Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?*

Simon S. Du
Institute for Advanced Study
ssdu@ias.edu

Sham M. Kakade
University of Washington, Seattle
sham@cs.washington.edu

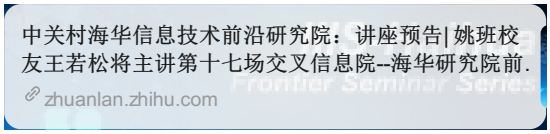
Ruosong Wang
Carnegie Mellon University
ruosongw@andrew.cmu.edu

Lin F. Yang
University of California, Los Angles
linyang@ee.ucla.edu

【强化学习 101】Representation Lower Bound

张楚珩
清华大学 交叉信息院博士在读

今天本文的作者之一，姚班毕业的同学，Ruosong Wang 来讲他们的一篇理论工作，还挺有意思的。



原文传送门

Du, Simon S., et al. "Is a Good Representation Sufficient for Sample Efficient Reinforcement Learning?." arXiv preprint arXiv:1910.03016 (2019).

特色

理论方面的工作，大家都基于很多种不同的假设设计了不同的 provably efficient 算法，给出了相应的 upper bound，这相当于找到了 efficient algorithm 的许多充分条件。但是不同论文里面又都用到了许多很不一样的假设，它们之间的关系又都很难说得清楚。这篇论文反过来思考哪些条件是 efficient algorithm 的必要条件，即如果缺少哪些条件，就不可能设计出 efficient algorithm。

讲座结束之后跟王若松同学聊了一下，整理梳理了一下最近读到的一些强化学习理论方向的相关工作；另外我发现这篇 paper 的 related work 写的也很棒，因此特别记录下来。特别注意到有哪些已经 solved，有哪些仍然是 open problem， 不同的流派和做法可以按照一些什么样的维度去区分他们的工作。

强化学习理论进展

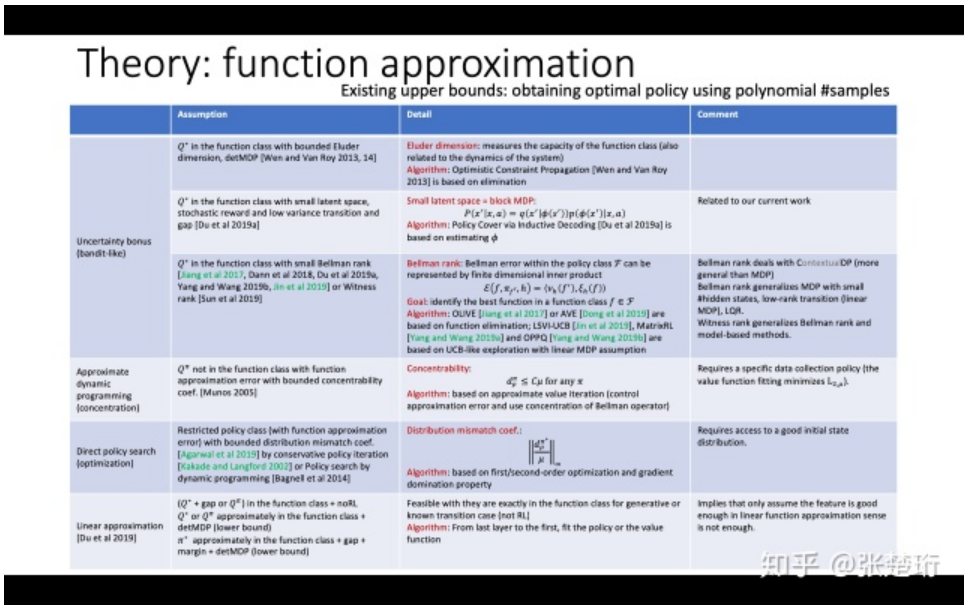
首先，强化学习理论问题可以被划分为两大块：

第一块是 tabular case，即状态空间是离散的情形。在这种情形下，相关的 RL 问题都基本上被解

决了，理论分析允许 stochastic transition、stochastic reward、unknown dynamics、arbitrary initial state distribution 等。已知的 upper bound 和 lower bound 只相差了一个 H（planning horizon），大致上 sample complexity 基本为 $\Theta(|S||A|\text{poly}(H)/\epsilon^2)$ ，具体的可以参见姜楠老师的 paper [1]。

第二块是目前大家比较关心的，approximated case，即当状态空间非常大或者连续的时候，我们就不能允许在 sample complexity 里面出现 $|S|$ 项了，因此需要考虑做 function approximation。这一块主要有三个流派：

