arXiv:2302.00727v1 [cs.LG] 1 Feb 2023

基于核的 Q 学习样本复杂度

Sing-Yuan Yeh[1,2] Fu-Chieh Chang[1,3] Chang-Wei Yueh[4] Pei-Yuan Wu[3,4] Alberto Bernacchia[1] Sattar Vakili[1]

[1]联发科技研究院

[2]国立台湾大学与中央研究院数据科学研究生学位项目[3]国立台湾大学通信工程研究所[4]国立台湾大学电机工程学系

摘要

现代强化学习（RL）通常面临庞大的状态-动作空间。现有的分析结果通常针对状态-动作数量较少的设定，或针对线性建模的 Q 函数等简单模型。为了推导出能够处理大规模状态-动作空间及更一般化 Q 函数的高效强化学习策略，近期一些研究采用核岭回归进行非线性函数逼近。在本研究中，当存在生成模型时，我们推导了基于核的 Q 学习的样本复杂度。我们提出了一种非参数 Q 学习算法，能够在任意大规模的折扣马尔可夫决策过程（MDP）中找到一个ϵ-最优策略。相对于ϵ 以及核的复杂度（以其信息增益衡量），所提出算法的样本复杂度具有阶最优性。据我们所知，这是首个在如此一般的模型下证明有限样本复杂度的结果。

1 引言

近年来，强化学习（RL）已成功应用于多个领域，包括游戏（Silver 等人，2016；Lee 等人，2018年；Vinyals et al.，2019）、自动驾驶（Kahn 等人，2017年）、微芯片设计（Mirhoseini 等人，2021）、机器人控制（Kalashnikovet al.，2018年）和算法搜索（Fawzi 等人，2022年）。现实世界的问题通常包含规模庞大的状态-动作空间，甚至可能是无限的。例如，围棋 有10^{170} 个状态（Silver 等人，2016），而动作的数量矩阵乘法算法空间中的算法数量为 10^{30}（Fawzi 等人，2022年）。目前，人们尚未完全理解强化学习算法如何学习成功的策略来解决这些问题。要取得这样的成功，似乎需要使用现代函数逼近器，例如基于核的学习和深度神经网络。

一个重要的理论问题如下。考虑一个转移概率分布未知的马尔可夫决策过程（MDP）。假设有一个生成模型（Kakade，2003年）可用，它能够从任意状态-动作对提供样本转移。要学习一个足够好的策略，需要多少样本？这被称为样本复杂度。

以表格型和线性马尔可夫决策过程等某些简单设定为条件，已有研究推导出了样本复杂度的理论界。在表格型设定下，研究表明，学习一个 ϵ-最优策略（即其价值函数与最优价值函数的差距至多为 ϵ）的样本复杂度为 $\mathcal{O}(\frac{|\mathcal{S}||\mathcal{A}|}{\epsilon^2})$，其中 $|\mathcal{S}|$ 和 $|\mathcal{A}|$ 分别是状态集合和动作集合的基数（Kearns 和 Singh，1998年；Azar 等人，2013年；Sidford 等人，2018a,b；Agarwal 等人，2020年）。这意味着，对于非常大的状态-动作空间，要获得良好性能，可能需要几乎无限多的样本。另一类研究考虑线性马尔可夫决策过程模型，其中转移概率可以表示为 d维状态-动作特征映射的线性表示。研究表明，在这种情况下，样本复杂度为 $\mathcal{O}(\frac{d}{\epsilon^2})$，即与状态空间和动作空间的规模无关（Yang 和 Wang，2019）。遗憾的是，线性假设相当不灵活，在实践中也并不常见。

为了弥补表格型和线性马尔可夫决策过程假设所带来的局限性（例如状态-动作空间较小或模型过于简单），近期一些研究采用核岭回归，在可能无限的状态-动作域上进行非线性函数逼近。在这些工作中，转移概率分布（有时还包括奖

第 26th届人工智能与统计学国际会议（AISTATS）2023年，西班牙瓦伦西亚。版权归作者所有，2023年。

励函数）通过基于核的模型灵活表示（Yang 和 Wang，2020年；Yang 等人，2020a，b）。基于核的模型可提供强大的回归器和不确定性估计，从而为强化学习算法提供指导。此外，基于核的模型具有很强的表示能力，能够对范围广泛的问题建模，因为紧子集上的所有连续函数 $\mathbb{R}^d$ 都可以用常见的核进行逼近（Srinivas 等人，2010年）。然而，现有研究并未解决本文所关注的特定样本复杂度问题，而是在分回合马尔可夫决策过程的设定下推导遗憾界。更详细的比较见第1.2节。

基于核的方法可以理解为一种线性模型，其状态-动作特征映射具有无限维度，例如，该映射对应于核的 Mercer 特征函数（参见第 2.2 节）。从这个意义上说，线性模型是采用线性核的基于核模型的一个特例。尽管如此，关于线性马尔可夫决策过程样本复杂度的结果并不能推广到基于核的模型，因为这些样本复杂度随特征映射的维度而变化（在核设定中，该维度可能是无限的）。相比之下，在核设定中，样本复杂度取决于由核的复杂度决定的某些核特定性质，后文将对此进行更详细的讨论。

1.1 贡献

针对折扣马尔可夫决策过程及其样本复杂度问题（类似于 Azar 等人，2013年；Sidford 等人，2018a，b；Yang 和 Wang，2019），我们对现有工作进行了扩展和推广，具体如下。

- 我们介绍基于核的 Q 学习（称为 KQLearn），这是一种样本采集算法，能够在非常广泛的模型类别上，以有限的样本复杂度返回一个 ϵ-最优策略。与表格型和线性马尔可夫决策过程设定相比，KQLearn 至少做出了两项创新性贡献。在表格型设定中，样本从所有状态-动作对中采集，这导致样本复杂度呈现 $|\mathcal{S}||\mathcal{A}|$ 的缩放关系。在**性设定中，样本从一组覆盖整个状态-动作空间的状态-动作中采集（使样本复杂度随维度 d 缩放）。随后，通过值迭代更新线性模型参数的估计值。在我们的情形中，状态-动作空间是无限的，且采用非参数基于核的模型，因此上述两种方法均不可行。相反，KQLearn 利用核模型提供的不确定性来创建一个有限的状态-动作集合，并用该集合采集样本。随后，这些样本通过核岭回归传入近似贝尔曼算子，以更新价值函数（即定义在整个状态-动作空间上的连续函数）。
- 我们推导了适用于广泛核模型的 KQLearn 有限样本复杂度。具体而言，我们考虑两类 Mercer 特征值呈指数衰减（$\sigma_m \sim \exp(-m^{\beta_e})$，$\beta_e > 0$）和多项式衰减（$\sigma_m \sim m^{-\beta_p}$，$\beta_p > 1$）的核$\sigma_m$（见定义 2）。我们分别证明了在这两种设定下，样本复杂度为 $\tilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ 和 $\tilde{\mathcal{O}}\left((\frac{1}{\epsilon})^{\frac{2\beta_p}{\beta_p-1}}\right)$[1]。据我们所知，这是在特征值呈多项式衰减的设定下，适用于所有 $\beta_p > 1$ 的首个有限样本复杂度，也是首个阶数最优的样本复杂度。我们在第 1.2 节中更详细地讨论了与相关工作的比较。作为一个特例，我们重新得到了 Yang 和 Wang（2019）报告的线性设定下的 $\tilde{\mathcal{O}}(\frac{d}{\epsilon^2})$ 样本复杂度。

我们承认，KQLearn 的样本复杂度上界在对折扣因子的依赖方面可能并非阶最优 γ。特别是，对于光滑核，我们的上界随$\frac{1}{(1-\gamma)^7}$ 增长，这与 Yang 和 Wang (2019) 为线性设定提出的 PPQ-Learning 算法类似。然而，在表格型和线性设定下，这种依赖关系分别在 $\frac{1}{(1-\gamma)^3}$，中改进为 Sidford 等人 (2018a) 和 Yang 和 Wang (2019) 得到改进。这里能否实现同样的改进，似乎是一个颇具挑战性的问题。如上所述，即使要证明有限的样本复杂度也很有挑战性；在困难核（例如第 1.2 节讨论的某些多项式核）下，现有工作的样本复杂度可能会发散。

1.2 相关工作

使用生成模型研究折扣马尔可夫决策过程中的样本复杂度这一具体问题，已有表格型和线性设定下的研究。相关结果汇总于表1。该问题的其他变体则考虑不存在生成模型的马尔可夫决策过程（例如，代表性研究包括 Azar 等人，2017年；Jin等人，2018年、2020年；Russo，2019；Yang 等人，2020a、b；Kakade 等人，2020年；Zhou 等人，2021；Domingues 等人，2021，以及其中的参考文献），这些研究通常采用分回合设定，每回合长度为 T，遗憾界依赖于 T 和 H。随后可将遗憾界转换为样本复杂度（例如，参见 Jin等人，2018年；Yang 等人，2020b）。这些结果也列于表1。强化学习中非线性函数近似的其他方法还包括具有有界 eluder 维度的模型（Wang 等人，2020年；Ayoub 等人，2020年）以及平滑

[1]符号 $\mathcal{O}$ 和 $\tilde{\mathcal{O}}$ 分别用于表示数学阶，以及忽略对数因子后的数学阶。

表1：第 1.2 节讨论了各种设定下已有的样本复杂度。

算法	MDP	设定	样本复杂度
(带置信上界的Q学习，Jin等人，2018年)	分回合	表格型	$\tilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert H^4}{\epsilon^2}\right)$
(LSVI-UCB,Jin等人，2020年)	分回合	线性	$\tilde{\mathcal{O}}\left(\frac{d^3H^4}{\epsilon^2}\right)$
(KOVI,Yang 等人，2020b)	分回合	基于核的，多项式特征值衰减	$\tilde{\mathcal{O}}\left(H^4(\frac{1}{\epsilon})^{\frac{2\beta_p}{\beta_p-2}}\right)$
		基于核的，指数特征值衰减	$\tilde{\mathcal{O}}\left(\frac{H^4}{\lvert\mathcal{S}\rvert^2\lvert\mathcal{A}\rvert}\right)$
(Sidford 等人，2018a，方差缩减 QVI)	折扣型	表格型	$\tilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{(1-\gamma)^3\epsilon^2}\right)$
(PPQ-Learning,Yang 和 Wang，2019)	折扣型	线性	$\tilde{\mathcal{O}}\left(\frac{d}{(1-\gamma)^3\epsilon^2}\right)$
(OPPQ-Learning,Yang 和 Wang, 2019)	折扣型	线性	$\tilde{\mathcal{O}}\left(\frac{d}{(1-\gamma)^3\epsilon^2}\right)$
KQLearn	折扣型	基于核的，多项式特征值衰减	$\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^{\frac{2\beta_p}{\beta_p-1}}(1-\gamma)^{\frac{2\beta_p-1}{\beta_p-1}}}\right)$
		基于核的，指数特征值衰减	$\tilde{\mathcal{O}}\left(\frac{1}{\epsilon^2(1-\gamma)^7}\right)$

