# DeepMDP: Learning Continuous Latent Space Models for Representation Learning

Carles Gelada [1] Saurabh Kumar [1] Jacob Buckman [2] Ofir Nachum [1] Marc G. Bellemare [1]

# 【强化学习 68】DeepMDP

张楚珩
清华大学 交叉信息院博士在读



DeepMDP可以看做是对于原来MDP的一个抽象。

## 原文传送门

Gelada, Carles, et al. "DeepMDP: Learning Continuous Latent Space Models for Representation Learning." International Conference on Machine Learning. 2019.

## 特色

前面讲了很多state abstraction的东西，看到了state abstraction有很多很好的性质，但是一直都没有讲如何去得到这样一个state abstraction $\phi: \mathcal{S} \to \phi(\mathcal{S})$ ，实际上这也是一个比较困难的问题。这篇文章讲了一个practical的学习得到state abstraction的方法，并且理论上说明了它和bisumulation的联系（参见本专栏【强化学习理论 63】StatisticalRL 7）。

另外，之前去俞扬老师组转了转，他们做的MindGame[1]相当于是人工做了一个这样的state abstraction，感觉也很有意思，不过还没来得及看。这里的DeepMDP相当于是自动学习state abstraction，不过实验效果上来说目前还是MindGame的做法见效更快。

## 过程

### 1. 目标

本文目标是希望把当前比较复杂的状态空间投影到一个比较低维度的连续状态空间上，即 $\phi: \mathcal{S} \to \bar{\mathcal{S}}$ ，其中 $\bar{\mathcal{S}} \subset \mathbb{R}^D$ 。对于一个MDP $M = \{\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma\}$ ，定义 $\bar{M} = \{\bar{\mathcal{S}}, \mathcal{A}, \bar{\mathcal{R}}, \bar{\mathcal{P}}, \gamma\}$ 为与 $\phi$ 相关的一个MDP。可以把 $(\bar{M}, \phi)$ 称作 $M$ 的隐空间模型（latent space model）。如果用一个参数化的模型（比如神经网络）来拟合这个隐空间模型，可以使用三个神经网络分别来代表 $\phi, \bar{\mathcal{R}}, \bar{\mathcal{P}}$ 。文章称这样的 $(\bar{M}, \phi)$ 为DeepMDP。

本文的目标是设计参数化的模型来表示DeepMDP，同时设计相应的损失函数；同时本文分析了在DeepMDP相比于原MDP有多大损失（注意到DeepMDP通常维度更低，学习起来更容易，但是一般带来一些approximation error）。

### 2. 做法

DeepMDP的做法很直接，就是在采集到的样本上来最小化以下两项损失函数

$$L_{\mathcal{R}}^{\xi} = \mathop{\mathbb{E}}_{s,a\sim\xi} |\mathcal{R}(s,a) - \bar{\mathcal{R}}(\phi(s),a)|, \tag{7}$$

$$L_{\mathcal{P}}^{\xi} = \mathop{\mathbb{E}}_{s,a\sim\xi} \left[ W\left(\phi\mathcal{P}(\cdot|s,a), \bar{\mathcal{P}}(\cdot|\phi(s),a)\right)\right]. \tag{8}$$

其中，可以把 $\xi$ 看做是采集到样本的分布，W表示Wasserstein距离。其训练的参数为 $\phi,\bar{\mathcal{R}},\bar{\mathcal{P}}$ 三个神经网络的网络参数。

当然实际中Wasserstein距离很难计算，因此本文就使用了确定性的 $\mathcal{P}$ 模型，这样Wasserstein距离就退化为了L2 loss。同时，后面的理论部分要求 $\bar{\mathcal{R}},\bar{\mathcal{P}}$ 需要满足Lipschitz条件，因此使用了gradient penalty来限制训练得到的神经网络满足该条件。

