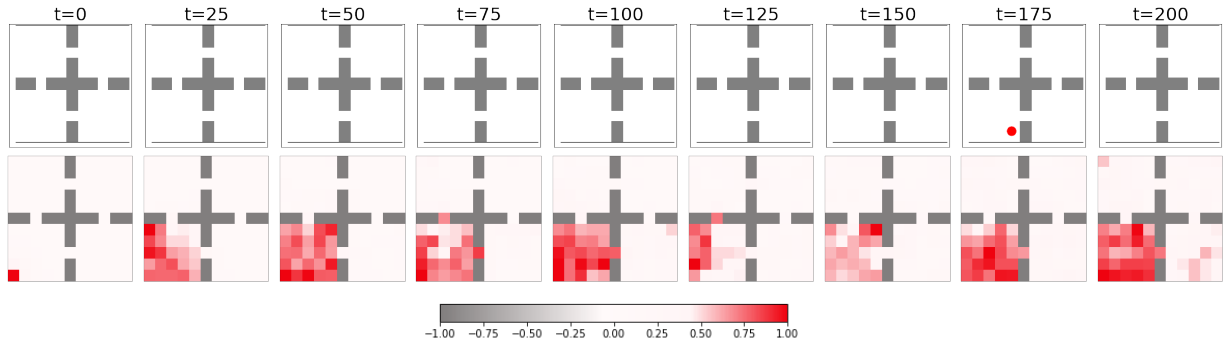

图*14.* 在部分可观测马尔可夫决策过程设置下，给定试验中的信念（带有 $p = 0.9$）。顶行展示了在时间步 $t = \{0, 25, 50, 75, 100, 125, 150, 175, 200\}$提供给智能体的观测快照。底行表示根据截至时间步 t的历史表示解码出的信念。更多详情请参见正文。 **图例：** 本页所有图表共用同一图例。对于底行，红色越深表示信念程度越高。对于顶行，红色表示智能体的位置。 p越高，表示部分可观测性越强，因此智能体缺失的帧也越多。

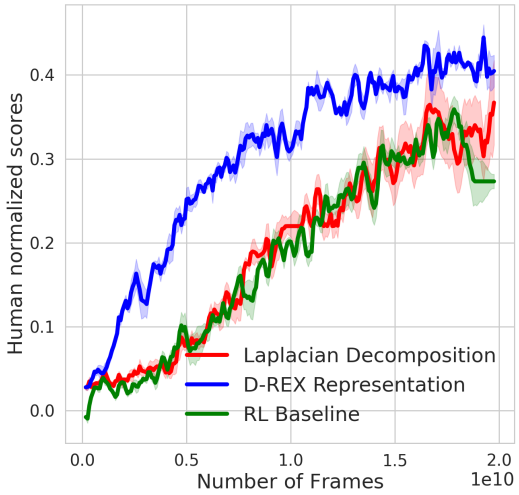

图*15*。人类归一化 在全部 30 项任务上汇总的标准化得分 DMLab30 的全部30项任务上汇总的得分（使用3个随机种子）

E. 更多大规模结果

E.1. DMLab30

我们使用人类归一化性能（$z_i - u_i$）/（$h_i - u_i$）和 $1 \leq i \leq 30$来衡量 DMLab30环境中每项任务的性能，其中 u_i 和 h_i 分别是随机策略和人类的原始得分，z_i 是智能体获得的得分。归一化得分为 1 表示智能体在该任务上的表现与人类相当。

在图 15 中，我们绘制了所有领域的汇总人类归一化得分；在图 18中，我们分别绘制了各项任务的学习曲线。从这两者中都可以观察到，在处理复杂环境时，基于分解的基线方法无法带来显著提升。相比之下，所提出的 De-Rex方法有望取得更好的表现。

E.2. DM-Hard-8

为了更好地了解仅学习表征对困难探索任务能有多大帮助，我们针对 SVD表示和拉普拉斯分解方法，在 DM-Hard-8 任务上进行了消融研究。对于这两种方法，我们绘制了不同 (a) w_{loss} 值和 (b) 执行分解的 f_{θ_1} 输出维度下的结果。

从图中的结果可以看出，16 和 17，仅使用表示学习时，D-Rex 最多只能解决 2 个任务，而拉普拉斯分解方法只能解决 1 个任务。这说明构建良好的伪计数以辅助探索至关重要，这一点在 De-Rex 探索方法的图 7 中也显而易见。

