arXiv:2301.00927v1 [stat.ML] 3 Jan 2023

深度谱 Q 学习及其在移动健康中的应用

Yuhe Gao*[1]、Chengchun Shi†[2]和Rui Song‡[1]

[1]北卡罗来纳州立大学统计系，美国罗利[2]伦敦政治经济学院统计系，英国伦敦

2023年1月4日

摘要

动态治疗方案根据患者的基线信息和时变协变量，随时间依次为患者分配个体化治疗。在移动健康应用中，这些协变量通常在较长的时间跨度内以不同频率收集。本文提出一种深度谱 Q 学习算法，将主成分分析（PCA）与深度 Q 学习相结合，以处理混合频率数据。理论上，我们证明了估计得到的最优策略下的平均回报收敛于最优策略下的平均回报，并建立了其收敛速率。我们还通过模拟和对糖尿病数据集的应用，进一步展示了该方法的实用性。

关键词： 动态治疗方案、混合频率数据、主成分分析、强化学习

*ygao32@ncsu.edu
†c.shi7@lse.ac.uk
‡rsong@ncsu.edu

1 引言

精准医学着重于考虑患者的个人信息，为其提供个体化治疗（例如，参见 Kosorok 和 Laber，2019年；Tsiatis 等，2019年）。相关研究已在众多领域发现了精准医学的多种应用，从心血管疾病研究到癌症治疗和基因治疗（Jameson 和 Longo，2015年）。动态治疗方案（DTR）由一系列治疗决策规则构成，这些规则针对每位患者在各个时点的状态量身定制，从数学上阐释了精准医学背后的理念。精准医学的主要目标之一，是确定能够带来平均而言最有利结局的最优动态治疗方案。

随着移动健康（mHealth）技术的快速发展，在医学研究中通过移动应用收集丰富的纵向数据已成为可能。OhioT1DM 数据集（Marling 和 Bunescu，2020年）提供了一个具有启发性的数据实例，其中包含 12 名 1 型糖尿病患者在 8 周内通过健身手环测量获得的数据。基于这些数据估计的数据驱动决策规则有望改善这些患者的健康状况（例如参见 Shi 等人，2020年；Zhu 等，2020；Zhou 等人，2022年）。然而，在这类移动健康研究中估计最优动态治疗方案仍然具有挑战性。首先，治疗阶段的数量（例如时间范围）不再固定，而患者数量可能十分有限。例如，在 OhioT1DM 数据集中，研究仅纳入了 12 名患者。尽管如此，假设治疗决策按小时做出，则时间范围超过 1000 个时段。现有动态治疗方案文献中的方法（Murphy，2003年；Zhang 等，2013年；Zhao 等，2015年；Song 等，2015年；Shi 等，2018年；Nie 等人，2021年；Fang 等，2022年；Qi 和 Liu，2018年；Mo 等，2021年；Ertefaie 等人，2021年；Guan 等，2020年；Zhang 等，2018年）在这类长时域或无限时域情形下会变得低效，并且需要大量患者才能保证一致性。其次，患者的时变协变量通常包含混合

频率数据。在 OhioT1DM 数据集中，一些变量（例如连续血糖监测（CGM）记录的血糖水平）每 5 分钟记录一次。同时，其他变量（例如餐食的碳水化合物估计量和运动强度）的记录频率则低得多。将每个一小时时间间隔内的这些高频变量进行拼接，会产生高维状态向量，而直接将这些状态作为治疗策略的输入，会导致决策规则噪声很大。一种朴素的方法是使用一些临时汇总指标对高频数据进行概括，以进行策略学习。然而，由于信息损失，这可能会产生次优策略。

近年来，统计学文献中关于无限时域策略学习和/或评估的研究日益增多。相关文献包括 Ertefaie 和 Strawderman（2018年）；Luckett 等人（2020年）；Shi 等人（2022年）；Liao 等人（2020年、2021年）；Shi 等人（2021c）；Chen 等人（2022年）；Ramprasad 等人（2022年）；Li等人（2022年）；Xu 等人（2020年）。计算机科学文献中则有大量关于开发无限时域强化学习（RL）算法的研究。这些算法可分为无模型算法和基于模型的算法。常见的无模型强化学习方法包括基于价值的方法：对从给定状态（或状态-动作对）出发的期望回报进行建模，并根据价值函数将最优策略计算为贪心策略（Ernst 等人，2005年；Riedmiller，2005年；Mnih 等人，2015年；Van Hasselt 等人，2016年；Dabney 等人，2018年）；以及基于策略的方法：通过策略梯度或动作评论家方法，在某个参数化类别中直接搜索最优策略（Schulman 等人，2015a,b；Koutn´ık 等人，2013年；Mnih 等人，2016年；Wang 等人，2017年b）。基于模型的算法与无模型算法的区别在于，它们会对环境的转移动力学建模，并利用环境模型推导或改进策略。常见的基于模型的强化学习方法包括 Guestrin 等人（2002年）、Janner 等人（2019年）、Lai 等人（2020年）、Li 等人（2020年）等。另请参阅 Arulkumaran 等人（2017年）；Sutton 和

Barto（2018年）；Luo 等人（2022年），了解更多详情。这些方法无法直接应用于 OhioT1DM 等数据集，因为它们尚未考虑混合频率数据。

在强化学习文献中，已有少数研究考虑通过降维来处理高维状态系统。具体而言，Murao 和 Kitamura（1997年）提出对状态空间进行分段，并学习状态的聚类表示。Whiteson 等人（2007年）提出将状态空间划分为铺砌，以表示各个状态。这两篇论文都提出通过离散化状态空间来实现降维。然而，这可能导致显著的信息损失（Wang 等人，2017年a）。Sprague（2007年）提出了一种使用邻域成分分析的迭代降维方法。该方法使用线性基函数对 Q 函数建模，无法支持更一般的非线性函数逼近。近期，有少数研究在强化学习中采用主成分分析（主成分分析）进行降维（Curran 等人，2015年、2016年；Parisi 等人，2017年）。然而，上述论文均未正式确立其方法的理论保证。此外，这些方法的研究动机源于游戏或机器人学中的应用，其推广到具有混合频率数据的移动健康应用是否可行，仍不得而知。

在动态治疗方案的文献中，少数研究考察了同时包含标量和函数型协变量的混合频率数据。具体而言，McKeague 和 Qian（2014年）提出了一种函数型回归模型，用于基于一个函数型协变量进行最优决策。Ciarleglio 等人（2015年）和 Ciarleglio 等人（2016年）将其方法扩展到包含多个标量和函数型协变量的更一般情形。Ciarleglio 等人（2018年）研究了如何通过变量选择处理混合频率数据。Laber 和 Staicu（2018年）将函数型主成分分析应用于函数型协变量，以实现降维。上述研究均针对单阶段决策。它们的方法无法直接应用于无限时域情形。

我们的贡献如下。在科学层面，混合频率数据常见于移动健康应用中。然而，这类数据在无限时域情形中尚未得到充分研究。因此，我们的方法填补了一个重要空白，并大幅拓展了现有动态治疗方案学习方法的适用范围。在方法层面，我们提出了一种用于降维的深度谱 Q 学习算法。与直接使用原始混合频率数据或其临时汇总结果作为治疗策略输入的方法相比，所提出的算法在经验表现上更优。在理论层面，我们推导了所提出策略的遗憾上界，并将该界分解为近似误差与估计误差之和。我们的理论为实践者选择所提出算法中的主成分数量提供了一些通用指南。

本文其余部分安排如下。我们将在第 2 节介绍动态治疗方案（DTR）和混合频率数据的相关背景。在第 3 节中，我们将介绍用于估计最优动态治疗方案的方法，并在第 4 节中研究其理论性质。第 5 节将开展实证研究。在第 5.3 节中，我们将把所提出的方法应用于 OhioT1DM 数据集。最后，我们将在第 6 节总结全文。

2 预备知识

2.1 数据与符号

假设该研究招募了 N 名患者。第 i名患者的数据集可概括为 $O_i \equiv \{(S_{i,t}, A_{i,t}, Y_{i,t}) : 1 \leq t \leq T_i\}$。为简便起见，我们假设对任意 i，$T_i = T$。每个状态$S_{i,t} = (X_{i,t}, \{Z_{i,t,j}\}_{j=1}^{J}) \in \mathcal{S}$ 由一组低频协变量 $X_{i,t} \in \mathbb{R}^{m_0}$ 和一组 J 高频协变量 $\{Z_{i,t,j}\}_{j=1}^{J}$ 构成，其中 $\mathcal{S}$ 为状态空间。对于每个高频变量$Z_{i,t,j}, j \in \{1, 2, .., J\}$，我们有$Z_{i,t,j} = (Z_{i\,t}^{(1)}, , _{,j}, Z_{i\,t}^{(2)}, , _{,j}, ..., Z^{(m_j)}$ $_{i,t,j})^T \in \mathbb{R}^{m_j}$

其中 m_j为某个较大的整数。令 τ 表示一个时间单位的长度。低频变量在时间点 $\tau, 2\tau, \cdots, t\tau, \cdots$ 记录。然而，第 j个高频变量的记录频率更高，记录时间点为$m_j^{-1}\tau, 2m_j^{-1}\tau, \cdots$。请注意，我们允许 J 高频变量以不同的频率记录。令 $m = \sum_{j=1}^{J} m_j$和Z_i^T，$_t = [Z_{i,t,1}^T, Z_{i,t,2}^T, \ldots, Z_{i,t,J}^T] \in \mathbb{R}^m$表示一个高维变量，该变量拼接了所有高频协变量。因此，状态$S_{i,t}$ 可表示为$(X_{i,t}, Z_{i,t}) \in \mathbb{R}^{m_0+m}$。此外，$A_{i,t}$ 表示第 t个时间点的治疗指标，$\mathcal{A}$ 表示所有可能治疗选项的集合。奖励$Y_{i,t}$ 对应于第i位患者在第 t个决策点之后获得的响应。按照惯例，我们假设$Y_{i,t}$ 的值越大表示结局越好。我们要求$|Y_{i,t}|$ 由某个常数 $R_{max} > 0$一致有界，并假设 $O_1, O_2, ..., O_N$为 $i.i.d.$；这些条件在强化学习文献中很常见（例如参见 Sutton 和 Barto，2018年）。最后，我们将函数关于给定分布函数 σ 聚合后的 l_p-范数记为$\|\cdot\|_{p,\sigma}$。对于任意整数 $q \in \mathbb{N}$，我们用 $[q]$ 表示指标集 $\{1, 2, 3, ...q\}$。

2.2 假设、策略与价值函数

我们要求系统满足马尔可夫假设，即

$$P(S_{0,t+1} \in \mathbb{S} | S_{0,t} = s, A_{0,t} = a, \{S_{0,t'}, A_{0,t'}\}_{0 \le t' < t}) = P(S_{0,t+1} \in \mathbb{S} | S_{0,t} = s, A_{0,t} = a) \ \forall t,$$

对任意 s、a 以及博雷尔集 $\mathbb{S} \in \mathcal{S}$ 均成立。换言之，下一状态的分布仅通过当前状态-动作对依赖于过去的数据历史。我们假设转移核相对于某个一致有界的转移密度函数 $q(s'|s, a)$ 绝对连续，使得对于某个常数 $c_q > 0$，$\sup_{s,a,s'} |q(s'|s, a)| \le c_q$。

此外，我们还施加以下条件均值独立假设：

$$E(Y_{0,t} | S_{0,t} = s, A_{0,t} = a, \{Y_{0,t'}, S_{0,t'}, A_{0,t'}\}_{0 \le t' < t}) = E(Y_{0,t} | S_{0,t} = s, A_{0,t} = a) = r(s, a) \ \forall t,$$

w这里我们将 $r(s,a) = r(x, \{z_j\}_{j=1}^J, a)$ 称为即时奖励函数 .