### 3. 定义

下面希望进一步说明最小化以上两项损失函数得到的DeepMDP性质较好，即DeepMDP上的价值函数、最优策略，相比于原MDP损失不是太大。直观来说，上面两项损失函数其实是要求了DeepMDP和原MDP在reward和dynamics上都相差不大，但即使这样，如果原MDP或者最优策略很不规则也可能导致DeepMDP和原MDP差距较大，因此我们需要要求MDP和策略都比较“平滑”。对于一个状态空间上的 metric $d_{\mathcal{S}}$ ，定义 “平滑”的 MDP 和“平滑”的策略。

**Definition 2.** *Let $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma \rangle$ be an MDP with a continuous, metric state space $(\mathcal{S}, d_{\mathcal{S}})$, where $d_{\mathcal{S}} : \mathcal{S}\times\mathcal{S} \to \mathbb{R}^+$, and a discrete action space $\mathcal{A}$. We say $\mathcal{M}$ is $(K_R, K_P)$-Lipschitz if, for all $s_1, s_2 \in \mathcal{S}$ and $a \in \mathcal{A}$:*

$$|\mathcal{R}(s_1,a) - \mathcal{R}(s_2,a)| \leqslant K_{\mathcal{R}} d_{\mathcal{S}}(s_1,s_2)$$
$$W\left(\mathcal{P}(\cdot|,s_1,a), \mathcal{P}(\cdot|s_2,a)\right) \leqslant K_{\mathcal{P}} d_{\mathcal{S}}(s_1,s_2)$$



**Definition 3.** *A policy $\pi \in \Pi$ is $K_V$-Lipschitz-valued if for all $s_1, s_2 \in \mathcal{S}$ and $a, a \in \mathcal{A}$:*

$$|V^{\pi}(s_1) - V^{\pi}(s_2)| \leqslant K_V d_{\mathcal{S}}(s_1,s_2)$$
$$|Q^{\pi}(s_1,a) - Q^{\pi}(s_2,a)| \leqslant K_V d_{\mathcal{S}}(s_1,s_2)$$



在 Lipschitz MDP 上的策略也是 Lipschitz 的，即以上两个定义之间存在关系

**Lemma 1.** *Let $\mathcal{M}$ be $(K_R, K_P)$-Lipschitz and let $\pi$ be any policy with the property that $\forall s_1, s_2 \in \mathcal{S}$,*

$$|V^\pi(s_1) - V^\pi(s_2)| \leqslant \max_{a \in \mathcal{A}} |Q^\pi(s_1, a) - Q^\pi(s_2, a)|$$

*then $\pi$ is $\frac{K_\mathcal{R}}{1-\gamma K_\mathcal{P}}$-Lipschitz-valued.*



（我没太想明白，能找到一个反例使得这个 property 不成立？）

由此可以推论，MDP 上的最优策略也是 Lipschitz 的。

**Corollary 1.** *Let $\mathcal{M}$ be $(K_R, K_P)$-Lipschitz, then $\pi^*$ is $\frac{K_\mathcal{R}}{1-\gamma K_\mathcal{P}}$-Lipschitz-Valued.*

### 4. Global DeepMDP bound

这里讲的 Global 的意思是， DeepMDP 和原 MDP 之间在所有的 state-action space 上 reward 和 dynamics 都相差不多时，考虑从原 MDP 到 DeepMDP 的差距。

即给定

$$L^\infty_\mathcal{R} = \sup_{s,a} L_{\bar{\mathcal{R}}}(s, a) \tag{5}$$

$$L^\infty_\mathcal{P} = \sup_{s,a} L_{\bar{\mathcal{P}}}(s, a) \tag{6}$$

***DeepMDP 和原 MDP 价值函数的差距不太大***

**Lemma 2.** *Let $\mathcal{M}$ and $\bar{\mathcal{M}}$ be an MDP and DeepMDP respectively, with an embedding function $\phi$ and global loss functions $L^\infty_\mathcal{R}$ and $L^\infty_\mathcal{P}$. For any $K_{\bar{V}}$-Lipschitz-valued policy $\bar{\pi} \in \bar{\Pi}$ the value difference can be bounded by*

$$\left|Q^{\bar{\pi}}(s, a) - \bar{Q}^{\bar{\pi}}(\phi(s), a)\right| \leqslant \frac{L^\infty_\mathcal{R} + \gamma K_{\bar{V}} L^\infty_\mathcal{P}}{1 - \gamma},$$