核（Domingues 等人，2021）。在这些研究中，与我们最相关的两项是 Yang 和 Wang (2019) 和 Yang 等人 (2020a, b)。

与Yang 和 Wang (2019)类似，我们也研究了使用生成模型的折扣马尔可夫决策过程中的样本复杂度。我们考虑的是非参数基于核的模型，而他们考虑的是参数线性模型。因此，他们的算法和结果都无法推广到我们的设定。线性设定是核设定的一种特殊情况，即使用线性核时，我们可以恢复 $\tilde{\mathcal{O}}(\frac{d}{\epsilon^2})$ 样本复杂度，见 Yang 和 Wang (2019)，适用于两种算法：PPQ-Learning 和 OPPQ-Learning。后者改善了关于折扣因子的样本复杂度。

与 Yang 等人 (2020b类似，我们也考虑一种基于核的模型。我们研究带有生成模型的折扣马尔可夫决策过程中的样本复杂度，而他们主要研究分回合马尔可夫决策过程中的遗憾界。他们还将样本复杂度作为遗憾界的直接推论予以报告。具体而言，在特征值呈指数衰减和多项式衰减这两种设定下，他们的样本复杂度分别可表示为 $\tilde{\mathcal{O}}(\frac{1}{\epsilon^2})$ 和$\tilde{\mathcal{O}}\left((\frac{1}{\epsilon})^{\frac{2\beta_p}{\beta_p-2}}\right)$。在多项式设定下，他们的样本复杂度界大于我们的界。此外，他们的样本复杂度不总是有限的；当 $1 < \beta_p \leq 2$ 时，它可能发散，而这涵盖了许多值得关注的情形。这种次优性以及可能出现的平凡结果，是当 $1 < \beta_p \leq 2$ 时超线性（因而平凡）的遗憾界所导致的。有关这一结果相关理论挑战的详细讨论，请参见 Vakili 等人 (2021d。

例如，考虑 Mat´ern 核族，它是最常用的核族之一（Snoek 等人，2012；Shahriari 等人，2015），也是理论上备受关注的核族（Srinivas 等人，2010年）。对于光滑度参数为´的 Mat ν ern 核，在 d 维输入域上，$\beta_p = 1 + \frac{2\nu}{d}$（Yang 等人，2020b）。这意味着Yang 等人（2020b）中的样本复杂度在 $d > 2\nu$ 时发散（使用 Mat´ern 核时通常会出现这种情况）。不过，我们强调，带生成模型的折扣马尔可夫决策过程与分回合马尔可夫决策过程属于不同的设定，不能直接比较。尽管如此，我们还是在一个非常一般的设定下，给出了首个始终有限的样本复杂度结果，该设定涵盖了所有具有多项式衰减特征值的核。

另一个相关问题是基于核的老虎机问题（Srinivas 等人，2010年），它对应于一个退化的马尔可夫决策过程，其中 $|\mathcal{S}| = 1$。基于核的老虎机问题是一个研究充分的问题，已有阶最优遗憾界（Salgia 等人，2021；Li 和 Scarlett，2022年）和样本复杂度（Vakili 等人，2021a）。平方指数核（SE）和 Mat´ern 核的样本复杂度下界见（Scarlett 等人，2017年），其与我们的结果具有相同的 ϵ 尺度（至多相差对数因子），这表明我们的样本复杂度在 ϵ 下是阶最优的（见第 4节）。

论文结构：第2节在概述马尔可夫决策过程和核岭回归的背景后，对问题进行了形式化定义。KQLearn 的介绍见第3节。第4节讨论研究结果。第5节提供了高层次分析，详细内容则留待附录讨论。

2 预备知识

本节概述马尔可夫决策过程和核岭回归的背景，然后正式阐述在该设定下 Q 学习的样本复杂度问题。

2.1 折扣马尔可夫决策过程

折扣马尔可夫决策过程（MDP）可由元组 $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ 描述，其中 $\mathcal{S}$ 是状态空间，$\mathcal{A}$ 是动作空间，$\gamma \in (0, 1)$ 是折扣因子，$r: \mathcal{S} \times \mathcal{A} \to [0, 1]$ 是奖励函数，$P(\cdot|s, a)$ 是从状态-动作对 (s, a) 出发到下一状态的转移概率分布[2]，定义在$\mathcal{S}$ 上。我们用记号 $\mathcal{Z} = \mathcal{S} \times \mathcal{A}$ 表示状态 - 动作空间。我们的结果通常适用于（可能非常大但）有限的 $\mathcal{Z}$ 或某些无限的 $\mathcal{Z}$。为确保正确性，我们假设 $\mathcal{Z}$ 是 $\mathbb{R}^d$ 的紧子集。

目标是找到一个（可能是随机的）策略 $\pi: \mathcal{S} \to \mathcal{A}$，使长期期望奖励最大化，即最大化价值函数，

$$V^\pi(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r\left(s_t, \pi\left(s_t\right)\right) \mid s_0 = s\right],$$

其中 $s_t \sim P(\cdot|s_{t-1}, \pi(s_{t-1}))$ 构成了状态的轨迹。可以证明（例如，参见 Puterman，2014），在温和的假设下（例如，P的连续性、$\mathcal{Z}$的紧致性，以及 r的有界性），存在一个最优策略π^*，它在每个状态下都能达到可能的最大价值 V^*。

$$\forall s \in \mathcal{S}: \quad V^*(s) := V^{\pi^*}(s) = \max_\pi V^\pi(s).$$

为简化记号，对于价值函数 $V: \mathcal{S} \to \mathbb{R}$，令

$$[PV](s, a) := \mathbb{E}_{s' \sim P(\cdot|s,a)}[V(s')].$$

策略 π 的 Q 函数（有时也称为状态-动作价值函数）以及最优 Q 函数定义如下

$$Q^\pi(s, a) = r(s, a) + \gamma[PV^\pi](s, a), \text{ and}$$
$$Q^*(s, a) = Q^{\pi^*}(s, a),$$

分别地，贝尔曼算子 $\mathcal{T}: \mathbb{R}^{\mathcal{S}} \to \mathbb{R}^{\mathcal{S}}$ 定义为

$$\forall s \in \mathcal{S}: \quad [\mathcal{T}V](s) = \max_{a \in \mathcal{A}} \{r(s, a) + \gamma[PV](s, a)\}.$$

ϵ-最优策略的样本复杂度：ϵ-最优策略定义如下。

[2]我们有意不对 P 使用标准术语“转移核”，以避免与基于核的学习中的“核”一词混淆。

定义 1.（ϵ-最优策略）如果策略 π 从任何初始状态出发都能达到如下所述的近似最优值，则称其为 ϵ-最优：

$$V^\pi(s) \geq V^*(s) - \epsilon, \quad \forall s \in \mathcal{S},$$

或者等价地 $\|V^\pi - V^*\|_\infty \leq \epsilon$。

我们的目标是利用少量样本学习 ϵ-最优策略。在本文中，我们遵循 Kearns 和 Singh（1998年）；Azar 等人（2013年）；Sidford 等人（2018a,b）；Yang 和 Wang（2019）的做法，假设给定一个生成模型（Kakade，2003年），强化学习算法可以针对任意状态-动作对（s, a）$\in \mathcal{Z}$查询转移样本 $s' \sim P(\cdot|s, a)$。强化学习算法的样本复杂度定义为该算法为获得 ϵ-最优策略所使用的此类样本数量。

2.2 再生核希尔伯特空间与核岭回归

在强化学习场景中实现有限样本复杂度的现有研究，通常假设状态-动作空间较小，或马尔可夫决策过程具有线性模型。利用基于核的学习，可以对这些结果进行推广和扩展。具体而言，一种自然的方法是使用已知再生核希尔伯特空间（RKHS）中的元素来对转移建模。本节概述再生核希尔伯特空间和核岭回归。

设 $K: \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ 是相对于有限博雷尔测度的一个已知正定核。设 $\mathcal{H}_K$ 是由 K 诱导的再生核希尔伯特空间，其中 $\mathcal{H}_K$ 包含一族定义在 $\mathcal{Z}$ 上的函数。令 $\langle \cdot, \cdot \rangle_{\mathcal{H}_K}: \mathcal{H}_K \times \mathcal{H}_K \to \mathbb{R}$ 以及 $\|\cdot\|_{\mathcal{H}_K}: \mathcal{H}_K \to \mathbb{R}$ 分别表示 $\mathcal{H}_K$ 的内积和范数。再生性质意味着，对于所有 $f \in \mathcal{H}_K$，以及 $z \in \mathcal{Z}$、$\langle f, K(\cdot, z) \rangle_{\mathcal{H}_K} = f(z)$。不失一般性地，我们假设对于所有 z，$K(z, z) \leq 1$。Mercer 定理表明，在某些温和条件下，K 可以使用无限维特征映射表示为：

$$K(z, z') = \sum_{m=1}^{\infty} \sigma_m \psi_m(z) \psi_m(z'). \tag{1}$$

正式表述及详细内容见附录 A。

核岭回归：基于核的模型可提供强大的回归器和不确定性估计器（粗略来说，即替代后验方差），可用于指导强化学习算法。具体而言，考虑一个未知函数 $f \in \mathcal{H}_K$。考虑一组$\mathcal{U}_J = \{z_j\}_{j=1}^J \subset \mathcal{Z}$ 个 J 输入。假设 J 带噪观测 $\{Y(z_j) = f(z_j) + \epsilon_j\}_{j=1}^J$ 已知，其中ϵ_j 为独立同分布的零均值次高斯噪声项。核岭回归分别给出以下回归器和不确定性估计（例如，参见 Sch¨olkopf 等人，