接下来，将策略 $\pi : \mathcal{S} \to P(\mathcal{A})$ 定义为一个函数，它将患者在每个时间点的状态映射为动作空间上的概率分布函数。给定 π，我们将其（状态）价值函数定义为

$$V^{\pi}(s) = \sum_{t=0}^{\infty} \gamma^t E^{\pi}\{Y_{0,t} | S_{0,0} = s\},$$

其中 $\gamma \in [0,1)$ 是一个折扣因子，用于平衡即时奖励和长期奖励。根据定义，状态价值函数刻画了在决策过程遵循给定目标策略 π 的情况下的期望回报。此外，我们将状态-动作价值函数（或 Q 函数）定义为

$$Q^{\pi}(s,a) = \sum_{t=0}^{\infty} \gamma^t E^{\pi}\{Y_{0,t} | S_{0,0} = s, A_{0,0} = a\},$$

它是 给定初始状态-动作对时的期望折扣累积奖励 .

在马尔可夫假设和条件均值独立假设下，存在一个最优策略 π^*，使得 $V^{\pi^*}(s) \geq V^{\pi}(s)$ $\forall \pi$， $s \in \mathcal{S}$（例如，参见 Puterman，2014年）。此外， π^* 满足以下贝尔曼最优性方程:

$$Q^{\pi^*}(s,a) = E\{Y_{0,t} + \gamma \max_{a' \in \mathcal{A}} Q^{\pi^*}(S_{0,t+1}, a') | S_{0,t} = s, A_{0,t} = a\}, \tag{1}$$

其中 Q^{π^*} 表示最优 Q 函数。

2.3 ReLU 网络

本文采用基于价值的方法，通过估计最优 Q 函数来学习最优策略 π^*。我们将使用带有修正线性单元（ReLU）激活函数的稀疏神经网络，即 $f(x) = \max(x, 0)$，对 Q 函数进行建模。

与简单的参数模型相比，使用神经网络的优势在于，神经网络能够更好地捕捉高维状态系统中潜在的非线性。

形式上，稀疏 ReLU 网络类定义如下：

$$\mathcal{F}_{SReLU}(L,\{d_j\}_{j=0}^{L+1},s,V_{max})=\{\mathbb{R}^{d_0}\to\mathbb{R}^{d_{L+1}}:f(x)=W_Lg_{L-1}\circ g_{q-1}\circ ...g_1\circ g_0(x),$$
$$g_j(x)=\sigma(W_jx+v_j),W_j\in\mathbb{R}^{d_{j+1}\times d_j},v_j\in\mathbb{R}^{d_{j+1}},j\in\{0,1,...,L\},$$
$$\max_{j=0,1,...,L}\{\max(\|W_j\|_\infty,|v_j|_\infty)\}\leq 1,\sum_{j=0}^{L}(\|W_j\|_0+|v_j|_0)\leq s,\max_{k\in\{1,2,...,d_{L+1}\}}\|f_k\|_\infty\leq V_{max}\}.$$

这里， L 表示神经网络的隐藏层数， d_j 表示每层的宽度。由于 Q 函数的输出是标量，输出维度 d_{L+1} 设为 1。 $\mathcal{F}_{ReLU}(L)$ 中的参数包括权重矩阵 W_j 和偏置向量 v_j。稀疏度 s 为模型中非零参数的总数提供上界。在实现中，可以使用 Dropout 层满足这一约束（Srivastava 等人，2014年）。理论上，稀疏 ReLU 网络能够以极小极大最优收敛速率拟合光滑函数（Schmidt-Hieber，2020年）。第 4 节中的主要定理将依赖这一性质。图 1 展示了稀疏 ReLU 网络的示意图。

3 谱拟合 Q 迭代

2.3 节中的 ReLU 激活函数神经网络常用于基于价值的强化学习算法。然而，在医学研究中，训练数据集的规模往往有限，总观测值只有几千或一万左右（例如，Marling 和 Bunescu，2020年；Liao 等人，2021年）。与此同时，数据包含高频状态变量，因此形成了高维状态系统。直接将这些状态用作输入会产生非常嘈杂的策略。这促使我们考虑在强化学习中进行降维。

一种简单的降维方法是将高频状态的一些汇总统计量用作策略学习的输入。例如，在 OhioT1DM 数据集上，可以将两个治疗决策点之间连续血糖监测血糖水平的平均值用作汇总统计量，Zhu 等人（2020年）、Shi 等人（2022年）和 Zhou 等人（2022年）均采用了这种做法。本文提出使用主成分分析来降低$\{Z_{i,t,j}\}_{j=1}^{J}$的维度。我们预计，与一些临时选取的汇总量（例如，平均值）相比，使用主成分分析能够保留更多信息。

要在无限时域设定中应用主成分分析，我们需要对拼接后的高维变量$Z_{i,\ t}^{T} \in \mathbb{R}^{m}$施加一些平稳性假设： $E[Z_{i,t}] = \mu$ 和$Cov[Z_{i,t}] =$**G** 满足某个均值向量 $\mu \in \mathbb{R}^{m}$ 和协方差矩阵 **G** $\in \mathbb{R}^{m\times m}$ ，且它们与 t相互独立。在真实数据应用中，我们可以检验拼接后的高频变量$Z_{i,\ t}$是否弱平稳（例如，参见 Dickey 和 Fuller，1979；Said 和 Dickey，1984；Kwiatkowski 等人，1992）。如果它弱平稳，拼接后的高频协变量$Z_{i,\ t}$将自动满足上述两个假设。文献中也广泛采用了类似假设（例如，参见 Shi 等人，2021b；Kallus 和 Uehara，2022）。不失一般性，为简化记号，我们假设 $\mu = 0_{m}$。协方差矩阵**G**通常是未知的。实践中，我们建议使用样本协方差估计量**G**。

ˆ 我们的步骤如下所述。根据谱分解，有 **G** $=\sum_{k=1}^{m}$ ˆλ_{k} ˆU_{k} ˆU_{k}^{T}，其中 $\lambda_{1} \geq \lambda_{2} \geq ... \geq 0$ 为特征值，ˆ U_{k}为对应的特征向量。因此，我们可以将 $Z_{i,t}$ 表示为 $\sum_{k=1}^{m}$ ˆ$\lambda_{k}^{1/2}$ ˆ$V_{(i,t)\ k}$ ˆU_{k}，其中 ˆ$V_{(i,t)\ k}$ 是估计的主成分得分，由 ˆ$\lambda_{k}^{-1/2} Z_{i}^{T}$,ˆ$U_{k}$给出。对于任意 κ，估计的主成分得分 ˆ$\mathbf{V}_{i,t,\kappa} = (\ \hat{}V_{(i,t)\ 1}\ ,\ \hat{}V_{(i,t)\ 2}\ , ...,\ \hat{}V_{(i,t)_{\kappa}})$ 对应于拼接后的高频变量 $Z_{i,t}$的 $\kappa \leq m$ 个最大特征值。当 $\kappa = m$时，使用这些主成分得分等同于使用原始高频变量 $Z_{i,t}$。

ˆ 我们将用 $Q^{\pi}(X_{i,\ t},\ \mathbf{V}_{i,\ t,\ \kappa},\ A_{i,\ t})$ 近似 $Q^{\pi}(X_{i,\ t},\ \mathbf{V}_{i,\ t,\ m},\ A_{i,\ t})$，并提议使用 Riedmiller（2005年）的神经拟合Q迭代算法来学习估计的最优策略。我们在算法 1 中详细说明这一过程。

算法 1 谱拟合Q迭代

输入： $\{S_{i,t} = (X_{i,t}, Z_{i,t}) A_{i,t}, Y_{i,t}, \gamma\}$，其中 $i \in \{1, 2, ..., N\}, t \in \{1, 2, ..., T\}$；ReLU 网络函数类 $\mathcal{F}_{SReLU} = \mathcal{F}_{SReLU}(L, \{d_j\}_{j=0}^{L+1}, s, V_{max})$；抽样分布 σ；样本量 n；主成分数量 κ；迭代次数 K；估计的ˆ 协方差矩阵 $\mathbf{G}$，针对 $Z_{i,t}$；

计算前 κ 个主成分分析 $\hat{\mathbf{V}}_{i,t,\kappa}$ 用于高频数据部分 $Z_{i,t}$；

˜对于每个 $a \in \mathcal{A}$，初始化一个稀疏 ReLU 网络 $Q_0(x, v_\kappa, a) \in \mathcal{F}_{SReLU}$；

对于 $k = 1$ **到** K，**执行**

采样 n 观测值 $I_k = \{(i,\ t):\ 1 \le i \le N,\ 1 \le t \le T-1\}$ 基于来自数据的 σ；˜ ˜ ˜ ˆ 定义响应 $R_{i,\ t}$ 基于 Q_k：$R_{i,\ t}(Q_k) = Y_{i,\ t} + \gamma \max_{a \in \mathcal{A}} Q_k(X_{i,\ t+1},\ \mathbf{V}_{i,\ t+1,\ \kappa},\ a)$；˜ ˜将 Q_k 更新为 Q_{k+1}：

$$\tilde{Q}_{k+1} \leftarrow \underset{f(\cdot,a) \in \mathcal{F}_{SReLU}}{\operatorname{argmin}} \frac{1}{n} \sum_{(i,t) \in I_j} [R_{i,t}(\tilde{Q}_k) - f(X_{i,t}, \hat{\mathbf{V}}_{i,t,\kappa}, A_{i,t})]^2$$