- Uncertainty-bonus-based algorithm：主要是 bandit problem 的研究思路，对 dynamics 和 function class 做一些假设，然后按照 UCB 或者 Elimination 的思路设计算法。
- Approximate dynamic programming-based algorithm：主要基于 Bellman operator 的 contraction 性质，然后仔细处理 approximation error，对 transition 有一些限制。
- Direct policy search-based algorithm：需要假设有比较均匀的初始状态分布，这样就能利用 gradient domination 和优化领域的一些结论来证明 sample complexity。



绿色标注了我比较熟悉的 paper。最近读 paper 都没有写知乎，这里统一标注一下，欢迎大家一起交流讨论！

这里的总结说明：目前大家需要对 transition 做各种 low-rank 的假设（第一大块），或者在状态分布上做一些超出 RL 范围的假设（第二、三大块）；如果仅仅只假设有一个好的特征，但是对 dynamics 不做假设（第四大块，本工作），是不可能存在有效的 RL 算法的。

这篇文章的贡献和之前工作的关系如下

Query Oracle	RL	Generative Model	Known Transition
Previous Upper Bounds			
Exact Linear Q^* + DetMDP [Wen and Van Roy, 2013]	✓	✓	✓
Exact Linear Q^* + Bellman-Rank [Jiang et al., 2017]	✓	✓	✓
Exact Linear Q^* + Low Var + Gap [Du et al., 2019a]	✓	✓	✓
Exact Linear Q^* + Gap (Open Problem / Theorem B.1)	?	✓	✓
Exact Linear Q^π for all π (Open Problem / Theorem C.1)	?	✓	✓
Approx. Linear Q^π for all π + Bounded Conc. Coeff. [Munos, 2005, Antos et al., 2008]	✗	✓	✓
Approx. Linear Q^π for all π + Bounded Dist. Mismatch Coeff. [Agarwal et al., 2019]	✗	✓	✓
Lower Bounds (this work)			
Approx. Linear Q^*+ DetMDP (Theorem 4.1)	✗	✗	✗
Approx. Linear Q^π for all π + DetMDP(Theorem 4.1)	✗	✗	✗
Exact Linear π^* + Margin + Gap + DetMDP (Theorem 4.2)	✗	✗	✗
Exact Linear Q^* (Open Problem)	?	?	?

注意到 RL（指定的初始状态分布，只能通过动作/策略来转移到各个不同的状态） > generative model（可以设置为任意的状态） > known transition（整个转移概率函数都完全知道）。因此只要证明前面的某个设定下的 upper bound，即在后面设定有相应的 upper bound；只要证明后面某个设定下的 lower bound，即在前面的设定下有相应的 lower bound。

过程

1. 定义

这篇文章假设如果有某种比较好的特征，使得 function approximation 能以最简单的线性模型表示的话，那么会有怎样的 upper bound 和 lower bound。那么一个好的特征具体是什么呢？对于 value-based 方法来说，一个好的特征能够使得最优价值函数（或者任意价值函数）都能够被线性表示出来，即

Assumption 4.1 (Q^* Realizability). *There exists $\delta > 0$ and $\theta_0, \theta_1, \ldots, \theta_{H-1} \in \mathbb{R}^d$ such that for any $h \in [H]$ and any $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, $|Q_h^*(s,a) - \langle \theta_h, \phi(s,a) \rangle| \leq \delta$.*

Assumption 4.2 (Value Completeness). *There exists $\delta > 0$, such that for any $h \in [H]$ and any policy π, there exists $\theta_h^\pi \in \mathbb{R}^d$ such that for any $(s,a) \in \mathcal{S}_h \times \mathcal{A}$, $|Q_h^\pi(s,a) - \langle \theta_h, \phi(s,a) \rangle| \leq \delta$.*

对于 policy-based 方法来说，一个好的特征是能够表示出最优策略，即

Assumption 4.3 (π^* Realizability). *For any $h \in [H]$, there exists $\theta_h \in \mathbb{R}^d$ that satisfies for any $s \in \mathcal{S}_h$, we have $\pi^*(s) \in \arg\max_a \langle \theta_h, \phi(s,a) \rangle$.*

这样的假设显然是比 Q* realizability 弱的。如果把前面的假设看做是 regression 存在合适的 regression 的话，这里相当于是假设在一个 classification 问题中存在一个分界面。考虑到有监督学习的一些通用假设，通常还会假设该分界面存在一个 margin，即