2002年），

$$\hat{f}_{\mathcal{U}_J}(z) = k_{\mathcal{U}_J}^\top(z)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1} Y_{\mathcal{U}_J},$$
$$\Sigma^2_{\mathcal{U}_J}(z) = K(z,z) - k_{\mathcal{U}_J}^\top(z)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1} k_{\mathcal{U}_J}(z), \tag{2}$$

其中 $k_{\mathcal{U}_J}(z) = [K(z,z_1)\ldots,K(z,z_J)]^\top$ 是一个向量，其元素为 z 与各观测之间的核值；$K_{\mathcal{U}_J} = [K(z_i,z_j)]_{i,j=1}^J$ 是核矩阵，$Y_{\mathcal{U}_J} = [Y(z_1)\ldots,Y(z_J)]^\top$ 是观测向量，I_J 是维度为 J 的单位矩阵，$\lambda > 0$ 是一个自由正则化参数。

置信区间：核岭回归提供的预测和不确定性，使我们能够在算法和分析中使用标准置信区间。特别是，有各种结果表明，以至少 $1-\delta$ 的概率，预测函数满足 $|f(z) - f_{\mathcal{U}_J}(z)| \le \beta(\delta)\Sigma_{\mathcal{U}_J}(z)$（针对固定的 z，或同时针对所有 z），其中置信区间宽度乘数 $\beta(\delta)$ 取决于观测噪声的性质，以及 f 的复杂度（以其再生核希尔伯特空间范数衡量）（Srinivas 等人，2010年；Abbasi-Yadkori，2013年；Vakili 等人，2021a，2022年）。如果定义域 $\mathcal{Z}$ 是有限的，则对固定 z 的置信区间应用并集界，即可直接得到一致置信界。对于连续域，通常会采用离散化论证，并考虑以下连续性假设。

假设 1. 对于每个 $n \in \mathbb{N}$，存在 $\mathcal{Z}$ 的一个离散化集合 $\mathbb{Z}$，使得对于任意满足 $\|f\|_{\mathcal{H}_K} \le C_K$ 的 $f \in \mathcal{H}_K$，都有 $f(z) - f([z]) \le \frac{1}{n}$ $[z] = \arg\min_{z' \in \mathbb{Z}} \|z' - z\|_{l^2}$ $|\mathbb{Z}| \le cC_K^d n^d$，其中 min 表示距离 $f \in \mathcal{H}_K$ 最近的点，且 c 是一个与 n 和 C_K 无关的常数。

假设1 是一个技术性且较弱的假设，对于平方指数核和 Matérn 等典型核均成立，其中 $\nu > 1$ (Srinivas 等人，2010年；Chowdhury和 Gopalan，2017年；Vakili 等人，2021a)。

在我们的分析中，我们使用以下再生核希尔伯特空间元素的置信区间。

引理 1 (Vakili 等人 (2021a，2022年))。考虑一组固定设计的观测点，其中 $\mathcal{U}_J$ 与观测噪声相互独立。当噪声项为参数为 R [3] 且 $\|f\|_{\mathcal{H}_K} \le C_K$ 的次高斯分布时，以下各项均对 $z \in \mathcal{Z}$ 一致成立，其概率为 $1-\delta$，

$$f(z) \le \hat{f}_{\mathcal{U}_J}(z) + \beta(\delta)\Sigma_{\mathcal{U}_J}(z),$$
$$f(z) \ge \hat{f}_{\mathcal{U}_J}(z) - \beta(\delta)\Sigma_{\mathcal{U}_J}(z), \tag{3}$$

where $\beta(\delta) = \mathcal{O}\left(C_K + \frac{R}{\lambda}\sqrt{d\log(\frac{JC_K}{\delta})}\right)$.

[3]若随机变量 X 的矩生成函数有一个均值为零、方差为 R^2的高斯分布的矩生成函数作为上界，则称其为参数为 R 的次高斯随机变量。

最大信息增益： 为了便于分析，我们定义最大信息增益 $\Gamma_{K,\lambda}$，它是核特定的复杂度项。借助与椭圆势引理（Carpentier 等人，2020年）类似的结果，它使我们能够界定核模型中的总体不确定性。具体而言，定义

$$\Gamma_{K,\lambda}(J) = \sup_{\mathcal{U}\subset\mathcal{Z},|\mathcal{U}|\le J} \frac{1}{2}\log\det\left(I_J + \frac{1}{\lambda^2}K_{\mathcal{U}}\right). \tag{4}$$

接下来，我们有如下结论。

引理 2 (Srinivas 等人 (2010年))。对于任意集合 $\mathcal{U}_J \subset \mathcal{Z}$，我们有

$$\sum_{j=1}^J \Sigma^2_{\mathcal{U}_{j-1}}(z_j) \le \frac{2}{\log(1+1/\lambda^2)}\Gamma_{K,\lambda}(J). \tag{5}$$

2.3 问题表述

考虑第 2.1 节所述的折扣型马尔可夫决策过程。我们希望设计一种样本复杂度较低的算法，以获得一个 ϵ-最优强化学习策略，并假设转移概率分布属于某个已知核的再生核希尔伯特空间。一般性不失，假设其再生核希尔伯特空间范数的上界为 1。

假设 2。假设转移概率分布满足：

$$\forall s' \in \mathcal{S}: \quad \|P(s' \mid \cdot,\cdot)\|_{\mathcal{H}_K} \le 1.$$

鉴于再生核希尔伯特空间具有一般性，该假设非常灵活。这是一个标准假设，Yang 等人 (2020b) 也采用了该假设。我们没有对与策略相关的 Q 函数作出任何明确假设。回顾第 PV 节中 2.1 对 3 的定义，我们在引理 V 中证明：对于任意 $\mathcal{S} \to [0, \frac{1}{1-\gamma}]$，$\|PV\|_{\mathcal{H}_K} \le \frac{c}{1-\gamma}$，根据假设 2，这一点对我们的分析至关重要。

一些通用记号：对于任意实数 x以及实数 a、b，记号 $\Pi_{[a,b]}[x]$ 表示将 x 投影到 $[a,b]$上。对于任意整数 J，I_J 表示 $J \times J$ 单位矩阵，而 $\mathbf{0}_J$ 表示 $J \times 1$ 零向量。

3 基于核的 Q 学习

在本节中，我们提出一种新颖的基于核的 Q 学习算法，称为 KQLearn。回顾 $Q(s,a) = r(s,a) + \gamma[PV](s,a)$。转移概率分布 P 和价值函数 V 对算法而言均未知。因此，算法在各轮次中递归地近似PV，并利用上一轮观测数据中的 PV进行核岭回归。也就是说，该算法

基于核模型提供的 PV预测，执行近似贝尔曼算子更新。样本的收集依据核模型中 PV的不确定性进行。为此，算法首先构建一个最大不确定性集，并用它来收集样本。

最大不确定性集 ($\mathcal{U}_J$)：算法首先创建一个大小为 J 的最大不确定性集，称为 $\mathcal{U}_J \subset \mathcal{Z}$。此集合根据核模型提供的不确定性构建。具体而言，根据以下规则将每个状态-动作加入此集合：选择核模型中不确定性最高的状态-动作。

$$(s_j, a_j) = \arg\max_{(s,a)\in\mathcal{Z}} \Sigma^2_{\mathcal{U}_{j-1}}(s,a). \tag{6}$$

然后，以 $\mathcal{U}_0 = \varnothing$ 为起点，递归地执行 $\mathcal{U}_j = \mathcal{U}_{j-1} \cup \{(s_j, a_j)\}$。随后使用集合 $\mathcal{U}_J$ 从生成模型中收集样本。

算法按轮次推进，轮次编号为 $\ell = 1, \ldots, L$。每一轮 ℓ 都会接收来自上一轮 $\ell-1$ 的含噪观测 $Y_{(\ell-1)\mathcal{U}_J}$，这些观测对应于 PV。随后使用这些观测进行核岭回归，以构建覆盖整个 $\mathcal{Z}$ 的 PV 回归器，并获得新的观测 $Y_{(\ell)}$。观测向量初始化为零向量 $Y_{(0)} = \mathbf{0}_J$。

在每一轮 ℓ中，对于每个状态-动作对 $(s_j, a_j) \in \mathcal{U}_J$，从生成模型中获取转移状态 $s'_j \sim P(\cdot|s_j, a_j)$。随后，观测 $Y_{(\ell)}(s_j, a_j)$ 如下所示：

$$Y^{(\ell)}(s_j, a_j) = \Pi_{[0,\frac{1}{1-\gamma}]} \max_{a\in\mathcal{A}} \Big\{ r(s'_j, a) + \gamma k_{\mathcal{U}_J}(s'_j, a)^\top (K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1} Y^{(\ell-1)}_{\mathcal{U}_J} \Big\}. \tag{7}$$

右侧的第二项是针对 PV进行核岭回归的回归器，使用 $Y_{(\ell-1)\mathcal{U}_J}$ 作为观测向量。在分析中，我们给出了该回归误差的高概率界。向量$Y_{(\ell)\mathcal{U}_J} = [Y_{(\ell)}(s_1, a_1) \ldots, Y_{(\ell)}(s_J, a_J)]^\top$可理解为对 PV的更新后的含噪观测，并将其传递至下一轮 $\ell+1$。根据价值函数的定义和有界奖励假设，可以很容易地验证，对于所有 $s \in \mathcal{S}$，$0 \le V^*(s) \le \frac{1}{1-\gamma}$。因此，我们将 PV的值投影到 $[0, \frac{1}{1-\gamma}]$ 区间上。

KQLearn 总共收集 $N = JL$ 个样本。[4] 算法 1 中提供了伪代码1。

收集完所有样本后，KQLearn 算法返回一个强化学习策略 π，该策略根据以下代理 Q函数选择动作：

$$\widehat{Q}^{(L)}(s,a) = r(s,a) + \gamma k^\top_{\mathcal{U}_J}(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1} Y^{(L)}_{\mathcal{U}_J}. \tag{8}$$

具体而言，当观测到状态 s时，策略 π选择动作 $\pi(s) = \arg\max_{a\in\mathcal{A}} \widehat{Q}^{(L)}(s,a)$。右侧的第二项是 KQlearn 算法在第 L 轮中使用观测对 PV 进行的核岭回归。