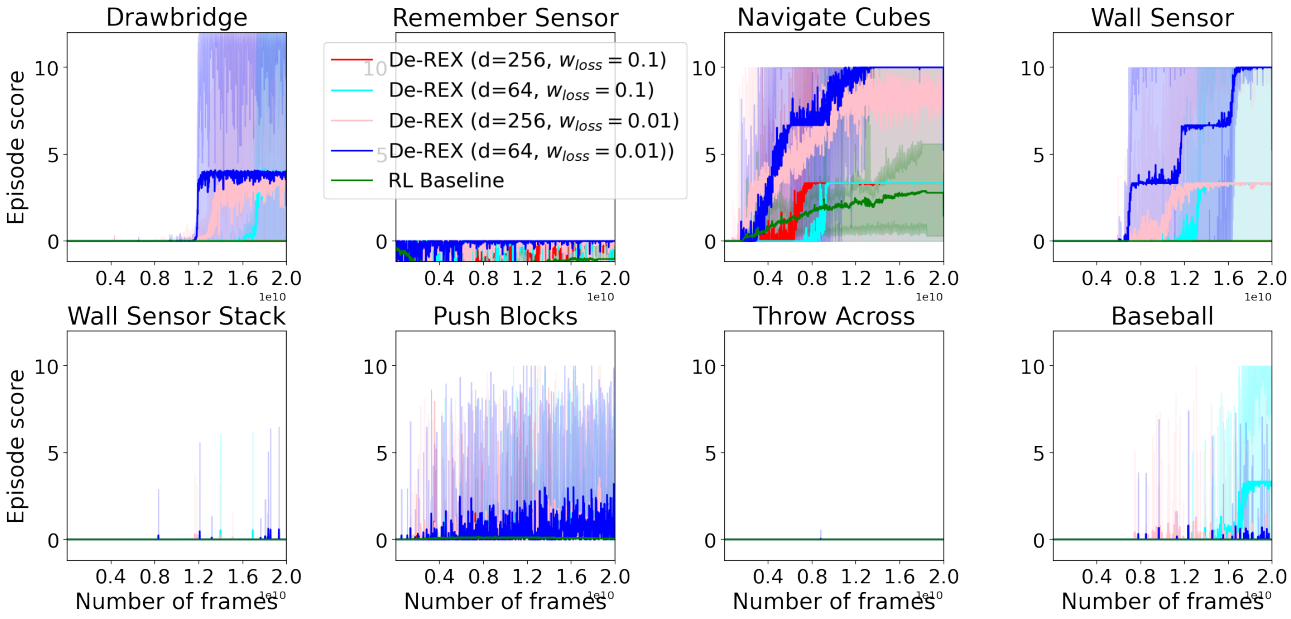

图*16*。参数研究：De-Rex 表示。表示中的奇异值分解损失权重（w_{loss}）和奇异值分解的维度（d）的敏感性

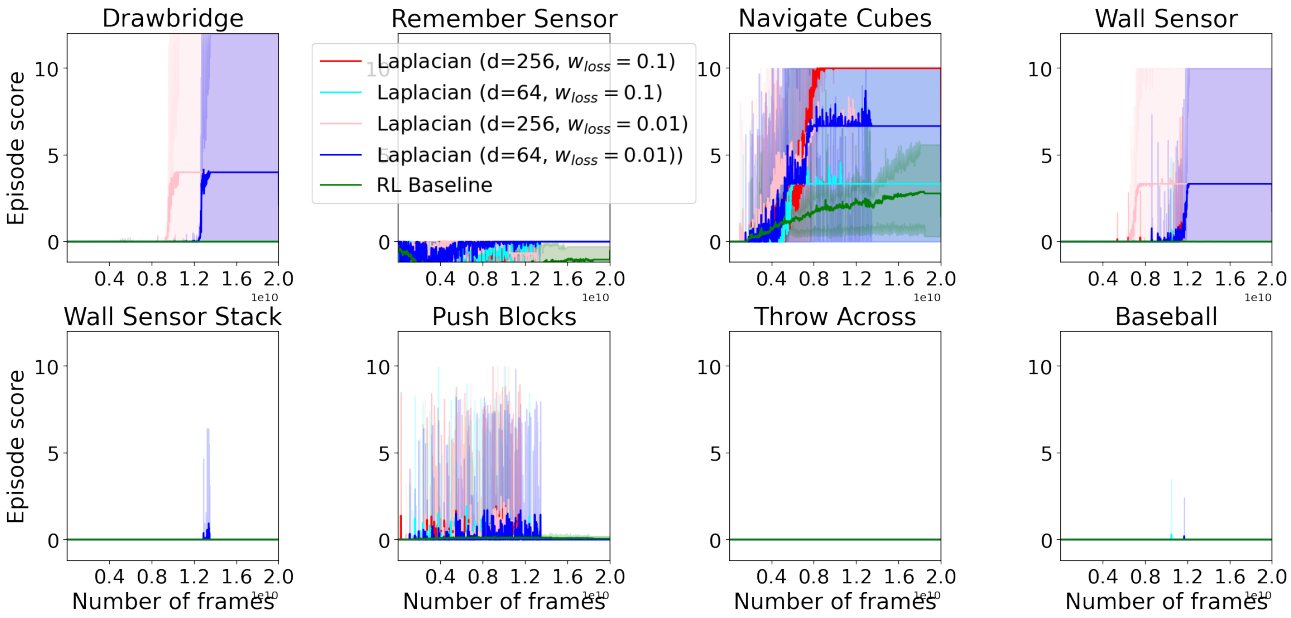