Assumption 4.4 (π^* Realizability + Margin). *We assume $\phi(s,a) \in \mathbb{R}^d$ satisfies $\|\phi(s,a)\|_2 = 1$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$. For any $h \in [H]$, there exists $\theta_h \in \mathbb{R}^d$ with $\|\theta_h\|_2 = 1$ and $\triangle > 0$ such that for any $s \in \mathcal{S}_h$, there is a* unique *optimal action $\pi^*(s)$, and for any $a \neq \pi^*(s)$, $\langle \theta_h, \phi(s, \pi^*(s)) \rangle - \langle \theta_h, \phi(s,a) \rangle \geq \triangle$.*

如果有两个 action 太过相似，那么 optimal policy 和 suboptimal policy 就变得难以分辨，所以一般还会假设存在一个 gap。

we first define the function gap : $\mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as $\text{gap}(s,a) = \max_{a' \in \mathcal{A}} Q^*(s,a') - Q^*(s,a)$.

Assumption 3.1 (Optimality Gap). *There exists $\rho > 0$ such that $\rho \leq \text{gap}(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ with $\text{gap}(s,a) > 0$.*

2. 算法思路（Upper bound）

下面我们证明表中的『exact Linear Q* + gap + generative model』 存在有效的算法。

Theorem B.1. *Under Assumption 3.1 and Assumption 4.2 with $\delta = 0$, in the Generative Model query model, there is an algorithm that finds π^* with* poly $\left(d, H, \frac{1}{\rho}\right)$ *trajectories with probability* 0.99.

gap　　delta

注意到 d 是 feature 的维度

当 $\delta = 0$ 并且存在一个 gap， 考虑到 optimal policy 能用 d 维特征的线性组合表示出来，那么用 $\text{poly}(d, \frac{1}{\rho})$ 的样本就能以 $1 - \frac{0.01}{H}$ 的概率分辨出某一层 h 上的 optimal policy；从 H-1 到 1 来跑这个算法；在第 h 层的时候，它后面 Q 的估计可以用已经学到的 h+1 到 H-1 层的 optimal policy 来 rollout。

Theorem C.1. *Under Assumption 4.2 with $\delta = 0$, in the Generative Model query model, there is an algorithm that finds an ϵ-optimal policy $\hat{\pi}$ using* poly $\left(d, H, \frac{1}{\epsilon}\right)$ *trajectories with probability* 0.99.

思路差不多。

3. Lower bound

下面我们来说明 Approximate Linear Q* + DetMDP 时为什么不能有一个多项式复杂度的算法？（sample complexity 会和 planning horizon H 成指数关系）

一个例子

先看一个直观的难的例子，说明了即使 feature 造的足够好，以至于线性函数拟合就能够表示真实的价值函数或者策略（value completeness assumption），也还是会有关于 planning horizon H 呈指数的困难。

- Dynamics：考虑有两个 action 的 DetMDP，如果选第一个 action 就走到左子节点，如果选择第二个就走到右子节点。
- Reward：在最后一层的某一个状态上 reward=1，其他状态 reward=0。

- 初步分析：直观来说，由于必须至少把最后一层所有状态都遍历一遍，才能够知道是那个状态上有 reward，因此至少需要 $\Omega(2^H)$ 的样本才够。这是由于，虽然假设了 DetMDP，其中 reward 虽然是确定性的，但是是未知的。

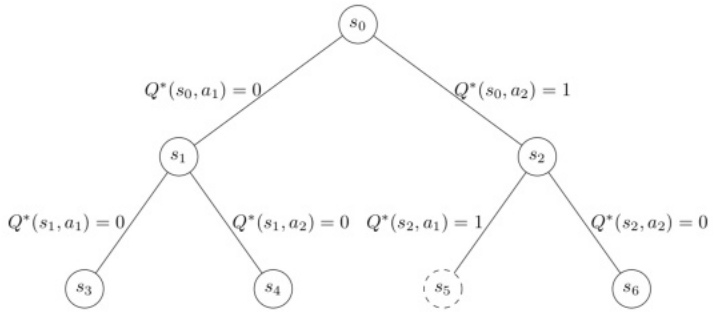


Figure 2: An example with $H = 3$. For this example, we have $\overline{r}(s_5) = 1$ and $\overline{r}(s) = 0$ for all other states s. The unique state s_5 which satisfies $\overline{r}(s) = 1$ is marked as dash in the figure. The induced Q^* function is marked on the edges.