算法 1 基于核的 Q 学习（KQLearn）

输入：折扣马尔可夫决策过程 M、核 K、正则化参数 $\lambda > 0$，以及 $N > 0$
输出 $\widehat{Q}^{(L)} : \mathcal{Z} \to \mathbb{R}$

初始化 $L, J \in \mathbb{N}$, $N = LJ$。
初始化 $Y^{(0)} = \mathbf{0}_J$和集合 $\mathcal{U}_0 = \varnothing$。
对于所有 $j = 1$、.....、J，执行
  使用公式 $\Sigma_{\mathcal{U}_{j-1}}(\cdot)$ 更新函数 2。
  选择 $(s_j, a_j) \leftarrow \arg\max_{(s,a)\in\mathcal{Z}} \Sigma^2_{\mathcal{U}_{j-1}}(s,a)$。
  $\mathcal{U}_j \leftarrow \mathcal{U}_{j-1} \cup \{(s_j, a_j)\}$。
结束循环
对所有 $\ell = 1, \ldots, L$ 执行 ▷ 轮次
  对于所有 $j = 1, \ldots, J$，执行
    获取 a 样本 转移 状态 $s' \sim P(\cdot|s_j, a_j)$
    更新 the $Y_{(\ell)}$ as 遵循。$Y_{(\ell)}(s_j, a_j) \leftarrow \Pi_{[0,\frac{1}{1-\gamma}]} \max_{a\in\mathcal{A}} \{r(s', a) + \gamma k^\top_{\mathcal{U}_J}(s', a)(K_{\mathcal{U}_J} + \lambda I_J)^{-1} Y_{(\ell-1)\mathcal{U}_J}\}$
  结束于
</style}结束循环
$\widehat{Q}^{(L)}(\cdot) = r(\cdot) + \gamma k^\top_{\mathcal{U}_J}(\cdot)(K_{\mathcal{U}_J} + \lambda^2 I^J_{\mathcal{U}_J})^{-1} Y_{(L)}$.

4 KQLearn 的样本复杂度

本节给出我们的理论结果。以下定理为 KQLearn 算法所获得的策略的价值函数误差建立了一个界。

定理 1. 考虑第 *2.1* 节所描述的折扣马尔可夫决策过程。考虑第 *3* 节所描述的 KQLearn 算法。在假设 *1* 和 *2* 下，以至少 $1-\delta$ 的概率，

$$\|V^\pi - V^*\|_\infty \le 2\beta(\delta)\left(\frac{\gamma}{1-\gamma}\right)^2 \sqrt{\frac{2\Gamma_{K,\lambda}(J)}{J}} + 2\gamma^{L-1}\left(\frac{1}{1-\gamma}\right)^2,$$

其中 $\beta(\delta) = \mathcal{O}\left(\frac{c}{1-\gamma} + \frac{1}{2\lambda(1-\gamma)}\sqrt{d\log\left(\frac{J}{(1-\gamma)\delta}\right)}\right)$ 且 c 是引理 *3* 中给出的常数。

[4] 为简化表述，我们假设 $N = JL$。当 J 不能整除 N 时，可以忽略最后一轮中的样本。

当 J和 L 都足够大时，上界中的两项都可以任意小。

4.1 样本复杂度

我们可以利用核特定的$\Gamma_{K,\lambda}$上界，得到 KQLearn 算法的显式样本复杂度；这些上界取决于 K 的 Mercer 特征值衰减率。具体而言，我们定义以下特征值衰减特征曲线（与 Chatterji 等人 (2019)；Vakili 等人 (2021c)；Yang 等人(2020b）中概述的类似）。

定义 2。 *[多项式和指数特征值衰减]*考虑 Mercer 特征值 $\{\sigma_m\}_{m=1}^{\infty}$，其对应于 K 方程*1*中按降序排列的特征值。

(i) 对于某些 $C_p > 0$、 $\beta_p > 1$，若对所有 $m \in \mathbb{N}$ 均有 $\sigma_m \leq C_p m^{-\beta_p}$，则称 K 具有 (C_p, β_p) 多项式特征值衰减。(ii) 对于某些 $C_{e,1}$、 $C_{e,2}$，若对所有 $m \in \mathbb{N}$ 均有 $\sigma_m \leq C_{e,1}\exp(-C_{e,2}m^{\beta_e})$，则称 K 具有 $(C_{e,1}、C_{e,2}、\beta_e)$ 指数特征值衰减。

现在，我们可以给出 Mercer 特征值呈多项式衰减和指数衰减的广义核类的样本复杂度显式界。

定理 2。 考虑第 *2.1* 节所述的折扣马尔可夫决策过程。考虑第 *3* 节所述的 KQLearn 算法，其中 $L = \Theta\left(\frac{\log(\epsilon(1-\gamma)^2}{1-\gamma}\right))$ 且 $J = \frac{N}{L}$。在假设 *1* 和 *2* 下，KQLearn 以至少 $1-\delta$ 的概率获得一个 ϵ-最优策略，样本复杂度至多为

- 对于具有（C_p， β_p）多项式特征值衰减的核，

$$N = \tilde{\mathcal{O}}\left(\frac{\left(\log(\frac{1}{\delta})\right)^{\frac{\beta_p}{\beta_p-1}}}{\epsilon^{\frac{2\beta_p}{\beta_p-1}}(1-\gamma)^{\frac{7\beta_p-1}{\beta_p-1}}}\right) \tag{9}$$

- 对于具有（$C_{e,1}$，$C_{e,2}$， β_e）指数特征值衰减的核，

$$N = \tilde{\mathcal{O}}\left(\frac{\log(\frac{1}{\delta})}{\epsilon^2(1-\gamma)^7}\right) \tag{10}$$

包含隐含常数及 $\tilde{}$ 中对数因子的详细表达式 $\mathcal{O}$ 记号见附录 B。

特定核函数：我们的样本复杂度界可针对多种核函数进行专门化，这些核函数的特征值衰减或 $\Gamma_{K,\lambda}$ 的界是已知的（例如 Srinivas 等人，2010年；Vakili 等人，2021c，a）。具体而言，对于 Mat´ern 核和 SE 核，我们分别有：

$$N = \tilde{\mathcal{O}}\left(\frac{\left(\log(\frac{1}{\delta})\right)^{1+\frac{d}{2\nu}}}{\epsilon^{2+\frac{d}{\nu}}(1-\gamma)^{7+\frac{3d}{\nu}}}\right), \text{ and } N = \tilde{\mathcal{O}}\left(\frac{\log(\frac{1}{\delta})}{\epsilon^2(1-\gamma)^7}\right).$$

4.2 样本复杂度的最优性

上述样本复杂度相对于 ϵ是阶最优的。我们将其与核老虎机样本复杂度的下界进行比较（当 $|\mathcal{S}| = 1$时，核老虎机是我们设定的一个特例）。具体而言，Scarlett 等人 (2017年)证明了 $\Omega\left((\frac{1}{\epsilon})^{2+\frac{d}{\nu}}\right)$ 以及$\Omega\left(\frac{1}{\epsilon^2}\right)$ 的样本复杂度下界，分别适用于 Mat´ern 核和 SE 核。我们的结果首次给出了非常一般情形下强化学习问题的有限样本复杂度，该情形涵盖所有具有多项式衰减特征值的核。

就折扣因子而言，在光滑核情形下，我们的样本复杂度按 $\frac{1}{(1-\gamma)^7}$ 的量级变化（与线性设定下的 PPQ-Learning 算法类似，Yang 和 Wang，2019）。然而，在表格型和线性设定下，这一结果已改进为 $\mathcal{O}(\frac{1}{(1-\gamma)^3})$。未来一个值得研究的问题是，在核设定下，样本复杂度对折扣因子的依赖是否也能改进为 $\mathcal{O}(\frac{1}{(1-\gamma)^3})$。正如引言中所讨论的，在核设定下，既无法观测所有状态-动作，也无法通过值迭代对模型进行参数化更新。因此，需要采用不同的算法设计与分析方法；在以下这些设定中，这一问题的挑战性依次递增：表格型→线性 → 基于核的。

5 分析

定理 2 是定理 1的推论，并利用了关于$\Gamma_{K,\lambda}$ 的核相关界。定理1的证明基于若干组成部分，包括跟踪核岭回归中的近似误差，以及近似贝尔曼算子的收敛误差。本节概述定理 1证明的主要步骤，详细内容则留待附录。

近似贝尔曼算子：回顾第2.1节中定义的贝尔曼算子。转移概率分布和价值函数是在连续域上复杂的非线性函数，算法无法获知。因此，我们定义一个近似贝尔曼算子$\hat{\mathcal{T}}$，它利用 PV在固定集合 $\mathcal{U}_J$上的含噪观测，并借助核岭回归执行近似贝尔曼算子更新。具体而言，对于所有 V： $\mathcal{S} \to [0, \frac{1}{1-\gamma}]$，以及固定集合 $\mathcal{U}_J \subset \mathcal{Z}$，

我们定义

$$[\widehat{\mathcal{T}}V](s) = \max_{a\in\mathcal{A}}\Big\{r(s,a) + \gamma k_{\mathcal{U}_J}^\top(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}V]_{\mathcal{U}_J}\Big\}, \quad (11)$$

其中，$[\widehat{PV}](s,a) = V(s')$ 是一个随机变量，$s' \sim P(\cdot|s,a)$ 是一个随机转移状态，并且

$$[\widehat{P}V]_{\mathcal{U}_J} = \Big[[\widehat{P}V](s_1,a_1),\ldots,[\widehat{P}V](s_J,a_J)\Big]^\top.$$

在 KQLearn 算法中，定义

$$\widehat{V}^{(\ell)}(s) = \max_{a\in\mathcal{A}}\Big\{r(s,a) + \gamma k_{\mathcal{U}_J}^\top(s,a)\left(K_{\mathcal{U}_J} + \lambda^2 I_J\right)^{-1} Y_{\mathcal{U}_J}^{(\ell)}\Big\}, \quad (12)$$

我们将其称为代理价值函数（类似于（8）中给出的代理 Q 函数）。于是，我们得到关于 ℓ 的以下递归关系。

$$\widehat{V}^{(\ell)} = \widehat{\mathcal{T}}\Pi_{[0,\frac{1}{1-\gamma}]}[\widehat{V}^{(\ell-1)}]. \quad (13)$$