结束循环

˜**返回** 贪心策略：$\pi_K(a|x,\ v_\kappa) = 0$，如果 $a \notin argmax_{a'} Q_K(x,\ v_\kappa,\ a')\ \forall x,\ v_\kappa,\ a$

在算法 1 中，我们为 $Q(s,\ a)$ 中的每个 a 拟合一个 $|\mathcal{A}|$ 神经网络。在动作空间较小的情况下，这样做是合理的。当数据集较小时（例如 OhioT1DM 数据集），我们建议将 n 设为 $N(T-1)$，这样每次迭代都会使用所有数据记录（而不是随机子样本）。

与 Riedmiller（2005年）提出的原始神经拟合Q迭代算法类似，该算法的直觉也基于贝尔曼最优性方程（1）。在每一步 k 中，

算法 1 中， $\tilde{Q}_k$ 估计 Q^{π^*} 和响应 R_i，$_t$（$\tilde{Q}_k$）$= Y_i$，$_t + \gamma \max_{a\in\mathcal{A}} \tilde{Q}_k$（$X_i$，$_{t+1}$，ˆ$\mathbf{V}_i$，$_{t+1}$，$_\kappa$，$a$）对应于方程（1）的右侧。因此，利用 Q_{k+1} 对$\tilde{}R_i$，$_t$ 进行回归拟合，就是求解贝尔曼最优性方程。算法 1 与原始神经拟合Q迭代算法的关键区别在于，高维输入 Z_i^T，$_t = [Z_{i,t,1}^T, Z_{i,t,2}^T, \ldots, Z_{i,t,J}^T]$ 被纳入状态空间，并在学习过程中映射为一个较低维度的ˆ ˜维向量 $\mathbf{V}_i$，$_t$，$_\kappa$，因此神经网络 Q_0（x，v_κ，a）ˆ以主成分 $\mathbf{V}_i$，$_t$，$_\kappa$ 而非原始高维 Z_i，$_t$ 作为输入。

4 渐近性质

在讨论我们提出的 Q-learning 方法的渐近性质之前，先介绍一些记号。

定义4.1。定义

$\mathcal{F}_0(L, \{d_i\}_{i=0}^{L+1}, s, V_{max}) = \{f\colon \mathcal{S}\times\mathcal{A}\to\mathbb{R}, f(.,a) \in \mathcal{F}_{SReLU}(L, \{d_i\}_{i=0}^{L+1}, s, V_{max})$ $\forall a\in\mathcal{A}\}$,其中 $\mathcal{F}_{SReLU}(L, \{d_j\}_{i=0}^{L+1}, s, V_{max})$ 表示具有 L 层、稀疏度参数为 s 的稀疏 ReLU 网络，$V_{max} = \frac{R_{max}}{1-\gamma}$ 表示累积奖励的统一上界。

——

我们将 $L_{\mathcal{F}_0} = \sup^{f\in\mathcal{F}_0}\sup_{x\neq y} \frac{|f_{(y=f\ x\ |^2)}()}{\|y-x\|^2}$定义为 $\mathcal{F}_0$ 中所用稀疏 ReLU 类 $\mathcal{F}_{SReLU}($ $L, \{d_j\}_{i=0}^{L+1}, s, V_{max})$ 的 Lipschitz 常数。注意，该函数类 $\mathcal{F}_0$ 用于原始神经网络拟合 Q 迭代算法中对 Q 函数进行建模，其中状态中的高频部分 Z 的维度不会通过主成分分析进行降维。我们进一步定义函数类 $\mathcal{F}_2$，它先将高频部分 Z 转换为其主成分得分，再使用稀疏 ReLU 神经网络得到相应的 Q 函数。具体而言，$\mathcal{F}_2$ 是由函数 $\{f_2\}$组成的集合，使得 $f_2(x, z, a) = f_0(x, \hat{}v_\kappa, a)$,

其中， $f_0 \in \mathcal{F}_0$ 和 $\hat{}v_\kappa$ 是包含 Z的前 κ 个主成分的向量。注意， $\mathcal{F}_2$ 是我们在算法 1 中用于对 Q 函数进行建模的函数类。也就是说， $Q_k \in \mathcal{F}_2, k \in \{1, 2, ..., K\}$ （ $\mathcal{F}_2$ 的正式定义以及此处未提及的另一个函数类 $\mathcal{F}_1$ 见附录的 A 节）。

此外，我们通过 $\mathcal{C}_r(\mathcal{D}, \beta, H)$ 引入 H¨older 光滑函数类，其中 $\mathcal{D} \in \mathbb{R}^r$表示函数输入集合。定义如下：

定义 4.2。定义

$$\mathcal{C}_r(\mathcal{D}, \beta, H) = \{f : \mathcal{D} \to \mathbb{R} : \sum_{\gamma<\beta} \|\partial^\gamma f\|_\infty + \sum_{\alpha\in\mathbb{N}_0^r:\|\alpha\|=\lfloor\beta\rfloor} \sup_{x,y\in\mathcal{D},x\neq y} \frac{|\partial^\alpha f(x) - \partial^\alpha f(y)|}{\|x-y\|_\infty^{\beta-\lfloor\beta\rfloor}} \leq H\},$$

其中， $\alpha \in \mathbb{N}_0^r$ 是用于表示偏导数的 r 元多重指标。

接下来，我们构造一个网络结构 $\mathcal{G}(\{p_j, t_j, \beta_j, H_j\}_{j\in[q]})$，其中该网络每一层的分量函数都属于 Hölder 光滑函数类 $\mathcal{C}(\mathcal{D}, \beta, \mathcal{H})$，称为 Hölder 光滑函数的复合。该复合网络包含 q 层，每一层为 g_j： $[a_j, b_j]^{p_j} \to [a_{j+1}, b_{j+1}]^{p_{j+1}}$，使得 g_{jk}第 k 个分量（ $k \in [p_{j+1}]$）在第 j 层满足 $g_{jk} \in \mathcal{C}_{t_j}([a_j, b_j]^{t_j}, \beta_j, H_j)$，其中有 $t_j \leq p_j$ 个输入。类似于 $\mathcal{F}_0$ 的定义，我们可以在 $\mathcal{S}\times\mathcal{A} \to \mathbb{R}$ 上定义函数类 $\mathcal{G}_0(\{p_j, t_j, \beta_j, H_j\}_{j\in[q]})$（简记为 $\mathcal{G}_0$），使得每个函数 $g \in \mathcal{G}_0$ 都满足 $g(\cdot, a) \in \mathcal{G}(\{p_j, t_j, \beta_j, H_j\}_{j\in[q]})$，其中 $\forall a \in \mathcal{A}$。函数类 $\mathcal{G}_0$ 与网络结构 $\mathcal{G}$ 之间的关系，类似于定义 4.1 中函数类 $\mathcal{F}_0$ 与神经网络 $\mathcal{F}_{SReLU}$ 之间的关系。关于 $\mathcal{G}_0(\{p_j, t_j, \beta_j, H_j\}_{j\in[q]})$ 的更多细节，请参阅 Fan 等人（2020）的定义 4.1。

接下来，我们将介绍定理所需的三个主要假设：

假设4.3。 $Cov(Z)$ 的特征值遵循指数衰减趋势 $\lambda_k = O(e^{-\zeta k})$ $k = 1, 2, ..., m$，其中 $\zeta > 0$ 为某个常数。

假设 4.4。 估计量 $\hat{\mathbf{G}} = \sum_{k=1}^{m} \hat{\lambda}_k \hat{U}_k \hat{U}_k^T$ 满足 $\left\| \hat{U}_k - U_k \right\|_2 = O_p(n^{-\Delta})$，其中 $1 \leq k \leq m$ 使得 $\Delta > 0$ 为某个常数。

假设 4.5。 首先，我们定义贝尔曼最优算子 $\mathcal{T}$ 如下

$$\mathcal{T} f(x, z, a) = E\{Y_{0,t} + \gamma \max_{a' \in \mathcal{A}} f(X_{0,t+1}, Z_{0,t+1}, a') | A_{0,t} = a, X_{0,t} = x, Z_{0,t} = z\}.$$

然后，对于 $f \in \mathcal{F}_2$，我们假设 $\mathcal{T} f \in \mathcal{G}_0$。

在这三个假设中，假设 4.3 中特征值的指数衰减结构在高维数据分析和函数型数据分析文献中很常见（例如，Reiß 和 Wahl，2020年；Crambes 和 Mas，2013年；Jirak，2016年）。该假设旨在控制仅使用 $Z_{i,t}$ 的前 κ 个主成分得分 $\hat{}$ 所造成的信息损失。假设 4.4 关于估计量 $\mathbf{G}$ 的一致性，函数型数据分析文献中也采用了类似的假设（例如，Laber 和 Staicu，2018年；Staicu 等人，2014年）。通过采用证明 Zhang 和 Wang（2016年）定理 5.2 时的类似论证，我们可以证明该假设在我们的设定中同样成立。该假设用于界定协方差矩阵估计所导致的误差。假设 4.5 在文献中称为完备性假设（例如，Chen 和 Jiang，2019年；Uehara 等人，2022年、2021年）。当转移核和奖励函数满足某些平滑性条件时，该假设会自动成立。

~我们的第一个定理关注算法 1 中 Q_K 的非渐近收敛速率。

定理 4.6（估计的 Q 函数的收敛性）。 设 μ 为 $\mathcal{S}$ 上的某个分布，使得 $d\mu(s)\,ds$ 有正下界。在假设 *4.3* 至 *4.5* 下，当 n 充分大时，存在一个用于函数类 $\mathcal{F}_2$ 建模 ~ ~$Q(s, a)$ 的稀疏 *ReLU* 网络结构，使得通过我们的算法 *1* 得到的 Q_K 满足：

$$\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}}\left\|Q^{\pi^*}(.,a)-\tilde{Q}_K(.,a)\right\|_{2,\mu}^2 = O_p(|\mathcal{A}|(n^{\alpha^*}\log^{\xi^*}n + d_1^*\kappa)n^{-1}\log^{\xi^*+1}n$$
$$+L_{\mathcal{F}_0}(e^{-\zeta\kappa}-e^{-\zeta m}+n^{-2\Delta})+\frac{\gamma^{2K}}{(1-\gamma)^2}R_{max}^2),$$

其中，$\xi^*>1$、$0<\alpha^*<1$ 是一些常数，$|\mathcal{A}|$ 是治疗选项的数量，d_1^* 是用于 $\mathcal{F}_2$ 的稀疏 *ReLU* 网络第一层的宽度，且满足界 $m_0+m\le d_1^*\le n^{\alpha^*}$。