图*17*。 参数研究：拉普拉斯分解表示。表示中分解损失权重（w_{loss}）和奇异值分解维度（d）的敏感性

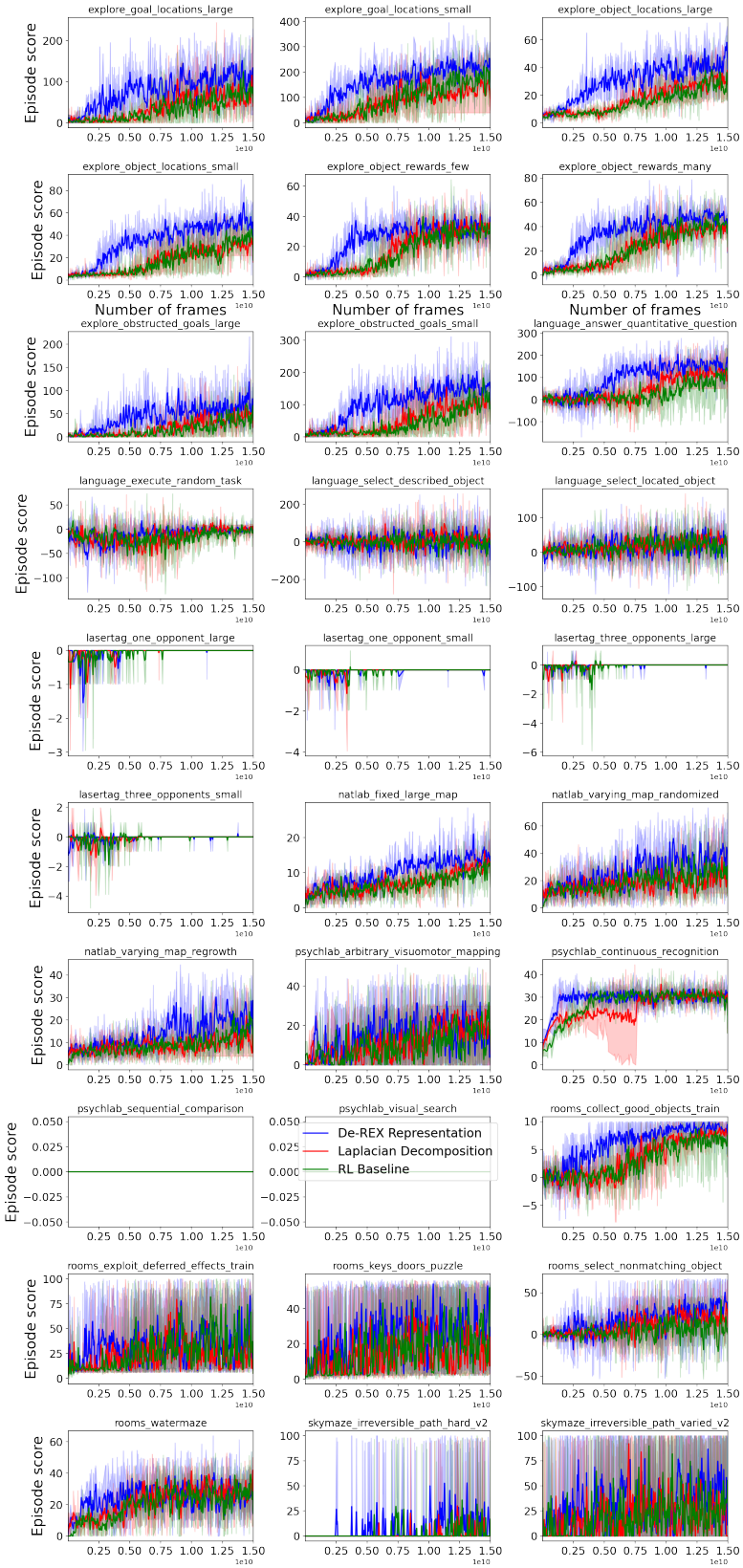

图*18*。DMLab30 任务中 De-Rex 表示、拉普拉斯分解和 VMPO 基线方法的单独学习曲线。