代理价值函数中的误差：为了界定由 KQLearn 得到的策略 π 的价值函数误差，$\|V^\pi - V^*\|$我们需要界定（8）中给出的代理 Q 函数的误差，该函数用于得到 π。代理 Q 函数的误差可以根据第 $L-1$ 轮代理价值函数中的误差进行界定。具体而言，我们有

$$\|\widehat{Q}^{(L)} - Q^*\|_\infty \le \|\widehat{V}^{(L-1)} - V^*\|_\infty. \quad (14)$$

因此，接下来我们界定代理价值函数中的误差。代理价值函数中的误差可以写成两项之和：近似贝尔曼算子中的误差，以及使用真实贝尔曼算子时价值函数中的误差。具体来说，

$$\begin{aligned}\left\|\widehat{V}^{(L-1)} - V^*\right\|_\infty &= \left\|\widehat{\mathcal{T}}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(L-2)} - V^*\right\|_\infty \\ &\le \left\|\widehat{\mathcal{T}}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(L-2)} - \mathcal{T}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(L-2)}\right\|_\infty \\ &+ \left\|\mathcal{T}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(L-2)} - V^*\right\|_\infty. \end{aligned} \quad (15)$$

第二项可以递归地进行界定，从而得到定理 $\frac{2\gamma^{L-1}}{(1-\gamma)^2}$（1中的误差界的第二项。第一项则引出了分析中的一个重要步骤，该步骤基于核岭回归的误差。

具体而言，令 $V: \mathcal{S} \to [0, \frac{1}{1-\gamma}]$ 为任意价值函数。对于所有 $s\in\mathcal{S}$，我们有：

$$\begin{aligned}&[\widehat{\mathcal{T}}V](s) - [\mathcal{T}V](s) \\ &= \max_{a\in\mathcal{A}}\Big\{r(s,a) + \gamma k_{\mathcal{U}_J}^\top(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}V]_{\mathcal{U}_J}\Big\} \\ &\quad - \max_{a\in\mathcal{A}}\{r(s,a) + \gamma[PV](s,a)\} \\ &\le \gamma\max_{a\in\mathcal{A}}\Big\{k_{\mathcal{U}_J}^\top(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}V]_{\mathcal{U}_J} \\ &\quad - [PV](s,a)\Big\}. \end{aligned} \quad (16)$$

核岭回归中的误差：公式 max中16 内部的项是核岭回归中的误差，其中 $\sim PV$是目标函数，$[PV]_{\mathcal{U}_J}$是一组 J 含噪观测，且$k_{\mathcal{U}_J}^\top(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}V]_{\mathcal{U}_J}$是回归器。为了应用引理1，我们需要给出 PV的再生核希尔伯特空间范数上界，以及观测噪声在 PV中的次高斯参数上界。以下引理将证明这些结论。

引理 3。考虑一个可积价值函数 $V: \mathcal{S} \to [0, \frac{1}{1-\gamma}]$。在假设2下，有

$$\|PV\|_{\mathcal{H}_K} \le \frac{c}{1-\gamma}, \quad (17)$$

其中 $c = \int_{\mathcal{S}} V(s)ds$ 是由 $\mathcal{S}$的体积确定的常数。

引理 4。考虑一个转移概率分布P和一个可积价值函数 $V: \mathcal{S} \to [0, \frac{1}{1-\gamma}]$。对于所有 (s,a)，均有 $\mathbb{E}\left[[\widehat{PV}](s,a)\right] = PV(s,a)$。此外，$[PV](s,a)$是一个参数为1 $2(1-\gamma)$的次高斯随机变量。

引理 4 由 PV的定义以及有界随机变量的 Hoeffding 引理可得。引理 3 的证明见附录 C。

应用引理 1，我们以概率 $1-\delta$得到，对所有 $(s,a)\in\mathcal{Z}$，

$$\begin{aligned}&\left|k_{\mathcal{U}_J}^\top(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}V]_{\mathcal{U}_J} - [PV](s,a)\right| \\ &\quad \le \beta(\delta)\Sigma_{\mathcal{U}_J}(s,a), \end{aligned} \quad (18)$$

where $\beta(\delta) = \mathcal{O}\left(\frac{c}{1-\gamma} + \frac{1}{2(1-\gamma)\lambda}\sqrt{d\log(\frac{Jc}{(1-\gamma)\delta})}\right)$.

最后，利用关于总不确定性的引理 2，并根据 $\mathcal{U}_J$的设计，我们对等式右侧的 $\Sigma_{\mathcal{U}_J}(s,a)$ 进行界定。

因此，我们对（21）中的两项给出了界，该式界定了代理价值函数的误差。定理和引理的证明详情分别见附录 B 和附录C。

6 结论

现代强化学习通常面临庞大的状态-动作空间和复杂的模型。我们研究了核设定下带生成模型的折扣马尔可夫决策过程中的样本复杂度问题，延续了文献中的一系列研究（例如，参见 Kearns 和 Singh，1998年；Azar 等人，2017年；Sidford 等人，2018a，b；Yang 和 Wang，2019）。我们提出了一种新颖的基于核的 Q 学习算法，称为 KQLearn，并证明了其在非常一般的核类别下具有有限的样本复杂度上界。据我们所知，这是一般核设定下首个有限样本复杂度结果（涵盖所有具有多项式衰减特征值的核）。此外，与核老虎机问题这一特殊情形的下界相比，我们的样本复杂度在求得一个 ϵ -最优策略方面，与 ϵ相匹配。然而，我们的样本复杂度关于折扣因子的缩放可能并非最优，这仍是一个值得未来研究的开放问题。

参考文献

Abbasi-Yadkori, Y.（2013）。线性参数化控制问题的在线学习。Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W.（2019）。强化学习：理论与算法。美国华盛顿州西雅图，华盛顿大学计算机科学系，技术报告，第 10–4 页。Agarwal, A., Kakade, S., a nd Yang, L. F.（2020）。基于生成模型的基于模型的强化学习是极小极大最优的。载于《学习理论会议》，第 67–83 页。PMLR。Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L.（2020）。基于价值目标回归的基于模型的强化学习。载于《国际机器学习会议》，第 463–474 页。PMLR。Azar, M. G., Munos, R., and Kappen, H. J.（2013）。具有生成模型的强化学习样本复杂度的极小极大 PAC 界。《机器学习》，91(3)：325–349。Azar, M. G., Osband, I., and Munos, R.（2017）。强化学习的极小极大遗憾界。载于《国际机器学习会议》，第 263–272 页。PMLR。Carpentier, A., Vernade, C., and Abbasi-Yadkori, Y.（2020）。重新审视椭圆势引理。arXiv 预印本 arXiv:2010.10182。Chatterji, N., Pacchiano, A., and Bartlett, P.（2019）。使用核损失的在线学习。载于《机器学习研究论文集》，第 97 卷，第 971–980 页，美国加利福尼亚州长滩。PMLR。

Chowdhury, S. R. and Gopalan, A.（2017）。论核化多臂老虎机。载于《国际机器学习会议》，第 844–853 页。PMLR。Christmann, A. and Steinwart, I.（2008）。《支持向量机》。施普林格，纽约。Domingues, O. D., M´enard, P., Pirotta, M., Kaufmann, E., and Valko, M.（2021）。基于核的强化学习：有限时间分析。载于《国际机器学习会议》，第 2783–2792 页。PMLR。

Fawzi, A., Balog, M., Huang, A., Hubert, T., Romera-Paredes, B., Barekatain, M., Novikov, A., R Ruiz, F. J., Schrittwieser, J., Swirszcz, G. 等（2022年）。利用强化学习发现更快的矩阵乘法算法。《自然》，610(7930):47–53。

Jin, C., Allen-Zhu, Z., Bubeck, S. 和 Jordan, M. I.（2018）。Q 学习是否可证明地高效？《神经信息处理系统进展》，31。Jin, C., Yang, Z., Wang, Z. 和 Jordan, M. I.（2020）。基于线性函数近似的可证明高效强化学习。载于《学习理论会议》，第 2137–2143 页。PMLR。Kahn, G., Villaflor, A., Pong, V., Abbeel, P. 和 Levine, S.（2017）。用于碰撞避免的不确定性感知强化学习。arXiv 预印本 arXiv:1702.01182。Kakade, S., Krishnamurthy, A., Lowrey, K., Ohnishi, M. 和 Sun, W.（2020）。在线非线性控制的信息论遗憾界。《神经信息处理系统进展》，33:15312–15325。Kakade, S. M.（2003）。论强化学习的样本复杂度。伦敦大学，伦敦大学学院（英国）。Kalashnikov, D., Irpan, A., Pastor, P., Ibarz, J., Herzog, A., Jang, E., Quillen, D., Holly, E., Kalakrishnan, M., Vanhoucke, V. 等（2018）。用于基于视觉的机器人操作的可扩展深度强化学习。载于《机器人学习会议》，第 651–673 页。PMLR。Kearns, M. 和 Singh, S.（1998）。Q 学习和间接算法的有限样本收敛率。载于《神经信息处理系统进展》，第 11 卷。麻省理工学院出版社。Lee, K., Kim, S.-A., Choi, J. 和 Lee, S.-W.（2018）。连续动作空间中的深度强化学习：模拟冰壶游戏案例研究。载于《国际机器学习会议》，第 2937–2946 页。PMLR。Li, Z. 和 Scarlett, J.（2022）。少批次高斯过程赌博机优化。载于《国际人工智能与统计会议》，第 92–107 页。PMLR。Mercer, J.（1909年）。正型与负型函数及其与积分方程理论的联系。《皇家学会哲学汇刊》