其中 $\bar{\Pi}$ 表示 DeepMDP 中的所有策略，它是原 MDP 中所有策略的子集。

***满足约束的状态表示（abstraction/representation）下距离相近的状态价值函数也相近***

这里想研究把满足(5)和(6)约束的 representation $\phi$ 做为一个 state abstraction 的质量如何。有一种比较显然的 representation 失败的情况：两个状态价值函数很不一样的状态，在representation $\phi$ 的作用下，映射到同一个状态。这里想说明的是，这样的情况不会发生，即representation $\phi$ 作用下距离较近的状态，其价值函数相差也不大。

**Theorem 1.** *Let $\mathcal{M}$ and $\overline{\mathcal{M}}$ be an MDP and DeepMDP respectively, with an embedding function $\phi$ and global loss functions $L^{\infty}_{\bar{\mathcal{R}}}$ and $L^{\infty}_{\bar{\mathcal{P}}}$. For any $K_{\bar{V}}$-Lipschitz-valued policy $\bar{\pi} \in \bar{\Pi}$ the representation $\phi$ guarantees that for any $s_1, s_2 \in \mathcal{S}$ and $a \in \mathcal{A}$,*

$$\left|Q^{\bar{\pi}}(s_1,a) - Q^{\bar{\pi}}(s_2,a)\right| \leqslant K_{\bar{V}} \left\|\phi(s_1) - \phi(s_2)\right\|_2 + 2\frac{\left(L^{\infty}_{\bar{\mathcal{R}}} + \gamma K_{\bar{V}} L^{\infty}_{\bar{\mathcal{P}}}\right)}{1-\gamma}$$



***DeepMDP 下的最优策略在原 MDP 中也不会太差***

**Theorem 2.** *Let $\mathcal{M}$ and $\overline{\mathcal{M}}$ be an MDP and a $(K_R, K_P)$-Lipschitz DeepMDP respectively, with an embedding function $\phi$ and global loss functions $L^{\infty}_{\bar{\mathcal{R}}}$ and $L^{\infty}_{\bar{\mathcal{P}}}$. For all $s \in \mathcal{S}$, the suboptimality of the optimal policy $\bar{\pi}^*$ of $\overline{\mathcal{M}}$ evaluated on $\mathcal{M}$ can be bounded by,*

$$V^*(s) - V^{\bar{\pi}^*}(s) \leqslant 2\frac{L^{\infty}_{\bar{\mathcal{R}}} + \gamma K_{\bar{V}} L^{\infty}_{\bar{\mathcal{P}}}}{1-\gamma}$$



*Where $K_{\bar{V}} = \frac{K_{\bar{R}}}{1-\gamma K_{\bar{P}}}$ is an upper bound to the Lipschitz constant of the value function $\bar{V}^{\bar{\pi}^*}$, as shown by Corollary 1.*

## 5. Local DeepMDP bounds

前面讲的 global bound 的要求是对于所有的 state-action space，DeepMDP 和原 MDP 差距都严格地不超过规定的数值。在实际中都是在样本上最小化误差的，没法保证全局的 bound，因此这里考

虑一个更为实际的情况，即在样本上 DeepMDP 和原 MDP 差距不太大时，相应的 bound。

这里要求在样本上 DeepMDP 和原 MDP 的 reward 和 dynamics 相差不大。

$$L_{\mathcal{R}}^{\xi} = \mathop{\mathbb{E}}_{s,a\sim\xi} |\mathcal{R}(s,a) - \bar{\mathcal{R}}(\phi(s),a)|, \tag{7}$$

$$L_{\mathcal{P}}^{\xi} = \mathop{\mathbb{E}}_{s,a\sim\xi} \left[ W\left(\phi\mathcal{P}(\cdot|s,a), \bar{\mathcal{P}}(\cdot|\phi(s),a)\right)\right]. \tag{8}$$