关于定理 4.6 的神经网络结构和样本量假设，更多细节请参见附录 B。定理 4.6 给出了˜估计的 Q 函数 Q_K的误差界。基于该定理，我们进一步建立了通过算法 1 获得的估计策略 π_K 的遗憾界。为此，我们还需要一个假设：

假设 4.7。 假设存在 $\eta>0,\delta_0>0$，使得

$$P(s:\max_a Q^{\pi^*}(s,a)-\max_{a\in\mathcal{A}-\arg\max_{a'}Q^{\pi^*}(x,a')}Q^{\pi^*}(s,a)\le\epsilon)=O(\epsilon^\eta)$$

对于 $0<\epsilon\le\delta_0$。

边际类型条件假设 4.7 在文献中被广泛采用。具体而言，在分类中，人们会施加边际条件来界定超额风险（Tsybakov，2004年；Audibert 和 Tsybakov，2007年）。在动态治疗方案中，也有类似的假设，用于证明有限时域设定下状态价值函数的收敛（Qian 和 Murphy，2011；Luedtke 和 van der Laan，2016年）。在强化学习中，Shi 等人（2022年）引入了这些假设，以获得估计最优策略的更紧遗憾界。

定理 4.8 (状态价值函数的收敛)。 在假设 *4.3* 至 *4.7* 以及定理 *4.6* 中的条件 $\mu, \mathcal{F}_2, n$ 下，我们有：

$$E_\mu[V^{\pi^*}(s) - V^{\pi_K}(s)] = O_p(\frac{1}{1-\gamma}\{|\mathcal{A}|(n^{\alpha^*}\log^{\xi^*} n + d_1^*\kappa)n^{-1}\log^{\xi^*+1} n + L_{\mathcal{F}_0}(e^{-\zeta\kappa} - e^{-\zeta m} + n^{-2\Delta}) + \frac{\gamma^{2K}}{(1-\gamma)^2}R_{max}^2\}^{\frac{\eta+1}{\eta+2}}).$$

这两个定理的证明见附录 C 节。下面总结我们的理论发现。首先，我们注意到，定理 4.8 中遗憾的收敛速度快于定理 4.6 中估计的 Q 函数的收敛速度。这是因为边际类型假设 4.7 使我们能够得到更精确的误差界。文献中已有类似结果。例如，见 Audibert 和 Tsybakov（2007）的定理 3.3、Qian 和 Murphy（2011）的定理 3.1，以及 Shi 等人（2022）的定理 3 和 4。

从定理 4.8 可以看出，遗憾界主要由三个参数决定：样本量 n、主成分数量 κ 和迭代次数 K。右侧的第一项 $|\mathcal{A}|(n^{\alpha^*}(\log^{\xi^*} n) + d_1^*\kappa)n^{-1}\log^{\xi^*+1} n$ 对应估计误差，它随 n 增加而减小，随 κ增加而增大。第二项 $L_{\mathcal{F}_0}(e^{-\zeta\kappa} - e^{-\zeta m} + n^{-2\Delta})$对应近似误差，它随 n 和 κ 增加而减小。剩余项 $\gamma^{K+1}{}_{(1-\gamma)_2}R_{max}$ 是优化误差，随着算法 1 中的迭代次数 K 增加而减小。

与现有关于深度拟合 Q 迭代算法收敛速率的结果（见 Fan 等人，2020年的定理 4.4）相比，我们的定理还刻画了对主成分数量的依赖关系。具体而言，选择前 κ个主成分会导致信息损失（例如偏差），其量级为 $e^{-\zeta\kappa} - e^{-\zeta m}$，但可将高频变量导致的模型复杂度从 $d_1^* m$ 降至 $d_1^*\kappa$ ，并且

因此降低策略估计量的方差。这体现了偏差-方差权衡。请注意，偏差按指数速率衰减；当训练数据较少时，降低模型复杂度可能更有益。因此，正如我们的数值研究所示，我们的算法有望在小样本情况下优于原始的拟合Q迭代算法。

最后，主成分的数量应随 n 发散，以确保所提算法的一致性。根据这两个定理，在偏差与方差之间取得平衡的最优 κ^* 应满足 $\kappa^* \asymp \log(n)$（详情见附录 D 节）。因此，当 n 趋于无穷时，我们最终将取 $\kappa^* = m$ ，此时算法 1 将等同于原始的神经网络拟合 Q 迭代。这只是一项用于选择主成分数量的渐近指南。下一节将提供一些实用指南。

5 实证研究

5.1 实用 主成分数量指南

在函数型数据分析中，已有多种准则用于选择主成分的数量，包括方差解释率、赤池信息准则（AIC）和贝叶斯信息准则（BIC）（例如，参见 Yao等人，2005年；Li等人，2013年）。在我们的设定中，AIC/BIC 难以应用，因为 Q函数估计不存在自然的目标函数（例如似然）。一种可能的做法是将 Shi 等人（2021a）为单阶段决策制定提出的价值信息准则扩展到无限时域。尽管如此，如何确定惩罚参数以实现一致的调参选择，仍不明确。

在此，我们根据方差解释率选择 κ 。也就是说，我们可以考察使主成分分析的总方差解释率达到某一水平（例如 95%）所需的最小 κ 。Laber 和 Staicu（2018）在单阶段决策制定中也采用了这种方法。为展示该方法的实证表现，我们使用 $\kappa \in \{2, 6, 10, 14..., 74\}$ 应用算法 1，并在以下数值研究中评估这些策略的期望回报。

数据生成过程可描述如下。我们将低频协变量 $X_{i,t}$ 设为一个二维向量，将高频变量 $Z_{i,t}$ 设为一个 108 维向量（$m = 108$）。二者均从零均值正态分布中采样。 Z 的协方差矩阵设定为满足假设4.3。动作空间为二元的，训练数据中用于生成动作的行为策略为均匀随机策略。奖励函数 $r(x, z, a)$ 设为 $= x\beta_{1,a} + z\beta_{2,a} + c \max\{(x\beta_{1,a} + z\beta_{2,a}) \ 0\}$ ，其中 c 为某个常数，$\beta_{1,a}$,$\beta_{2,a}$ 为系数向量，因此该函数是线性函数与具有单层和 ReLU 激活函数的神经网络的混合。下一状态（$X_{i,t+1}$,$Z_{i,t+1}$）将从正态分布中生成，其均值为状态（$X_{i,t}$,$Z_{i,t}$）和动作 $A_{i,t}$的线性函数。轨迹 N 的数量固定为 6，轨迹 T 的长度设为 80。

该 ReLU 网络由 3 个隐藏层组成，宽度为 $d_1 = 15, \ d_2 = d_3 = 5$。在第 1、2 和 3 层之间添加了 Dropout 层，Dropout率为 10%。训练过程中，Dropout 层以 10% 的概率将前一层的输出随机设为 0，从而提高神经网络的稀疏度并减少过拟合（Srivastava 等人，2014年）。神经网络结构的超参数可以通过交叉验证进行调优。折扣因子 γ 固定为 0。5。

为了评估策略表现，我们可以使用蒙特卡洛方法来近似每个估计策略下的期望回报。具体而言，对于每个估计策略 π,

我们生成 $N_{mc} = 100$ 条轨迹，每条长度为 $T_{mc} = 20$ （在我们的设置中，超过 $T_{mc} = 20$ 后的累积奖励可以忽略不计）。初始状态分布与训练数据集中的相同。动作根据 π分配。随后，可以通过计算这100条轨迹的经验累积奖励的平均值来近似期望回报。

对于列表 $\{2, 6, 10, 14..., 74\}$ 中的每个 κ，我们使用算法 1，在 80 个随机种子上学习最优策略，并通过蒙特卡罗方法评估其期望回报。然后，我们计算这 80 个期望回报的样本均值和标准误，以估计策略的价值并构建误差范围。图2展示了这些期望回报的估计值及其置信区间。可以看出，将 κ 从 2 增加到 6 会带来显著提升。然而，进一步增加 κ 会使性能变差。这一趋势与我们的理论一致，因为当 κ 的值较小时，偏差项将主导估计误差。随着 κ 增大，偏差会以指数速度衰减，而模型复杂度项则成为主导项。同时，当 $\kappa \leq 6$ 增大时，方差解释率会迅速上升，并在此后保持稳定。因此，使用这一准则来选择 κ 是合理的。在我们的实现中，我们选择满足方差解释率至少为 95% 的最小 κ 。

5.2 模拟研究

在模拟研究中，我们将所提出的策略 π_K^{PCA}与两种基准策略进行比较：一种是直接使用原始高频变量$Z_{i,\ t,\ j}$ 作为输入（记为 π_K^{ALL}），另一种是使用其均值作为输入（记为 π_K^{AVE}）。这两种策略均以类似方式，基于深度拟合 Q 迭代算法计算得出。此外，我们还纳入了另一种基准策略，记为 π_K^{BOTTLE}：该策略在神经网络架构中原始 Z 的输入层之后添加一个瓶颈层，使该瓶颈层的宽度

与所提出的策略 π_K^{PCA}的输入维度相同。该策略与所提出的策略的区别在于，它使用这个瓶颈层而非主成分分析进行降维。数据生成设定和神经网络结构与第 5.1 节相同。不过，在本模拟研究中，我们改变样本量以及高频变量的维度。具体而言，我们考虑 9 种训练样本量（$N = 6$， $T = 60$，75，90，105，120 和 $N = 7$，8，9，10， $T = 120$），以及 5 种不同的高频部分维度 $m = 27$，54，108，162，216。此外，我们设置了两种生成高频变量的方式，将在下文讨论。我们同样将所提出的策略与 π_K^{ALL}、 π_K^{BOTTLE} 和 π_K^{AVE}进行比较，并使用蒙特卡罗方法评估其价值。

在第一种设定中，我们考虑 5 种情况，其中 J （高频变量的数量）分别等于 1，2，4，6，8，且每个高频变量 Z_j， $j \in \{1, 2, ..., J\}$ 的维度为 27。在此设定中，所有 J 高频变量彼此相关，拼接后的高频变量 Z 的特征值按指数级衰减。我们发现，如图 3 所示，在这 5 种情况下，前 5 个主成分解释的方差均超过 95%。因此，我们将主成分数量设为 $\kappa = 5$ ，并在图 4 中绘制结果。

在第二种设定中， J 高频变量彼此独立。对于每个 j， Z_j 中的所有元素彼此相关，且 Z_j 的特征值与第一种设定下拼接后的高频变量 Z 的特征值以相同的指数级衰减。因此，随着 J 增加，为确保解释的方差比例超过 95%，需要更多主成分。具体而言，当高频变量数量 $J = 1$，2，4，6，8 时，相应的 κ 分别为 5，10，20，30，40。详情见图 3。所有估计得到的最优策略的期望回报绘于图 5。