伦敦。A辑，含数学或物理性质论文，209:415–446。Mirhoseini, A., Goldie, A., Yazgan, M., Jiang, J. W., Songhori, E., Wang, S., Lee, Y.-J., Johnson, E., Pathak, O., Nazi, A. 等（2021）。用于快速芯片设计的图布局方法。《自然》，594(7862):207–212。Puterman, M. L.（2014）。马尔可夫决策过程：离散随机动态规划。约翰·威立父子公司。Russo, D.（2019）。基于随机化价值函数进行探索的最坏情形遗憾界。《神经信息处理系统进展》，32。Salgia, S., Vakili, S. 和 Zhao, Q.（2021）。一种基于定义域收缩、具有阶最优遗憾表现的贝叶斯优化算法。《神经信息处理系统进展》，34:28836–28847。Scarlett, J., Bogunovic, I. 和 Cevher, V.（2017）。带噪高斯过程赌博机优化的遗憾下界。载于《学习理论会议》，第 1723–1742 页。PMLR。Sch¨olkopf, B., Smola, A. J., Bach, F. 等（2002）。核方法学习：支持向量机、正则化、优化及其他。麻省理工学院出版社。Shahriari, B., Swersky, K., Wang, Z., Adams, R. P. 和 De Freitas, N.（2015）。让人类退出循环：贝叶斯优化综述。IEEE会议录，104(1):148–175。Sidford, A., Wang, M., Wu, X., Yang, L. 和 Ye, Y.（2018a）。利用生成模型求解马尔可夫决策过程的近最优时间与样本复杂度。《神经信息处理系统进展》，31。Sidford, A., Wang, M., Wu, X. 和 Ye, Y.（2018b）。方差缩减值迭代及更快的马尔可夫决策过程求解算法。载于《第二十九届ACM-SIAM离散算法年度研讨会》，第 770–787 页。工业与应用数学学会。Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. 等（2016）。利用深度神经网络和树搜索掌握围棋。《自然》，529(7587):484–489。Snoek, J., Larochelle, H. 和 Adams, R. P.（2012）。机器学习算法的实用贝叶斯优化。《神经信息处理系统进展》，25。Srinivas, N., Krause, A., Kakade, S. 和 Seeger, M.（2010）。赌博机设定中的高斯过程优化：无遗憾与实验设计。载于《ICML 2010——第 27 届国际机器学习会议论文集》，第 1015–1022 页。

Vakili, S., Bouziani, N., Jalali, S., Bernacchia, A., 和 Shiu, D.-S. （2021a）。高斯过程多臂老虎机的最优阶简单遗憾。《神经信息处理系统进展》，34:21202–21215。Vakili, S., Bromberg, M., Garcia, J., Shiu, D.-s., 和 Bernacchia, A.（2021b）。过参数化神经网络的一致泛化界。arXiv 预印本 arXiv:2109.06099。Vakili, S., Khezeli, K., 和 Picheny, V.（2021c）。关于高斯过程多臂老虎机中的信息增益与遗憾界。载于《国际人工智能与统计会议》，第 82–90 页。PMLR。Vakili, S., Scarlett, J., 和 Javidi, T.（2021d）。开放问题：再生核希尔伯特空间元素的紧致在线置信区间。载于《学习理论会议》，第 4647–4652 页。PMLR。Vakili, S., Scarlett, J., Shiu, D.-S., 和 Bernacchia, A.（2022）。基于核的学习中稀疏近似方法的改进收敛速率。arXiv 预印本 arXiv:2202.04005。Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., Choi, D. H., Powell, R., Ewalds, T., Georgiev, P., 等（2019）。使用多智能体强化学习达到《星际争霸II》宗师级水平。《自然》，575（7782）：350–354。Wang, R., Salakhutdinov, R., 和 Yang, L. F.（2020）。通过一般价值函数逼近实现可证明高效的强化学习。arXiv 预印本 arXiv:2005.10804。Yang, L. 和 Wang, M.（2019）。使用线性加性特征的样本最优参数化 Q 学习。载于《国际机器学习会议》，第 6995–7004 页。PMLR。Yang, L. 和 Wang, M.（2020）。特征空间中的强化学习：矩阵多臂老虎机、核与遗憾界。载于《国际机器学习会议》，第 10746–10756 页。PMLR。Yang, Z., Jin, C., Wang, Z., Wang, M., 和 Jordan, M.（2020a）。通过核函数与神经函数逼近实现可证明高效的强化学习。《神经信息处理系统进展》，33：13903–13916。Yang, Z., Jin, C., Wang, Z., Wang, M., 和 Jordan, M. I.（2020b）。强化学习中的函数逼近：大状态空间下的乐观主义。arXiv 预印本 arXiv:2011.04622。Zhou, D., He, J., 和 Gu, Q.（2021）。使用特征映射的折扣马尔可夫决策过程的可证明高效强化学习。载于《国际机器学习会议》，第 12793–12802 页。PMLR。

在附录中，我们提供了一些因篇幅限制而未在正文中展开的细节和证明。在附录 A 中，我们给出 Mercer 定理的正式表述以及再生核希尔伯特空间的构造性定义。定理 1 和 2的证明以及辅助引理分别见附录 B 和附录 C。

Mercer 定理

Mercer 定理（Mercer，1909年）给出了核的无限维特征映射表示（例如，参见Christmann 和 Steinwart（2008 年），定理 4.49）。设 $\mathcal{Z}$ 是一个紧致度量空间，且 μ 是 $\mathcal{Z}$ 上的有限博雷尔测度（我们考虑欧几里得空间上的勒贝格测度）。令 L^2_μ（$\mathcal{Z}$）为 $\mathcal{Z}$上关于 μ的平方可积函数集合。进一步地，如果一个核满足

$$\int_{\mathcal{Z}}\int_{\mathcal{Z}} K(z,z')^2\, d\mu(z) d\mu(z') < \infty.$$

定理 3.（Mercer 定理）设 $\mathcal{Z}$ 是一个紧致度量空间，且 μ 是 $\mathcal{Z}$上的有限博雷尔测度。令 K 为一个连续且平方可积的核，它诱导出积分算子 T_K： $L^2_\mu(\mathcal{Z}) \to L^2_\mu(\mathcal{Z})$，定义如下：

$$\left(T_K f\right)(\cdot) = \int_{\mathcal{Z}} K(\cdot, z') f(z')\, d\mu(z'),$$

其中 $f \in L^2_\mu(\mathcal{Z})$。那么，存在一列特征值—特征函数对 $\{(\sigma_m, \psi_m)\}_{m=1}^{\infty}$，使得 $\sigma_m > 0$，并且 $T_K\psi_m = \sigma_m\psi_m$，对任意 $m \geq 1$成立。此外，核函数可以表示为

$$K(z,z') = \sum_{m=1}^{\infty} \sigma_m \psi_m(z)\psi_m(z'),$$

其中，该级数在 $\mathcal{Z}\times\mathcal{Z}$ 上一致收敛。

根据 Mercer 表示定理（例如参见 Christmann 和 Steinwart (2008 年)，定理 4.51），由 K 诱导的再生核希尔伯特空间因此可以用 $\{(\sigma_m, \psi_m)\}_{m=1}^{\infty}$来表示。

定理 4.（Mercer 表示定理）设 $\{(\sigma_m, \psi_m)\}_{i=1}^{\infty}$ 为 Mercer 特征值—特征函数对。那么， K 的再生核希尔伯特空间为

$$\mathcal{H}_K = \left\{ f(\cdot) = \sum_{i=1}^{\infty} w_i \sigma_i^{\frac{1}{2}} \psi_i(\cdot) : w_i \in \mathbb{R}, \|f\|^2_{\mathcal{H}_K} := \sum_{i=1}^{\infty} w_i^2 < \infty \right\}$$

Mercer 表示定理指出，缩放特征函数 $\{\sqrt{\sigma_i}\psi_i\}_{i=1}^{\infty}$ 构成 $\mathcal{H}_K$ 的标准正交基。

B 定理证明

本节将给出主要定理的证明。

B.1 定理1的证明。

定理 1 的证明基于一个近似贝尔曼算子，该算子使用 KQLearn 轮次中对 PV的含噪观测，以及核岭回归。我们证明了该近似贝尔曼算子的误差界，并据此界定 KQLearn 所获得策略的价值函数误差。

近似贝尔曼算子：回顾第2.1节中定义的贝尔曼算子。转移概率分布和价值函数是在连续域上复杂的非线性函数，算法无法获知。因此，我们定义一个近似贝尔曼算子 $\mathcal{T}$，它使用 PV在固定集合 $\mathcal{U}_J$上的含噪观测，并利用核岭回归执行近似贝尔曼算子更新。具体而言，对于所有 V： $\mathcal{S} \to [0, \frac{1}{1-\gamma}]$，以及固定集合 $\mathcal{U}_J \subset \mathcal{Z}$，定义如下：

$$[\widehat{\mathcal{T}}V](s) = \max_{a\in\mathcal{A}} \left\{ r(s,a) + \gamma k^\top_{\mathcal{U}_J}(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}V]_{\mathcal{U}_J} \right\},$$

^其中 $[PV](s,a) = V(s')$ 是一个随机变量，$s' \sim P(\cdot|s,a)$ 是一个随机转移状态，且 d

$$[\widehat{P}V]_{\mathcal{U}_J} = \left[[\widehat{P}V](s_1,a_1),\ldots,[\widehat{P}V](s_J,a_J)\right]^\top.$$

在 KQLearn 算法中，定义

$$\widehat{V}^{(\ell)}(s) = \max_{a\in\mathcal{A}}\left\{r(s,a) + \gamma k_{\mathcal{U}_J}^\top(s,a)\left(K_{\mathcal{U}_J} + \lambda^2 I_J\right)^{-1} Y_{\mathcal{U}_J}^{(\ell)}\right\},$$

我们将其称为代理价值函数（类似于（8）中给出的代理 Q 函数）。于是，我们得到关于 ℓ 的如下递归关系。

$$\widehat{V}^{(\ell)} = \widehat{\mathcal{T}}\Pi_{[0,\frac{1}{1-\gamma}]}[\widehat{V}^{(\ell-1)}]. \tag{19}$$

代理价值函数中的误差：接下来，我们界定代理价值函数中的误差。代理价值函数中的误差可以写成两项之和：近似贝尔曼算子中的误差，以及使用真实贝尔曼算子时价值函数中的误差。具体而言，对于 $l > 1$,

$$\begin{aligned}\left\|\widehat{V}^{(l-1)} - V^*\right\|_\infty &= \left\|\widehat{\mathcal{T}}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(l-2)} - V^*\right\|_\infty \\ &\le \underbrace{\left\|\widehat{\mathcal{T}}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(l-2)} - \mathcal{T}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(l-2)}\right\|_\infty}_{\text{Term I}} \\ &\quad + \underbrace{\left\|\mathcal{T}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(l-2)} - V^*\right\|_\infty}_{\text{Term II}}.\end{aligned} \tag{20}$$