***DeepMDP 和原 MDP 价值函数的差距不太大***

**Lemma 3.** *Let $\mathcal{M}$ and $\bar{\mathcal{M}}$ be an MDP and DeepMDP respectively, with an embedding function $\phi$. For any $K_{\bar{V}}$-Lipschitz-valued policy $\bar{\pi} \in \bar{\Pi}$, the expected value function difference can be bounded using the local loss functions $L_{\mathcal{R}}^{\xi_{\bar{\pi}}}$ and $L_{\mathcal{P}}^{\xi_{\bar{\pi}}}$ measured under $\xi_{\bar{\pi}}$, the stationary state action distribution of $\bar{\pi}$.*

$$\mathop{\mathbb{E}}_{s,a\sim\xi_{\pi}} \left|Q^{\bar{\pi}}(s,a) - \bar{Q}^{\bar{\pi}}(\phi(s),a)\right| \leqslant \frac{\left(L_{\mathcal{R}}^{\xi_{\pi}} + \gamma K_{\bar{V}} L_{\mathcal{P}}^{\xi_{\pi}}\right)}{1-\gamma}$$



***满足约束的状态表示（abstraction/representation）下距离相近的状态价值函数也相近***

**Theorem 3.** *Let $\mathcal{M}$ and $\bar{\mathcal{M}}$ be an MDP and DeepMDP respectively, with an embedding function $\phi$. Let $\bar{\pi} \in \bar{\Pi}$ be any $K_{\bar{V}}$-Lipschitz-valued policy with stationary distribution $d_{\bar{\pi}}(s)$ and let $L_{\mathcal{R}}^{\xi_{\bar{\pi}}}$ and $L_{\mathcal{P}}^{\xi_{\bar{\pi}}}$ be the local loss functions measured under $\xi_{\bar{\pi}}$, the stationary state action distribution of $\bar{\pi}$. For any two states $s_1, s_2 \in \mathcal{S}$, the local representation similarity can be bounded by*

$$|V^{\bar{\pi}}(s_1) - V^{\bar{\pi}}(s_2)| \leqslant K_{\bar{V}} \|\phi(s_1) - \phi(s_2)\|_2 + \frac{L_{\mathcal{R}}^{\xi_{\pi}} + \gamma K_{\bar{V}} L_{\mathcal{P}}^{\xi_{\pi}}}{1-\gamma} \left(\frac{1}{d_{\bar{\pi}}(s_1)} + \frac{1}{d_{\bar{\pi}}(s_2)}\right)$$

这说明，对于策略访问较频繁的状态，如果其表示比较相近，那么其价值函数也差不太多；但是对于样本较少的区域，其表示就不是特别准确了。

### 6. 与 bisimulation 的联系

Bisimulation 对应的映射 $\phi$ 可以最大程度地压缩状态空间，可以证明在压缩之后的 MDP 上找到的最优策略性能和原 MDP 上最优策略性能差不多。这样，策略空间实际上可以被大大缩小到 $\Pi_\phi$（注意这上面是一弯，不是一横），从而提高了找到一个好的策略的效率；具体说来，这里的策略空间就不允许在被 $\phi$ 映射到同一个状态下的两个状态下采取不同的行动；如果是近似的 bisimulation，即不允许在被 $\phi$ 映射到相似状态下的两个状态下采取非常不同的行动。这里想要证明的是对于任何 $\Pi_\phi$ 中的策略，都可以找到 DeepMDP 中的一个策略，使得这两个策略相差不多。

下面先定义 bisimulation

**Definition 4** (Givan et al. (2003)). *Given an MDP $\mathcal{M}$, an equivalence relation $B$ between states is a bisimulation relation if for all states $s_1, s_2 \in \mathcal{S}$ equivalent under $B$ (i.e. $s_1 B s_2$), the following conditions hold,*

$$R(s_1, a) = R(s_2, a)$$
$$\mathcal{P}(G|s_1, a) = \mathcal{P}(G|s_2, a), \forall G \in \mathcal{S}/B$$