注意到，tabular case 下认为状态数目是固定的，这里认为状态会很多（比如这里就有 $2^{H+1}-1$ 个状态），但是可以用有限维的 feature 来表示。

先看一个简单的设定：假设如果只有一维 feature，并且这个 feature 是 binary 的，那么假设到达一个状态，其 feature=1，如果它上面 reward=0，那么就可以排除其他所有 feature=1 的状态。这个想法比较粗糙，但是说明一个问题，如果 representation 维度比较低， 就可以通过 representation 的相似程度来基于已知的样本来做泛化，这使得我们不需要把每个状态都访问到就可以探索到 reward function 的形态。

有了 value completeness assumption 之后真的可以减少我们对于每个状态的访问次数么？最关键的来了，答案是否！我们可以找到一个低维的特征表示 $d = \Omega(H/\delta^2)$ 使得其中能包含 2^H 个近似的标准正交基！注意到，一个近似的标准正交基一定能对于任意的 reward function 都满足 value completeness assumption；同时，如果为标准正交基，那么探索到一个标准正交向量对应的状态对于标准正交基中其他的向量并不会带来任何的信息量。因此，我们即使有了这样的一个满足 value completeness assumption 的表示，也仍然需要 2^H 次查询，才能够找到有 reward=1 的那个状态。

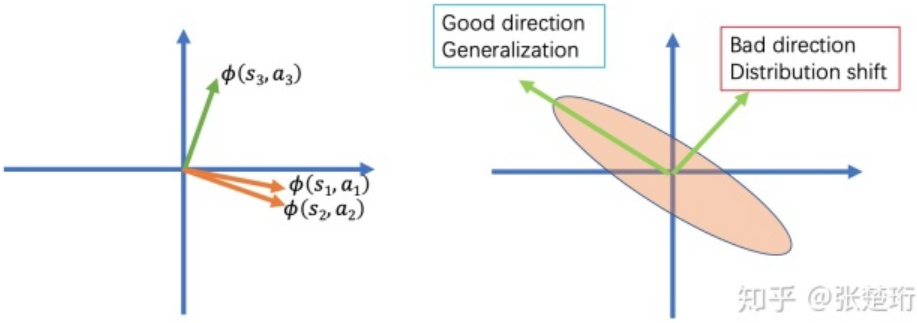
分析

下面从另外一个角度来分析。

前面讲了当 $\delta = 0$ 时的一个算法设计思路，最重要的是它利用了 $\delta = 0$ 这样一个已知条件。这种情况下，考虑一个 d 维的特征，只要我们知道 d 个线性无关的特征和对应的 Q 值，我们就能知道，它们拟合得到的 $\hat{Q}$ 一定等于真实的 Q^* 。

但是这样情况并不适用于 $\delta > 0$ 的情形，考虑下图中左边的例子。考虑 d=2，在 $\delta = 0$ 和 DetMDP 的情况，如果知道了橙色的两个向量的特征和 Q 值，那么我们就能拟合出线性函数的系数，从而非常安全地去预测任意一个特征（比如绿色的特征）对应的 Q 值。但当 $\delta > 0$ 时，我们只知道这样最优的一个拟合对于橙色两个特征的拟合精度在 δ 范围之内，这样系数 w 的取值范围就在一个狭长

的区域内了（大家在纸上笔画一下就知道了）。这种情况下，预测一个和橙色向量垂直的向量时，就会产生任意大的误差。

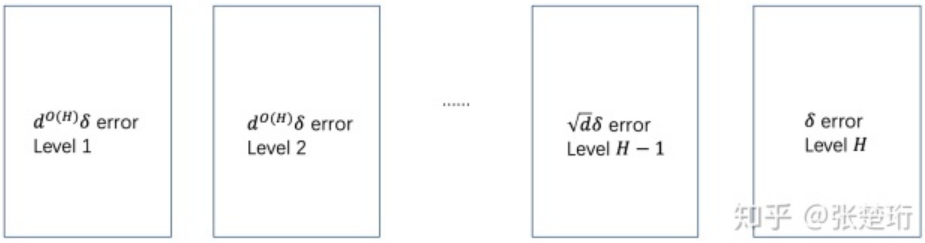


考虑一个更为一般的情形，看上图右边的例子。在 $\delta=0$ 时，采集到差不多 d 个点就能够非常安全地在整个空间上泛化了；但是在 $\delta>0$ 情况下，如果数据采样的分布在上图橙色椭球范围内，我们只能在椭球的主轴方向有比较好的泛化，但是在其垂直的方向上，泛化误差就会很大。