从图 4 和图 5 可以看出，所提出的策略 π_K^{PCA}始终取得比三种基准策略更大的值。同时， π_K^{ALL}和 π_K^{BOTTLE}的表现

相当。两者的值都受到训练样本量 n的显著影响。此外，在小样本下， π_K^{AVE}优于 π_K^{ALL} 和 π_K^{BOTTLE}，但当样本量较大时，其表现不如这两种策略。在第二种设定中，当 $J=4$、6、8 时， π_K^{ALL}和 π_K^{BOTTLE}往往比 π_K^{AVE}表现好得多，因为对多个高频变量取平均会损失更多与策略学习相关的信息。

最后，我们针对大型训练数据集开展了额外的模拟研究，其中$N=200$ 或 4000，且 $T=120$。这种设置在移动健康数据集中可能并不现实。纳入该设置仅是为了测试 π_K^{ALL} 的性能。由于 π_K^{ALL}也具有一致性，我们预计随着样本量趋于无穷大， π_K^{ALL}下的值与所提出的策略之间的差异将可以忽略不计。图6展示了结果。正如预期，当 $N \geq 200$时，我们观察到 π^{PCA} 与 π^{ALL} 之间没有显著差异。

5.3 OhioT1DMDataset上的应用

我们在 Marling 和 Bunescu（2020）更新后的 OhioT1DM 数据集上应用所提出的算法 1。在真实数据情形下，我们仍希望比较四种策略的行为： π_K^{PCA} 、 π_K^{ALL} 、 π_K^{BOTTLE} 和 π_K^{AVE}。OhioT1DM 数据集包含 12 名 1 型糖尿病患者的医疗信息，包括患者的连续血糖监测血糖水平、这段时间内使用的胰岛素剂量、患者自我报告的膳食和运动信息，以及由手机应用程序和生理传感器记录的其他变量。OhioT1DM 数据集中的高频变量（例如连续血糖监测血糖水平）每 5 分钟记录一次。运动和膳食数据的采集频率低得多，例如每隔几小时记录一次。此外，考虑到该数据集中的基础胰岛素率，尽管这一变量也每 5 分钟采集一次，但它通常会在一天中的数小时内保持不变。因此，取其平均值后，基础胰岛素率也可视为低频标量变量。两个

决策点之间的时间间隔设为 3 小时，因为我们只考虑非紧急情况，此时患者无需立即进行推注。其他使用 OhioT1DM 数据集的研究中，治疗决策频率也远低于连续血糖监测血糖水平的记录频率（例如，Shi 等，2022；Zhou 等，2021；Zhu 等，2020）。

对于低频协变量 $X_{i,\ t} = (X_{i\,t}^{(1)}{}_{,}, \ X_{i\,t}^{(2)}{}_{,})$， $X_{i\,t}^{(1)}{}_{,}$ 是根据过去 3 小时区间内第 i 位患者自我报告的该餐碳水化合物估计量构建的$[t-1, t)$。$X_{i,\ t}$ 中的第二个标量变量定义为过去区间内胰岛素剂量基础率的平均值 $[t-1, t)$。我们考虑一个高频元素 $Z_{i,\ t}$，其中包含过去 3 小时内每 5 分钟记录一次的连续血糖监测血糖水平（其维度为 $m = 36$）。当过去区间内给予第 i 位患者的胰岛素总量大于一个单位时，动作变量 $A_{i,\ t}$ 设为 1。响应变量 $Y_{i,\ t}$ 根据血糖控制指数（IGC，Rodbard，2009）定义，该指数是后续阶段血糖水平的非线性函数。IGC 值越高，表示该患者的血糖水平越接近或处于适宜的血糖范围内。

在本研究中，训练 π_K^{PCA} 时选择 $\kappa = 5$，因为前 5 个主成分解释的方差比例超过 99%，如图 7 所示。这里的 ReLU 网络有 2 个隐藏层，宽度为 $d_i = 6, \ i = 1, \ 2$（在第 1 层、第 2 层与输出层之间加入 Dropout 层，Dropout率为 10%）。为了估计四种策略的价值 $V^{\pi}(x, \ z)$，我们使用 Le 等（2019）提出的拟合 Q 评估算法。应用拟合 Q 评估算法时，使用随机森林模型拟合待评估策略的估计 Q 函数。将 12 位患者分为包含 9 位患者的训练集和包含 3 位患者的测试集后，不同患者组合共进行 220 次重复。在每次重复中，使用 9 位患者的数据训练

策略，并为该策略对应的拟合 Q 评估拟合随机森林。其余 3 位患者的数据则用于通过拟合 Q 评估得到的估计 Q 函数近似该策略的价值。将全部 220 次重复所得估计值的样本均值作为主要结果，并使用标准误构造误差范围。为了比较我们提出的策略 π_K^{PCA} 与 π_K^{ALL}、 π_K^{AVE} 和 π_K^{BOTTLE} 的表现，我们在表 1 中列出 π^{PCA} 与另外三种策略的估计价值之差，其中误差范围为平均差值的标准误乘以临界值 1。96。四种策略的估计价值见表2。

根据该结果可以看出， π_K^{PCA}的估计值高于所有三种基准方法。文献中常用通过连续血糖监测血糖水平的平均值获得的策略 π_K^{AVE}（Zhou 等，2021；Zhu 等，2020）。 π_K^{AVE}表现不太理想，可能是因为简单地用连续血糖监测血糖水平的平均值替代原始数据造成了信息损失。另一方面，训练数据的规模相对较小，因为由于存在大段缺失值，我们没有使用八周内记录的全部数据。最终，我们使用了约连续 5 周的训练数据来训练动态治疗方案。在这种情形下，使用原始高频向量 $Z_{i,\ t}$ 会显著增加 ReLU 网络结构的复杂度，使得待训练参数的数量相对于训练数据规模过大。因此， π_K^{ALL}和 π_K^{BOTTLE}无法优于通过主成分分析降低输入维度的 π_K^{PCA}。表 1 和表2中的结果与第5.2节的结果一致。

6 讨论

总之，我们提出了一种深度谱拟合Q迭代算法，用于处理无限时域情形下的混合频率数据。该算法依靠主成分分析进行降维，并利用深度神经网络捕捉高维系统中的非线性。在理论方面，我们为估计得到的最优策略建立了遗憾界。我们的定理为主成分数量的选择提供了渐近指导。在实证研究中，我们证明了所提算法优于未进行降维或采用状态的临时汇总方法等基准方法。我们还提供了选择主成分数量的实用指南。本文建立在主成分分析和深度 Q 学习的基础上。值得进一步研究其他降维方法和强化学习方法的表现。我们将此留待未来研究。

参考文献

Arulkumaran, K.、Deisenroth, M. P.、Brundage, M. 和 Bharath, A. A.《深度强化学习：简要综述》。*IEEE*信号处理杂志，34(6):26–38（2017年）。

Audibert, J. 和 Tsybakov, A.《插件分类器的快速学习率》。《统计年刊》，35(2):608–633（2007年）。

Chen, E. Y.、Song, R. 和 Jordan, M. I.《基于异质数据的强化学习：估计与推断》。*arXiv* 预印本 *arXiv:2202.00088*（2022年）。

Chen, J. 和 Jiang, N. “批量强化学习中的信息论考量。”国际机器学习会议，1042–1051（2019年）。

Ciarleglio, A.、Petkova, E.、Ogden, R. T. 和 Tarpey, T. “基于标量和函数型基线协变量的治疗决策。”《生物统计学》，71(4):884–894（2015年）。

Ciarleglio, A., Petkova, E., Ogden, T. 和 Tarpey, T. “当治疗效应调节变量未知时，基于标量和函数型预测变量构建治疗决策规则。”《皇家统计学会会刊：*C*辑（应用统计）》，67(5):1331（2018年）。

Ciarleglio, A., Petkova, E., Tarpey, T. 和 Ogden, R. T. “用于估计个体化治疗规则的灵活函数型回归方法。”《*Stat*》（国际统计学会），5(1):185–199（2016年）。

Crambes, C. 和 Mas, A. “具有函数型输出的函数型线性回归中的预测渐近性质。”《伯努利》，2627–2651（2013年）。

Curran, W., Brys, T., Aha, D., Taylor, M. 和 Smart, W. D. “面向辅助机器人的降维强化学习。” 收录于 *2016*年 *AAAI*秋季研讨会系列 （2016年）。

Curran, W.、Brys, T.、Taylor, M. 和 Smart, W.「使用主成分分析高效表示状态空间。」*arXiv* 预印本 *arXiv:1505.00322* （2015年）。

Dabney, W.、Rowland, M.、Bellemare, M. 和 Munos, R.「采用分位数回归的分布式强化学习。」第三十二届 *AAAI* 人工智能大会（2018年）。

Dickey, D. A. 和 Fuller, W. A.：“具有单位根的自回归时间序列的估计量分布。”《美国统计学会会刊》，74(366a):427–431（1979年）。

Ernst, D.、Geurts, P. 和 Wehenkel, L.：“基于树的批处理强化学习。”《机器学习研究杂志》，6:503–556（2005年）。

Ertefaie, A., McKay, J.R., Oslin, D., andStrawderman, R.L. “稳健 Q-learning。”《美国统计学会会刊》, 116(533):368–381 (2021年)。

Ertefaie, A. and Strawderman, R. L. “在无限时间跨度内构建动态治疗方案。” *Biometrika*,105(4):963–977(2018年)。

Fan, J., Wang, Z., Xie, Y., and Yang, Z. “深度 Q 学习的理论分析。”动力学与控制学习, 486–489 (2020年)。

Fang, E. X., Wang, Z., and Wang, L. “面向公平的最优个体化治疗规则学习。”《美国统计学会会刊》, 1–14 (2022年)。

Guan, Q., Reich, B. J., Laber, E. B., and Bandyopadhyay, D. “贝叶斯非参数策略搜索及其在牙周复诊间隔中的应用。”《美国统计学会会刊》, 115(531):1066–1078 (2020年)。

Guestrin, C.、Patrascu, R. 和 Schuurmans, D. “因子化马尔可夫决策过程中的基于模型的强化学习：算法引导的探索。”载于国际机器学习会议，235–242。Citeseer（2002年）。

Jameson, J. L. 和 Longo, D. L. “精准医学——个性化、问题重重且前景广阔。”产科与妇科调查，70(10):612–614（2015年）。