*Where $\mathcal{S}/B$ denotes the partition of $\mathcal{S}$ under the relation $B$, the set of all sets of equivalent states, and where $\mathcal{P}(G|s, a) = \sum_{s' \in G} \mathcal{P}(s'|s, a)$.*



bisimulation 只给出了两个状态等价不等价的定义，为了便于分析，我们还需要定义两个状态相似的程度。

**Definition 5.** *Given an MDP $\mathcal{M}$, a bisimulation metric $\tilde{d}$ satisfies the fixed point:*

$$\tilde{d}(s_1, s_2) = \max_a (1 - \gamma) |\mathcal{R}(s_1, a) - \mathcal{R}(s_2, a)| + \gamma W_{\tilde{d}}(\mathcal{P}(\cdot|s_1, a), \mathcal{P}(\cdot|s_2, a))$$

bisimulation 下比较容易学到的策略集合定义如下

**Definition 6.** *We denote by* $\widetilde{\Pi}_K$ *the set of* $K$***-Lipschitz-bisimilar policies****, s.t. for all* $s_1, s_2 \in \mathcal{S}, a \in \mathcal{A}$*,*

$$\{\pi : \pi \in \Pi, |\pi(a|s_1) - \pi(a|s_2)| \leqslant K\widetilde{d}(s_1, s_2)\}.$$

一个理想的 state abstraction 产生的效果应该是对应的策略集合包括下面的策略集合（最优策略在此集合中），但是大小尽可能小。下面这个定理说明了这一点。

**Theorem 4.** *Let* $\mathcal{M}$ *be an MDP and* $\overline{\mathcal{M}}$ *be a* $(K_{\overline{\mathcal{R}}}, K_{\overline{\mathcal{P}}})$*-Lipschitz DeepMDP, with an embedding function* $\phi$*, and global loss functions* $L^{\infty}_{\overline{\mathcal{R}}}$ *and* $L^{\infty}_{\overline{\mathcal{P}}}$*. Denote by* $\overline{\Pi}_K$ *the set of* $K$***-Lipschitz deep policies*** $\{\bar{\pi} : \bar{\pi} \in \overline{\Pi}, |\bar{\pi}(a|s_1) - \bar{\pi}(a|s_2)| \leqslant K \left\|\phi(s_1) - \phi(s_2)\right\|_2, \forall s_1, s_2 \in \mathcal{S}, a \in \mathcal{A}\}$*. Finally define the constant* $C = \frac{(1-\gamma)K_{\overline{\mathcal{R}}}}{1-\gamma K_{\overline{\mathcal{P}}}}$*. Then for any* $\widetilde{\pi} \in \widetilde{\Pi}_K$ *there exists a* $\bar{\pi} \in \overline{\Pi}_{CK}$ *which is close to* $\widetilde{\pi}$ *in the sense that, for all* $s \in \mathcal{S}$ *and* $a \in \mathcal{A}$*,*

$$|\widetilde{\pi}(a|s) - \bar{\pi}(a|s)| \leqslant L^{\infty}_{\overline{\mathcal{R}}} + \gamma L^{\infty}_{\overline{\mathcal{P}}} \frac{K_{\overline{\mathcal{R}}}}{1 - \gamma K_{\overline{\mathcal{P}}}}$$



## 7. 实验

文章实验效果还不错，在一个 demo 的环境中说明了该方法确实能够找到一个有意义的 state abstraction；用这种方法做辅助能够在 Atari 游戏上有性能提升。不过文章还是提到如果直接训练 DeepMDP 的三个神经网络，可能导致 $\phi$ 网络什么都学不到，因为 $\bar{\mathcal{P}}$ 网络为了预测的更准，会促使 $\phi$ 网络生成没有信息量的状态抽象，即整体的训练过程有一定的冲突。

---

参考文献

[1] Liu, Ruo-Ze, et al. "Efficient Reinforcement Learning with a Mind-Game for Full-Length StarCraft II."*arXiv preprint arXiv:1903.00715*(2019).

发布于 2019-06-08

强化学习 (Reinforcement Learning)



**文章被以下专栏收录**

**强化学习前沿**
读呀读paper