现在我们来界定（20）右侧的两项。

项 I：第一项引出了分析中的一个重要步骤，该步骤基于核岭回归中的误差。具体而言，令 $V\colon \mathcal{S}\to[0,\frac{1}{1-\gamma}]$ 为任意价值函数。对于所有 $s\in\mathcal{S}$，我们有，

$$\begin{aligned}[\widehat{\mathcal{T}}V](s) - [\mathcal{T}V](s) &= \max_{a\in\mathcal{A}}\left\{r(s,a) + \gamma k_{\mathcal{U}_J}^\top(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}V]_{\mathcal{U}_J}\right\} - \max_{a\in\mathcal{A}}\{r(s,a) + \gamma[PV](s,a)\} \\ &\le \gamma\max_{a\in\mathcal{A}}\left\{k_{\mathcal{U}_J}^\top(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}V]_{\mathcal{U}_J} - [PV](s,a)\right\}.\end{aligned} \tag{21}$$

核岭回归中的误差：方程max中的21内部的项是核岭回归中的误差，其中PV是目标函数，$[\widehat{P}V]_{\mathcal{U}_J}$是一组 J含噪观测，而$k_{\mathcal{U}_J}^\top(s,\ a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}V]_{\mathcal{U}_J}$是回归器。为了应用引理1，我们需要给出 PV的再生核希尔伯特空间范数上界，以及^中观测噪声的次高斯性参数上界。我们分别在引理3 和4中证明了这两点。具体而言，有

$$\|PV\|_{\mathcal{H}_K} \le \frac{c}{1-\gamma}. \tag{22}$$

并且，$[\widehat{P}V](s,a)$ 是一个参数为1 2(1−γ)的次高斯随机变量。 ——

应用引理1，我们以 $1-\delta$的概率，对所有$(s,\ a)\in\mathcal{Z}$,

$$\left|k_{\mathcal{U}_J}^\top(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}V]_{\mathcal{U}_J} - [PV](s,a)\right| \le \beta(\delta)\Sigma_{\mathcal{U}_J}(s,a), \tag{23}$$

其中 $\beta(\delta) = \mathcal{O}\left(\frac{c}{1-\gamma} + {}_2\ \underline{{}^1{}_{(1-\gamma)\lambda}}\ \underline{\sqrt{d\log(}}\ {}^{Jc}{}_{(1-\gamma)}\right.$
$\left._\delta)\right)$。

对 $\Sigma_{\mathcal{U}_J}(s,a)$ 进行界定：对于所有 $j \le J$，在较小的观测子集上进行条件化会降低方差 $\Sigma_{\mathcal{U}_j}(s,a) \ge \Sigma_{\mathcal{U}_J}(s,a)$（这是由于核矩阵的正定性）。根据观测点的选择规则：

$$(s_j, a_j) = \arg\max_{(s,a)\in\mathcal{Z}} \Sigma^2_{\mathcal{U}_{j-1}}(s,a), \tag{24}$$

我们有 $\Sigma_{\mathcal{U}_{j-1}}(s,a) \le \Sigma_{\mathcal{U}_{j-1}}(s_j,a_j)$。因此，对于所有 $(s,a) \in \mathcal{Z}$,

$$\begin{aligned}\Sigma^2_{\mathcal{U}_J}(s,a) &\le \frac{1}{J}\sum_{j=1}^{J}\Sigma^2_{\mathcal{U}_{j-1}}(s,a)\\ &\le \frac{1}{J}\sum_{j=1}^{J}\Sigma^2_{\mathcal{U}_{j-1}}(s_j,a_j)\\ &\le \frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J},\end{aligned}$$

其中最后一行由引理2得出。

替换 $\Sigma_{\mathcal{U}_J}(s,a)$ 的上界后，对于所有 $V\colon \mathcal{S}\to[0,\frac{1}{1-\gamma}]$, $s\in\mathcal{S}$,

$$[\widehat{\mathcal{T}}V](s) - [\mathcal{T}V](s) \le \gamma\beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}}. \tag{25}$$

因此,

$$\left\|\widehat{\mathcal{T}}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(l-2)} - \mathcal{T}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(l-2)}\right\|_\infty \le \gamma\beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}}. \tag{26}$$

项 II: 我们现在式（20）右侧的第二项，根据贝尔曼算子的压缩性进行界定 r.

$$\begin{aligned}\left\|\mathcal{T}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(l-2)} - V^*\right\|_\infty &\le \gamma\left\|\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(l-2)} - V^*\right\|_\infty\\ &\le \gamma\left\|\widehat{V}^{(l-2)} - V^*\right\|_\infty.\end{aligned} \tag{27}$$

第二个不等式由以下观测可得：对于所有 $s\in\mathcal{S}$, $V^*(s) \in [0,\frac{1}{1-\gamma}]$。

结合对项 I 和项 II 的界定，我们得到

$$\left\|\widehat{V}^{(l-1)} - V^*\right\|_\infty \le \gamma\beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}} + \gamma\left\|\widehat{V}^{(l-2)} - V^*\right\|_\infty.$$

递归地利用轮次 $l-1$、...、$L-1$ 的误差来界定轮次 l 的代理价值函数误差，我们有：

$$\begin{aligned}\left\|\widehat{V}^{(L-1)} - V^*\right\|_\infty &\le \beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}}\left(\sum_{i=1}^{L-1}\gamma^i\right) + \gamma^{L-1}\left\|\widehat{V}^{(0)} - V^*\right\|_\infty\\ &\le \beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}}\left(\sum_{i=1}^{L-1}\gamma^i\right) + \frac{\gamma^{L-1}}{1-\gamma},\end{aligned}$$

其中，第二个不等式来自 $\left\|\widehat{V}_{(0)} - V^*\right\|_\infty \le \frac{1}{1-\gamma}$。

回顾以下定义 式（$\widehat{Q}^{(L)}$ 8）中给出的代理 Q 函数 $\widehat{Q}^{(L)}$ 的误差界如下。对于所有 $(s,a) \in \mathcal{Z}$,

$$
\begin{aligned}
\left\|\widehat{Q}^{(L)}(s,a) - Q^*(s,a)\right\|_\infty &= \left\|\left(r(s,a) + \gamma k_{\mathcal{U}_J}^\top(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1} Y_{\mathcal{U}_J}^{(L)}\right) - \left(r(s,a) + \gamma[PV^*](s,a)\right)\right\|_\infty \\
&\le \left\|\gamma\left(k_{\mathcal{U}_J}^\top(s,a)(K_{\mathcal{U}_J} + \lambda^2 I_J)^{-1}[\widehat{P}\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(L-1)}]_{\mathcal{U}_J} - [P\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(L-1)}](s,a)\right)\right\|_\infty \\
&\quad + \left\|\gamma\left([P\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(L-1)}](s,a) - [PV^*](s,a)\right)\right\|_\infty \\
&\le \gamma\beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}} + \gamma\left\|\mathbb{E}_{s'\sim P(\cdot|s,a)}[\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(L-1)}] - \mathbb{E}_{s'\sim P(\cdot|s,a)}[V^*]\right\|_\infty \\
&\le \gamma\beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}} + \gamma\left\|\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(L-1)} - V^*\right\|_\infty \\
&\le \gamma\beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}} + \gamma\left\|\widehat{V}^{(L-1)} - V^*\right\|_\infty \\
&\le \gamma\beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}} + \gamma\left(\beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}}\left(\sum_{i=1}^{L-1}\gamma^i\right) + \frac{\gamma^{L-1}}{1-\gamma}\right) \\
&= \beta(\delta)\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}}\left(\sum_{i=1}^{L}\gamma^i\right) + \frac{\gamma^L}{1-\gamma} \\
&\le \frac{\gamma\beta(\delta)}{1-\gamma}\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}} + \frac{\gamma^L}{1-\gamma}.
\end{aligned}
$$

第一个不等式是在加上并减去项 $\gamma[P\Pi_{[0,\frac{1}{1-\gamma}]}\widehat{V}^{(L-1)}](s,a)$ 后，根据三角不等式得到的。第二个不等式是项 I 中核岭回归误差的上界。第三个不等式利用最大差值对期望之差进行界定。第四个不等式由观测到 $V^* \in [0, \frac{1}{1-\gamma}]$ 区间这一事实得出。第五个不等式则利用了上文给出的代理价值函数误差界。

π 的价值函数由代理 Q 函数得到：以下引理表明，利用代理 Q 函数的误差，可以界定 π 的价值函数误差。

引理 5. 考虑任意满足 $\|\tilde{Q} - Q^*\|_\infty \le \epsilon$ 的 Q 函数 $\tilde{Q}: \mathcal{S}\times\mathcal{A} \to [0, \frac{1}{1-\gamma}+1]$。定义策略 $\pi_{\tilde{Q}}$，使得 $\pi_{\tilde{Q}}(s) = \arg\max_{a\in\mathcal{A}} Q(s,a)$。那么，对所有 $s \in \mathcal{S}$,

$$
V^*(s) - V^{\pi_{\tilde{Q}}}(s) \le \frac{2\epsilon}{1-\gamma}.
$$

应用引理 5 于 Q （由 KQLearn 返回），可知以概率 $1-\delta$,

$$
\|V^\pi - V^*\|_\infty \le \frac{2\gamma\beta(\delta)}{(1-\gamma)^2}\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+1/\lambda^2)J}} + \frac{2\gamma^L}{(1-\gamma)^2}.
$$

证明至此完成。

B.2 定理 2 的证明

定理 2 是定理 1 以及针对特定核的 $\Gamma_{K,\lambda}$ 上界的推论。已有多项研究针对各种核给出了 $\Gamma_{K,\lambda}$ 的上界（Srinivas 等人，2010年；Vakili 等人，2021c,b）。对于具有多项式和指数特征值衰减的核，我们采用 Vakili 等人(2021c) 的结果。

多项式特征值衰减： 考虑一个具有 (C_p, β_p) 多项式特征值衰减的核。我们对$\Gamma_{K,\lambda}$ (Vakili等人，2021c) 有如下界。对所有 $J \in \mathbb{N}$,

$$
\Gamma_{K,\lambda}(J) = \mathcal{O}(J^{\frac{1}{\beta_p}}\log^{1-\frac{1}{\beta_p}}(J)) \tag{28}
$$