Janner, M., Fu, J., Zhang, M., and Levine, S. “何时该信任你的模型：基于模型的策略优化。”神经信息处理系统进展，32（2019年）。

Jirak, M. “最优特征展开与一致有界。”《概率论及相关领域》，166(3):753–799（2016年）。

Kallus, N. 和 Uehara, M. “利用双重强化学习高效打破离策略评估中的时间范围诅咒。”《运筹学》（ 2022 年）。

Kosorok, M. R. 和 Laber, E. B. “精准医学。”《统计学及其应用年度评论》，6:263–286（2019年）。

Koutn´ık, J.、Cuccu, G.、Schmidhuber, J. 和 Gomez, F. “为基于视觉的强化学习演化大规模神经网络。” 第*15*届遗传与进化计算年度会议论文集，1061–1068（2013年）。

Kwiatkowski, D.、Phillips, P. C.、Schmidt, P. 和 Shin, Y. “检验平稳性的原假设与单位根备择假设：我们有多大把握认为经济时间序列存在单位根？”《计量经济学杂志》，54(1-3):159–178（1992年）。

Laber, E. 和 Staicu, A. “个体化治疗方案的函数型特征构造。”《美国统计学会会刊》，113(523):1219–1227（2018年）。

Lai, H., Shen, J., Zhang, W. 和 Yu, Y. “双向基于模型的策略优化。” 收录于 国际机器学习会议，5618–5627。PMLR（2020年）。

Lazaric, A., Ghavamzadeh, M. 和 Munos, R. “基于分类的策略迭代算法分析。”《机器学习研究杂志》，17:583–612（2016年）。

Le, H., Voloshin, C. 和 Yue, Y. “约束下的批量策略学习。”国际机器学习会议，3703–3712（2019年）。

Li, G., Wei, Y., Chi, Y., Gu, Y. 和 Chen, Y. “利用生成模型打破基于模型的强化学习中的样本量壁垒。”神经信息处理系统进展，33:12861–12872（2020年）。

Li, Y., Wang, C.-h., Cheng, G. 和Sun, W. W.《速率最优上下文在线匹配老虎机》。*arXiv* 预印本 *arXiv:2205.03699* （2022年）。

Li, Y., Wang, N. 和 Carroll, R. J.《函数型数据中主成分数量的选择》。《美国统计学会会刊》，108(504):1284–1294（2013年）。

Liao, P., Klasnja, P. 和 Murphy, S.《长期平均结果的离策略估计及其在移动健康中的应用》。《美国统计学会会刊》，116(533):382–391（2021年）。

Liao, P., Qi, Z., Klasnja, P. 和 Murphy, S.《平均奖励马尔可夫决策过程中的批量策略学习》。*arXiv* 预印本 *arXiv:2007.11771* （2020年）。

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E. 和 Kosorok, M. R.《使用 V 学习估计移动健康中的动态治疗方案》。《美国统计学会会刊》，115(530):692–706（2020年）。PMID: 32952236。URL https://doi.org/10.1080/01621459.2018.1537919

Luedtke, A. R. 和 van der Laan, M. J.《可能不唯一的最优治疗策略下平均结果的统计推断》。《统计年刊》，44(2):713 – 742（2016年）。URL https://doi.org/10.1214/15-AOS1384

Luo, F.-M.、Xu, T.、Lai, H.、Chen, X.-H.、Zhang, W. 和 Yu, Y.《基于模型的强化学习综述》。*arXiv* 预印本 *arXiv:2206.09328*（2022年）。

Marling, C. 和 Bunescu, R. “用于血糖水平预测的 OhioT1DM 数据集：2020年更新。” *CEUR* 研讨会论文集，2675:71–74（2020年）。

McKeague, I. W. 和 Qian, M. “基于函数型预测变量的治疗策略估计。”《中华统计学杂志》，24(3):1461（2014年）。

Mnih, V.、Badia, A.、Mirza, M.、Graves, A.、Lillicrap, T.、Harley, T.、Silver, D. 和 Kavukcuoglu, K. “用于深度强化学习的异步方法。” 国际机器学习会议，1928–1937（2016年）。

Mnih, V.、Kavukcuoglu, K.、Silver, D.、Rusu, A.、Veness, J.、Bellemare, M.、Graves, R.、A.、M.、F.、A.K.、G.、Ostrovski 和 Petersen, S. “通过深度强化学习实现人类水平的控制。”《自然》，518(7540):529–533（2015年）。

Mo, W.、Qi, Z. 和 Liu, Y.《学习最优的分布鲁棒个体化治疗规则》。《美国统计学会会刊》，116(534)：659–674（2021年）。

Murao, H. 和 Kitamura, S.《采用自适应状态分割的 Q-learning（QLASS）》。载于《*1997*年 *IEEE* 国际机器人与自动化计算智能研讨会论文集 *CIRA’97*：面向机器人学与自动化的新计算原理》，179–184（1997年）。

Murphy, S. A.《最优动态治疗方案》。《英国皇家统计学会会刊：*B*辑（统计方法学）》，65(2):331–355（2003年）。

Nie, X.、Brunskill、E. 和 Wager、S.《学习何时治疗策略》。《美国统计学会会刊》，116(533):392–409（2021年）。

Parisi, S., Ramstedt, S. 和 Peters, J. “面向强化学习的目标驱动降维。”载于*2017*年*IEEE/RSJ*智能机器人与系统国际会议（*IROS*），4634–4639。IEEE（2017年）。

Puterman, M. L. 马尔可夫决策过程：离散随机动态规划。John Wiley & Sons（2014年）。

Qi, Z. 和 Liu, Y. “利用 D学习估计最优个体治疗规则。”电子统计学期刊，12(2):3601–3638（2018年）。

Qian, M. 和 Murphy, S. A. “个体化治疗规则的性能保证。”统计学年鉴，39(2):1180（2011年）。

Ramprasad, P., Li, Y., Yang, Z., Wang, Z., Sun, W. W. 和 Cheng, G. “用于强化学习策略评估的在线自助法推断。”《美国统计学会会刊》，1–14（2022年）。

Reiß, M. 和 Wahl, M. “主成分分析重构误差的非渐近上界。”《统计年刊》，48(2):1098–1123（2020年）。

Riedmiller, M. “神经网络拟合Q迭代——数据高效神经强化学习方法的初步实践。”欧洲机器学习会议，317–328（2005年）。

Rodbard, D.《连续血糖监测数据的解读：血糖变异性与血糖控制质量。》糖尿病技术与治疗，11(S1):S–55 (2009)。

Said, S. E. 和 Dickey, D. A.《未知阶自回归移动平均模型的单位根检验。》$Bio\ metrika$，71(3):599–607 (1984)。

Schmidt-Hieber, J. “使用深度神经网络和 ReLU 激活函数进行非参数回归。”《统计年刊》, 48(4):1875–1897 (2020年)。

Schulman, J., Levine, S., Abbeel, P., Jordan, M. 和 Moritz, P. “信赖域策略优化。” 国际机器学习会议, 1889–1897 (2015a)。

Schulman, J., Moritz, P., Levine, S., Jordan, M. 和 Abbeel, P. “使用广义优势估计进行高维连续控制。” *arXiv* 预印本 *arXiv:1506.02438*(2015b)。

Shi, C., Song, R. 和 Lu, W. “最优治疗决策的一致性与价值信息准则。”《统计年刊》, 49(1):49–75 (2021a)。

Shi, C., Song, R., Lu, W. 和 Fu, B. “针对具有异质个体治疗效应的最优治疗决策的极大极小投影学习。”《英国皇家统计学会会刊：*B*辑（统计方法学）》, 80(4):681–702 (2018年)。

Shi, C., Wan, R., Chernozhukov, V., and Song, R. “深度去偏的离策略区间估计。”载于国际机器学习会议，9580–9591。PMLR（2021年b）。

Shi, C., Wan, R., Song, R., Lu, W., and Leng, L. “马尔可夫决策过程是否适合这些数据：序贯决策中马尔可夫性质的检验。”载于国际机器学习会议，8807–8817。PMLR（2020年）。

Shi, C., Wang, X., Luo, S., Zhu, H., Ye, J. 和 Song, R. 《动态因果效应评估

在强化学习框架下进行 A/B 测试》。《美国统计学会会刊》，已录用（2021c）。

Shi, C.、Zhang, S.、Lu, W. 和 Song, R. “无限时域情形下强化学习价值函数的统计推断。”英国皇家统计学会期刊：*B*辑，84:765–793（2022年）。

Song, R.、Wang, W.、Zeng, D. 和 Kosorok, M. R. “用于动态治疗方案的惩罚型 q 学习。”《中华统计学杂志》，25(3):901（2015年）。

Sprague, N. “基于奖励的降维基迭代。”*2007*年电气电子工程师学会第六届国际发展与学习会议，187–192（2007年）。

Srivastava, N.、Hinton, G.、Krizhevsky, A.、Sutskever, I. 和 Salakhutdinov, R. “Dropout：一种防止神经网络过拟合的简单方法。”机器学习研究期刊，15(1):1929–1958（2014年）。

Staicu, A.-M., Li, Y., Crainiceanu, C. M. 和 Ruppert, D. “用于相关数据的似然比检验及其在纵向和函数型数据分析中的应用。”斯堪的纳维亚统计学杂志，41(4):932–949（2014年）。

Sutton, R. S. 和 Barto, A. G. 强化学习：导论。麻省理工学院出版社（2018年）。

Tsiatis, A. A.、Davidian, M.、Holloway, S. T. 和 Laber, E. B. 动态治疗方案：精准医学的统计方法。Chapman and Hall/CRC（2019年）。

Tsybakov, A.《统计学习中分类器聚合的最优方法》。《统计年刊》, 32(1):135–166（2004年）。

Uehara, M., Imaizumi, M., Jiang, N., Kallus, N., Sun, W. 和 Xie, T.《极小极大离线强化学习的有限样本分析：完备性、快速收敛率与一阶效率》。*arXiv* 预印本 *arXiv:2102.02981*（2021年）。

Uehara, M., Kiyohara, H., Bennett, A., Chernozhukov, V., Jiang, N., Kallus, N., Shi, C. 和 Sun, W. “部分可观测马尔可夫决策过程中的未来依赖的基于价值的离策略评估。” *arXiv*预印本 *arXiv:2207.13081*（2022年）。

Van Hasselt, H., Guez, A. 和 Silver, D. “采用双重Q学习的深度强化学习。” *AAAI*人工智能会议论文集，30(1)（2016年）。

Wang, L., Laber, E. B. 和 Witkiewitz, K. “采用交替深度神经网络的充分马尔可夫决策过程。” *arXiv*预印本，arXiv:1704.07531（2017年a）。