考虑有 H 层的情形。在第 h 层的时候，各个 state-action pair 对应的 Q 函数是通过已经学习好的 $h+1:H$ 的 Q* 函数得到的，考虑此时在 $h+1:H$ 上已经有了 δ 的误差。考虑到上面所讲的原因，对于这一层上 (s',a') 的估计误差 δ' 可以写作

$$\delta' \leq \sqrt{\#\text{data points}} \times \sqrt{\phi(s',a')^T C^{-1} \phi(s',a')} \times \delta$$

其中 C 表示样本点的 covariance matrix。为了控制红色的这一项比较小（红色的部分小于 1），我们需要采集的样本数目差不多为 $O(d)$ 。这样每一层的误差都会被放大 $\sqrt{d}$ ，这样误差就会呈指数增长。



4. 总结

- 在一些 model-based 的假设下（比如前面蓝色表中列的一些工作），已有一些算法能够有效对抗拟合误差。
- 困难并不来自于有监督学习的过程，比如我们可以假设 gap、margin 等，但是仍然不解决问题。
- 困难也并不来自于环境的 dynamics 未知。如果一个环境的 dynamics 难（比如前面提到的那个二叉树例子），即使把它告诉你，但是不告诉你 reward function，你也需要指数级的样本去探索 reward function。
- 最大的困难来自于 distribution mismatch。如果把 RL 看做是 SL（有监督学习），那么智能体一开始并不知道要在哪个分布上做优化。这一点在 Agarwal et al 2019 上也可以看得到，只要做了一个关于 distribution 的简单假设，很多问题都能迎刃而解。

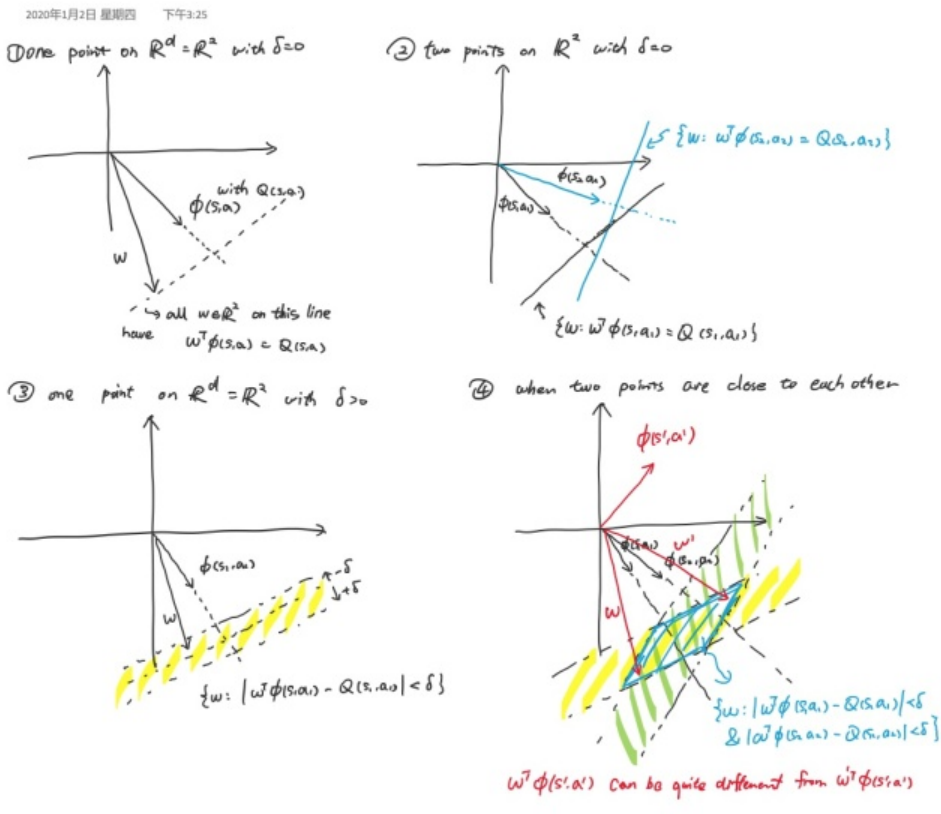
参考文献

[1] Jiang, Nan, and Alekh Agarwal. "Open problem: The dependence of sample complexity lower bounds on planning horizon." Conference On Learning Theory. 2018.

Acknowledgement

感谢王若松同学的 PPT 和 talk！

关于 @李英儒 的疑问



编辑于 2020-01-04

清华大学 计算机科学

文章被以下专栏收录

强化学习前沿
读呀读paper

进入专栏