We 将此界代入 KQLearn 得到的 π 的价值函数误差中，见定理1

$$\|V^{\pi}-V^{*}\|_{\infty} \quad \leq \quad \frac{2\gamma\beta(\delta)}{(1-\gamma)^{2}}\sqrt{\frac{2\Gamma_{K,\lambda}(J)}{\log(1+\frac{1}{\lambda^{2}})J}}+\frac{2\gamma^{L}}{(1-\gamma)^{2}}.$$

于是我们得到

$$\|V^{\pi}-V^{*}\|_{\infty} \quad = \quad \mathcal{O}\left(\frac{\gamma}{(1-\gamma)^{3}}\left(c+\frac{1}{\lambda}\sqrt{d\log(\frac{Jc}{(1-\gamma)\delta})}\right)\sqrt{J^{\frac{1}{\beta_{p}}-1}\log^{1-\frac{1}{\beta_{p}}}(J)}\right)+\mathcal{O}\left(\frac{\gamma^{L}}{(1-\gamma)^{2}}\right).$$

我们选择足够大的 J和 L，使得右侧的每一项都不超过 $\epsilon/2$。对

$$\begin{aligned} J &= \mathcal{O}\left(\left(\frac{1}{\epsilon}\right)^{\frac{2\beta_{p}}{\beta_{p}-1}}\frac{\gamma^{\frac{2\beta_{p}}{\beta_{p}-1}}}{(1-\gamma)^{\frac{6\beta_{p}}{\beta_{p}-1}}}\left(c+\frac{1}{\lambda}\sqrt{d\log(\frac{c}{(1-\gamma)\delta})}\right)^{\frac{2\beta_{p}}{\beta_{p}-1}}\log^{\frac{\beta_{p}}{\beta_{p}-1}}\left(\frac{1}{\epsilon(1-\gamma)}\right)\right), \\ L &= \mathcal{O}\left(\frac{\log(\frac{1}{\epsilon})+\log(\frac{1}{1-\gamma})}{(1-\gamma)}\right), \end{aligned} \tag{29}$$

采用适当的常数可确保 $\|V^{\pi}-V^{*}\|_{\infty}\leq\epsilon$。省略对数项和常数项后，该表达式可简化为 $J=\tilde{\mathcal{O}}\left(\frac{\log(\frac{1}{\delta}\quad\frac{\beta_{p}}{\beta_{p}-1})\big)\,(1-\gamma)^{\frac{6\beta_{p}}{\beta_{p}-1}}}{\epsilon^{\frac{2\beta_{p}}{\beta_{p}-1}}}\right)$和$L=\tilde{\mathcal{O}}\left(\frac{1}{1-\gamma}\right)$。由此得到

$$N=\tilde{\mathcal{O}}\left(\frac{\left(\log(\frac{1}{\delta})\right)^{\frac{\beta_{p}}{\beta_{p}-1}}}{\epsilon^{\frac{2\beta_{p}}{\beta_{p}-1}}\,(1-\gamma)^{\frac{7\beta_{p}-1}{\beta_{p}-1}}}\right). \tag{30}$$

指数特征值衰减：考虑一个具有（C_{e_1}， C_{e_2}， β_e）多项式特征值衰减的核。我们有关于 $\Gamma_{K,\lambda}$ 的如下界。对于所有 $J\in\mathbb{N}$,

$$\Gamma_{K,\lambda}(J)=\mathcal{O}(\log^{1+\frac{1}{\beta_{e}}}(J)) \tag{31}$$

我们将其替换为 KQLearn 得到的价值函数误差上界见 π 定理1，并得到

$$\|V^{\pi}-V^{*}\|_{\infty} \quad = \quad \mathcal{O}\left(\frac{\gamma}{(1-\gamma)^{3}}\left(c+\frac{1}{\lambda}\sqrt{d\log(\frac{Jc}{(1-\gamma)\delta})}\right)\sqrt{\frac{\log^{1+\frac{1}{\beta_{e}}}(J)}{J}}\right)+\mathcal{O}\left(\frac{\gamma^{L-1}}{(1-\gamma)^{2}}\right).$$

我们选择 J和 L 足够大，使得右侧的每一项都被 $\epsilon/2$所界定。选择 o f

$$\begin{aligned} J &= \mathcal{O}\left(\left(\frac{1}{\epsilon}\right)^{2}\frac{\gamma^{2}}{(1-\gamma)^{6}}\left(c+\frac{1}{\lambda}\sqrt{d\log(\frac{c}{(1-\gamma)\delta})}\right)^{2}\log^{2+\frac{1}{\beta_{e}}}\left(\frac{1}{\epsilon(1-\gamma)}\right)\right), \\ L &= \mathcal{O}\left(\frac{\log(\frac{1}{\epsilon})+\log(\frac{1}{1-\gamma})}{(1-\gamma)}\right), \end{aligned} \tag{32}$$

取适当的常数即可保证 $\|V^{\pi}-V^{*}\|_{\infty}\leq\epsilon$。该表达式可简化为 $J=\tilde{\mathcal{O}}\left(\frac{\log(\frac{1}{\delta})}{\epsilon^{2}(1-\gamma)^{6}}\right)$ 以及 $L=\tilde{\mathcal{O}}\left(\frac{1}{1-\gamma}\right)$，省略对数项和常数项。由此可得

$$N=\tilde{\mathcal{O}}\left(\frac{\log(\frac{1}{\delta})}{\epsilon^{2}\,(1-\gamma)^{7}}\right). \tag{33}$$

C 引理证明

本节将给出辅助引理的证明。

C.1 引理证明 3 [再生核希尔伯特空间范数 PV]

我们有

$$\begin{aligned}\|PV\|_{\mathcal{H}_K} &= \left\|\int_{s'\in\mathcal{S}} P(s'|\cdot,\cdot)V(s')ds'\right\|_{\mathcal{H}_K} \\ &\le \int_{s'\in\mathcal{S}} \|P(s'|\cdot,\cdot)V(s')\|_{\mathcal{H}_K}\, ds' \\ &= \int_{s'\in\mathcal{S}} \|P(s'|\cdot,\cdot)\|_{\mathcal{H}_K} V(s')ds' \\ &\le \int_{s'\in\mathcal{S}} V(s')ds' \end{aligned} \tag{34}$$

其中，最后一个不等式由假设 2 得出。注意，$\int_{s'\in\mathcal{S}} V(s')ds' \le \frac{1}{1-\gamma}\int_{s'} ds' \le \frac{c}{1-\gamma}$ 其中 c 是 $\mathcal{S}$ 的体积。

C.2 引理证明 4 [次高斯性 PV]

第一部分，$\mathbb{E}\left[[\widehat{P}V](s,a)\right] = PV(s,a)$，由 $\widehat{P}V$ 的定义可得。对于第二部分，注意到根据定义，$\widehat{P}V$ 是一个取值范围有界于 $[0,\frac{1}{1-\gamma}]$ 的随机变量。Hoeffding 引理指出：设 X 是任意满足 $0\le X\le a$ 的随机变量，则对任意 $\zeta\in\mathbb{R}$，$\mathbb{E}\left[\exp\left(\zeta(X-\mathbb{E}[X])\right)\right] \le \exp(\frac{\zeta^2 a^2}{8})$。应用 Hoeffding 引理可知，$\widehat{P}V$ 是参数为 $\frac{1}{2(1-\gamma)}$ 的次高斯变量。

C.3 引理 5 [关于代理 Q 函数误差所导致的价值函数误差的证明]

证明步骤与 Agarwal 等人 (2019) 中引理 1.11 的证明类似，该证明考虑了有限状态-动作对。首先，回顾第 2 节中的以下定义，

$$Q^{\pi_{\tilde{Q}}}(s,a) = r(s,a) + \gamma[PV^{\pi_{\tilde{Q}}}](s,a) \quad \text{and} \quad Q^*(s,a) = \max_{\pi} Q(s,\pi(s))$$

固定状态 $s\in\mathcal{S}$，并令 $\tilde{a}_s = \pi_{\tilde{Q}}(s)$。我们有

$$\begin{aligned} V^*(s) - V^{\pi_{\tilde{Q}}}(s) =& Q^*\left(s,\pi^*(s)\right) - Q^{\pi_{\tilde{Q}}}(s,\tilde{a}_s) \\ =& Q^*\left(s,\pi^*(s)\right) - Q^*(s,\tilde{a}_s) + Q^*(s,\tilde{a}_s) - Q^{\pi_{\tilde{Q}}}(s,\tilde{a}_s) \\ =& Q^*\left(s,\pi^*(s)\right) - Q^*(s,\tilde{a}_s) + \gamma[PV^*](s,\tilde{a}_s) - \gamma[PV^{\pi_{\tilde{Q}}}](s,\tilde{a}_s) \\ =& Q^*\left(s,\pi^*(s)\right) - Q^*(s,\tilde{a}_s) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,\tilde{a}_s)}\left[V^*\left(s'\right) - V^{\pi_{\tilde{Q}}}\left(s'\right)\right] \\ \le& Q^*\left(s,\pi^*(s)\right) - \tilde{Q}\left(s,\pi^*(s)\right) + \tilde{Q}(s,\tilde{a}_s) - Q^*(s,\tilde{a}_s) \\ &+ \gamma\mathbb{E}_{s'\sim P(s,\tilde{a}_s)}\left[V^*\left(s'\right) - V^{\pi_{\tilde{Q}}}\left(s'\right)\right] \\ \le& 2\left\|\tilde{Q} - Q^*\right\|_\infty + \gamma\left\|V^* - V^{\pi_{\tilde{Q}}}\right\|_\infty, \end{aligned}$$

其中，第一个不等式根据 $\tilde{a}_s$ 的定义，使用了 $Q\left(s,\pi^*(s)\right) \le Q(s,\tilde{a}_s)$。因此我们有

$$\left\|V^* - V^{\pi_{\tilde{Q}}}\right\|_\infty \le 2\left\|\tilde{Q} - Q^*\right\|_\infty + \gamma\left\|V^* - V^{\pi_{\tilde{Q}}}\right\|_\infty. \tag{35}$$

重新整理该不等式，即可得到引理

$$\left\|V^* - V^{\pi_{\tilde{Q}}}\right\|_\infty \le \frac{2}{1-\gamma}\left\|\tilde{Q} - Q^*\right\|_\infty. \tag{36}$$