Wang, Z., Bapst, V., Heess, N., Mnih, V., Munos, R., Kavukcuoglu, K. 和 de Freitas, N. “采用经验回放的样本高效演员-评论家。” 国际学习表征会议 （2017年b）。

Whiteson, S., Taylor, M. E., 和 Stone, P.《用于价值函数近似的自适应瓦片编码》。（2007年）。

Xu, Z., Laber, E., Staicu, A.-M., 和 Severus, E.《用于精准医学的潜在状态模型》。*arXiv* 预印本 *arXiv:2005.13001*（2020年）。

Yao, F., M¨uller, H.-G., 和 Wang, J.-L.《用于稀疏纵向数据的函数型数据分析》。《美国统计学会会刊》，100(470):577–590（2005年）。

Zhang, B., Tsiatis, A. A., Laber, E. B., and Davidian, M. “序贯治疗决策中最优动态治疗方案的稳健估计。”*Biometrika*，100(3):681–694（2013年）。

Zhang, X. and Wang, J. “从稀疏函数型数据到稠密函数型数据及更进一步。”《统计年刊》，44(5):2281–2321（2016年）。

Zhang, Y., Laber, E. B., Davidian, M., 和 Tsiatis, A. A.《可解释动态治疗方案》。《美国统计学会会刊》，113(524):1541–1549（2018年）。

Zhao, Y.-Q., Zeng, D., Laber, E. B., 和 Kosorok, M. R.《用于估计最优动态治疗方案的新统计学习方法》。《美国统计学会会刊》，110(510):583–598（2015年）。

Zhou, W., Zhu, R., 和 Qu, A.《通过 pT-Learning 估计最优无限时域动态治疗方案》。*arXiv* 预印本 *arXiv:2110.10719*（2021年）。

—。《通过 pT-Learning 估计最优无限时域动态治疗方案》。《美国统计学会会刊》，已接收（2022年）。

Zhu, L., Lu, W., 和 Song, R.《移动健康中的因果效应估计与最优剂量建议》。载于国际机器学习会议，11588–11598。PMLR（2020年）。

A 补充定义

本节将介绍更多定义，以便在补充信息的其余部分中使用。

注意， $\mathcal{F}_0$ 是一个用于建模 Q 函数而不使用主成分分析的函数类。我们需要基于高维输入 z的主成分分析，再定义两个函数类：

定义 A.1。

$$\mathcal{F}_1(L,\{d_i\}_{i=0}^{L+1},s,V_{max},G,\kappa)=\{f_1:\mathcal{S}\times\mathcal{A}\to\mathbb{R},f_1(x,z,a)=$$
$$f_0(x,(\sum_{k}^{\kappa}U_kU_k{}^T)z,a)=f_0(x,z^*,a),f_0\in\mathcal{F}_0(L,\{d_i\}_{i=0}^{L+1},s,V_{max})\},$$

其中， $\mathcal{F}_0$ 在第 4 节中定义， $G=\sum_{k=1}^{m}\lambda_k\mathbf{U_k}\mathbf{U_k}{}^T$ 是z的协方差矩阵。 $z^*=($ $\sum_k^\kappa U_kU_k{}^T)z$ 是从原始高频向量 z的前 κ 个主成分分析值中恢复得到的向量。

$\mathcal{F}_2$ 的定义已在第 4 节中提及。此处我们将对其进行如下正式定义：

$$\mathcal{F}_2(L,\{d_i\}_{i=0}^{L+1},s,V_{max},G,\kappa)=\{f_2:\mathcal{S}\times\mathcal{A}\to\mathbb{R},f_2(x,z,a)=$$
$$f_0(x,\begin{pmatrix}I_\kappa & 0_{\kappa\times(m-\kappa)}\end{pmatrix}\Sigma^{-\frac{1}{2}}U^Tz,a)=f_0(x,v_\kappa,a),f_0\in\mathcal{F}_0(L,\{d_i\}_{i=0}^{L+1},s,V_{max})\},$$

其中， v_κ 是包含原始高频向量 z的前 κ 个主成分分析值的向量。

注意， $\mathcal{F}_2(L,\{d_i\}_{i=0}^{L+1},s,V_{max},G,\kappa)$是我们在 Spectrum Q-迭代学习算法中用于对 Q 函数进行建模的函数类。在实践中，我们将使用ˆ估计值 G 及其对应的主成分分析向量ˆv_κ ，如 $\mathcal{F}_2$所示。两个类 $\mathcal{F}_1$ 和 $\mathcal{F}_2$ 之间的关系将用于主要定理（附录 C 节中的命题 1 和引理 2）的证明。

这里列出集中系数的定义如下：

定义 A.2。 设 σ_1 和 σ_2 是定义在 $\mathcal{S}\times\mathcal{A}$上的两个绝对连续分布。设 $\{\pi_t\}_{t\geq1}=\{\pi_1,\ldots,\pi_m\}$ 是一列策略。对于分布 σ_1中的一个动作-状态对 S_0， A_0，将第 t 个决策点的状态记为 S_t， $t=1$， 2，.....， m。假设我们根据 π_t在第 t个决策点采取动作 A_t 。那么，我们将 $(S_t,\ A_t)$ 的边缘分布记为 $P^{\pi_m}P^{\pi_{m-1}}...P^{\pi_1}\sigma_1$。由此可以定义集中系数

$$\omega_\infty(m;\sigma_1,\sigma_2)=\sup_{\pi_1,\ldots,\pi_m}\left\|\frac{dP^{\pi_m}P^{\pi_{m-1}}...P^{\pi_1}\sigma_1(s,a)}{d\sigma_2(s,a)}\right\|_{\infty,\sigma_2}. \tag{2}$$

强化学习文献中常见类似的集中系数定义（Fan 等人，2020年；Lazaric 等人，2016年）。

B 第一定理的详细版本

定理 4.6 的详细版本如下：设 μ 是 $(\mathcal{S} \times \mathcal{A})$ 上的某个分布，使得 $\mu(\mathcal{S} \times \mathcal{A}) = \mu_1(\mathcal{S}) \times \mu_2(\mathcal{A})$，其中 $\mu_2(a) > 0,\ \forall a \in \mathcal{A}$ 且 $\frac{d\mu_1}{ds}(s)$ 与 0 保持正距离（这里我们希望对动作空间 $\mathcal{A}$采用更一般的测度 μ_2 ）。定义$\beta_j^* = \beta_j \times \prod_{l=j+1} \min(\beta_l, 1)$ 和 $\alpha^* = \max_{j \in [q]} \frac{t_j}{2\beta_j^* + t_j}$，其中 β_j，q，t_j，p_j来自定义 4.2 以及第 4 节中 $\mathcal{G}_0(\{p_j, t_j, \beta_j, H_j\}_{j \in [q]})$ 的定义。

在假设 4.3 至 4.5 下，并且满足条件：n 足够大，使得 $\max\{\sum_j (t_j + \beta_j + 1)^{3+t_j}, \max_j p_j\} \leq C_\xi (\log(n))^\xi$ 对某个 $\xi > 0,\ C_\xi > 0$成立，则存在一个函数类 $\mathcal{F}_2(L^*, \{d_j^*\}_{i=0}^{L+1},\ s^*,\ V_{max},\ \hat{}G,\ \kappa)$，用于建模 $Q(s,\ a)$，其组成部分均为稀疏 ReLU 网络，且超参数满足：层数为 $L^* = O(\log^{\xi^*} n)$，第一层宽度满足界 $d_1^* \leq n^{\alpha^*}$，输出维度为 $d_{L^*+1}^* = 1$，其他层的宽度满足界 $d_0^* = m_0 + m \leq min_{j \in \{1,2,\ldots,L\}} d_j^* \leq max_{j \in \{1,2,\ldots,L\}} d_j^* = O(n^{\xi^*})$，稀疏度的阶为 $s^* \asymp d_1^* \kappa + n^{\alpha^*} \log^{\xi^*}(n)$，其中某个 $\xi^* > 1 + 2\xi$。在这种情况下，我们可以证明˜：由我们的算法 1 得到的 Q_K （由 $\mathcal{F}_2$ 建模）满足：

$$\left\| Q^{\pi^*} - \tilde{Q}_K \right\|_{2,\mu}^2 = O_p(|\mathcal{A}|(n^{\alpha^*} \log^{\xi^*} n + d_1^* \kappa) n^{-1} \log^{\xi^*+1} n + L_{\mathcal{F}_0}(e^{-\zeta\kappa} - e^{-\zeta m} + n^{-2\Delta}) + \frac{\gamma^{2K}}{(1-\gamma)^2} R_{max}^2),$$

其中，$|\mathcal{A}|$ 是治疗选项的数量。

C 定理证明

C.1 定理 4.6 的证明概要

定理 4.6 的详细表述如下：设 μ 是 $(\mathcal{S} \times \mathcal{A})$ 上的某个分布，使得 $\mu(\mathcal{S} \times \mathcal{A}) = \mu_1(\mathcal{S}) \times \mu_2(\mathcal{A})$，其中 $\mu_2(a) > 0,\ \forall a \in \mathcal{A}$ 且 $\frac{d\mu_1(s)}{ds}$ 与 0 保持正距离（这里我们希望对动作空间 $\mathcal{A}$采用更一般的测度 μ_2 ）。定义$\beta_j^* = \beta_j \times \prod_{l=j+1} \min(\beta_l, 1)$ 和 $\alpha^* = \max_{j \in [q]} \frac{t_j}{2\beta_j^* + t_j}$，其中 β_j，q，t_j，p_j来自定义 4.2 以及第 4 节中 $\mathcal{G}_0(\{p_j, t_j, \beta_j, H_j\}_{j \in [q]})$ 的定义。

在假设 4.3 至 4.5 下，并且满足条件：n 足够大，使得 $\max\{\sum_j (t_j + \beta_j + 1)^{3+t_j}, \max_j p_j\} \leq C_\xi (\log(n))^\xi$ 对于某个 $\xi > 0,\ C_\xi > 0$，存在一个函数类 $\mathcal{F}_2(L^*, \{d_j^*\}_{i=0}^{L+1}, s^*, V_{max}, \hat{}G, \kappa)$，用于建模 $Q(s, a)$，其组成部分均为稀疏 ReLU 网络，且其超参数满足：层数为 $L^* = O(\log^{\xi^*} n)$，第一层宽度满足界 $d_1^* \leq n^{\alpha^*}$，输出维度为 $d_{L^*+1}^* = 1$，其他层的宽度满足界 $d_0^* = m_0 + m \leq min_{j \in \{1,2,\ldots,L\}} d_j^* \leq max_{j \in \{1,2,\ldots,L\}} d_j^* = O(n^{\xi^*})$，稀疏度的阶为 $s^* \asymp d_1^* \kappa + n^{\alpha^*} \log^{\xi^*}(n)$，其中某个 $\xi^* > 1 + 2\xi$。在这种情况下，我们可以证明~，由我们的算法 1 得到的（由 $\mathcal{F}_2$ 建模的） Q_K 满足：

$$\left\| Q^{\pi^*} - \tilde{Q}_K \right\|_{2,\mu}^2 = O_p(|\mathcal{A}|(n^{\alpha^*} \log^{\xi^*} n + d_1^* \kappa) n^{-1} \log^{\xi^*+1} n + L_{\mathcal{F}_0}(e^{-\zeta\kappa} - e^{-\zeta m} + n^{-2\Delta}) + \frac{\gamma^{2K}}{(1-\gamma)^2} R_{max}^2),$$

其中，$|\mathcal{A}|$ 是治疗选项的数量。

首先，我们需要将其记作

$$\mathcal{F}_1(L^*, \{\{d_i^*\}_{i=1}^{L+1}, d_0 = m_0 + m\}, s^* - d_1\kappa, V_{max}, \hat{G}, \kappa)$$

$\mathcal{F}_1$,

$$\mathcal{F}_2(L^*, \{d_i^*\}_{i=0}^{L+1}, s^*, V_{max}, \hat{G}, \kappa)$$

如 $\mathcal{F}_2$ 和

$$\mathcal{F}_0(L^*, \{\{d_i^*\}_{i=1}^{L+1}, d_0 = m_0 + m\}, s^* - d_1\kappa, V_{max})$$

如 $\mathcal{F}_0$。

注意，假设转移密度函数 $q(s'|s,\ a)$ 满足 $\sup_{s',\ s,\ a} q(s'|s,\ a) \leq c_q$（如第2.2节所述）。令 $q^j(s'|s,\ a)$ 表示遵循策略序列 π_1，...，π_{j-1} 的决策后状态 s' 的密度。于是可以证明

$$\begin{aligned} q^j(s'|s,a) &= \int_{x\in\mathcal{S}} \sum_{a_x\in\mathcal{A}} q(s'|x,a)\pi_{j-1}(a_x|x)q^{j-1}(x|s,a)dx \\ &\leq c_q \int_{x\in\mathcal{S}} \sum_{a_x\in\mathcal{A}} \pi_{j-1}(a_x|x)q^{j-1}(x|s,a)dx = c_q. \end{aligned}$$

$\{\pi_1, ..., \pi_m\}$ 之后，s'，a' 的联合分布密度为

$$\int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \pi_m(a'|s')q^{(m)}(s'|s,a)d\mu(s,a) \leq c_q \int_{(s,a)\in\mathcal{S}\times\mathcal{A}} d\mu(s,a) = c_q.$$

因此，~~只要抽样分布 σ 有~~正的下界，$dP^{\pi_m}P^{\pi_{m-1}}...P^{\pi_1}\mu(s,\ a)\,d\sigma(s,\ a)$ 就有界。也就是说，对于某个常数 $c_{\mu,\ \sigma} > 0$，有 $\omega_\infty(m; \mu,\ \sigma) \leq c_{\mu,\ \sigma}$（请回忆 $\omega_\infty(m;\ \mu,\ \sigma)$ 来自定义2）。

然后可以证明

$$\frac{1}{(1-\gamma)^2}\sum_{m=0}^{\infty}\gamma^{m-1}(m+1)[\omega_\infty(m;\mu,\sigma)]^{\frac{1}{2}} \leq \frac{1}{(1-\gamma)^2}\sum_{m=0}^{\infty}\gamma^{m-1}(m+1)c_{\mu,\sigma}{}^{\frac{1}{2}} = \phi_{\mu,\sigma},$$

对某个常数 $\phi_{\mu,\ \sigma} > 0$。

~然后，我们可以得出关于 Q_K 近似误差的如下引理:

~**引理 C.1.** 在算法 *1* 的第 K 次迭代中得到的估计的 *Q* 函数 Q_K 满足

$$\left\|Q^* - \tilde{Q}_K\right\|_{2,\mu}^2 \leq 2\gamma^2(1-\gamma)^4\epsilon_{max}^2\phi_{\mu,\sigma}^2 + 8\frac{\gamma^{2K}R_{max}^2}{(1-\gamma)^2}, \tag{3}$$

其中 $\epsilon_{max} =$ 最大值$_{k\in\{1,2\ldots,K\}}\left\|T\tilde{Q}_{k-1} - \tilde{Q}_k\right\|_\sigma$。

~也就是说，在我们的算法中，第 K 次迭代的结果 Q_K 与对应于最优策略的 Q^* 之间的距离，可以由 Spectrum Q 迭代算法所有 K 次迭代中的最大近似误差 $\left\|T\tilde{Q}_{k-1} - \tilde{Q}_k\right\|_\sigma^2$ ~ ~ 界定。我们知道 Q_{k-1}，$Q_k \in \mathcal{F}_2$。

为了界定每次迭代中的项 $\left\|T\tilde{Q}_{k-1} - \tilde{Q}_k\right\|_\sigma^2$，我们将应用 Fan 等人（2020）的定理 6.2，并且可以证明

$$\left\|T\tilde{Q}_{k-1} - \tilde{Q}_k\right\|_\sigma^2 \leq (1+\epsilon)^2\omega(\mathcal{F}_2) + C\frac{V_{max}^2}{n\epsilon}\log N_{\delta,2} + C'V_{max}\delta$$

对于 $\forall\epsilon \in (0,1]$, 其中 $\omega(\mathcal{F}_2) = \sup^{g\in\mathcal{F}_2}\inf_{f\in\mathcal{F}_2}\|f - Tg\|_\sigma^2$。这里，$N_{\delta,\ 2}$ 是函数类 $\in \mathcal{F}_2$ 相对于 l_∞ 范数的最小 σ-覆盖集的基数。

如果我们取 $\epsilon = 1$ 并固定 $\delta = \frac{1}{n}$，则有

$$\left\|T\tilde{Q}_{k-1} - \tilde{Q}_k\right\|_\sigma^2 \leq 4\omega(\mathcal{F}_2) + C\frac{V_{max}^2}{n}\log N_{\frac{1}{n},2} + \frac{C'V_{max}}{n}. \tag{4}$$

那么我们需 要分别界定 $\omega(\mathcal{F}_2) = \sup^{g\in\mathcal{F}_2}\inf_{f\in\mathcal{F}_2}\|f - Tg\|_\sigma^2$和 $\log N_{\frac{1}{n},_2}$.

项 $\sup^{g\in\mathcal{F}_2}\inf_{f\in\mathcal{F}_2}\|f - Tg\|_\sigma^2$ 与使用函数 $f \in \mathcal{F}_2$ 近似同一函数类中的函数 g 时产生的误差有关，其中采用了最优贝尔曼运算。注意，在 Q 迭代算法的每次迭代中，我们使用类 ReLU 函数类中的 Q_k 来拟合响应值 $Y_i = R_i + \gamma\max_{a\in\mathcal{A}}\tilde{Q}_{k-1}(S_i',\ a)$。如果对 Y_i取期望，我们知道响应值 Y_i 背后的真实模型是来自 Tg 类的一个函数：$g \in \mathcal{F}_2$。正如 Fan 等人（2020）所论述的，当我们使用 $Q_k \in \mathcal{F}_2$拟合 Y_i 时，项 $\sup^{g\in\mathcal{F}_2}\inf_{f\in\mathcal{F}_2}\|f - Tg\|_\sigma^2$ 可以表示为偏差项 ~。

我们首先　　围绕这个与偏差相关的术语。我们需要证明以下命题：

命题 C.2。 函数类 $\mathcal{F}_1(L, \{\{d_i\}_{i=1}^{L+1}, d_0 = m_0 + m\}, s, V_{max}, \hat{G}, \kappa)$ 是 $\mathcal{F}_2(L, \{\{d_i\}_{i=1}^{L+1}, d_0 = \kappa + m_0\}, s + d_1\kappa, V_{max}, \hat{G}, \kappa)$的子集。

由此命题可得 $\mathcal{F}_1 \subset \mathcal{F}_2$。接下来，我们可以提出以下引理，它依赖于我们刚刚证明的命题 C.2：

引理 C.3。 在定理 *4.6* 的相同假设下，

$$\sup_{g\in\mathcal{F}_2}\inf_{f\in\mathcal{F}_2}\|f - Tg\|_\sigma^2 \le 2\sup_{f'\in\mathcal{G}_0}\inf_{f_0\in\mathcal{F}_0}\left\|f_0 - f'\right\|_\sigma^2 + 2L_{\mathcal{F}_0}\|Z - Z^*\|_\sigma^2,$$

其中 $L_{\mathcal{F}_0} = \text{上确界}_{f\in\mathcal{F}_0}\text{上确界}_{x\neq y}\frac{|f(y) - f(x)|^2}{\|y-x\|^2}$。 $\{Z\}_{j=1}^J$ 是分布 σ 中的高频向量，而 $Z^* = (\sum_{k=1}^{\kappa}\hat{U}_k\hat{U}_k^T)Z$ 是根据 Z的主成分分析值恢复得到的向量。

应用引理 C.3 后，我们现在可以将项 $\sup_{g\in\mathcal{F}_2}\inf_{f\in\mathcal{F}_2}\|f - Tg\|_\sigma^2$ 界定为 $\sup_{f'\in\mathcal{G}_0}\inf_{f_0\in\mathcal{F}_0}\left\|f_0 - f'\right\|_\sigma^2$ 和 $\|Z - Z^*\|_\sigma^2$。

回顾一下， $\mathcal{F}_0$ 是

$$\mathcal{F}_0(L^*, \{\{d_i^*\}_{i=1}^{L+1}, d_0 = m_0 + m\}, s^* - d_1\kappa, V_{max}),$$

正如本节开头所述。我们知道， $\mathcal{F}_0$ 的稀疏度满足 $s^* - d_1\kappa \asymp n^{\alpha^*}\log^{\xi^*}(n)$，且 $\mathcal{F}_0$ 的网络架构满足 $d_1^* \le n$ 和 $m_0 + m \le min_{j\in\{1,2,\ldots,L\}}d_j^* \le max_{j\in\{1,2,\ldots,L\}}d_j^* = O(n^{\xi^*})$。因此，我们可以使用 Fan 等人（2020）的公式 (4.18) 来界定项 $\sup_{f'\in\mathcal{G}_0}\inf_{f_0\in\mathcal{F}_0}\left\|f_0 - f'\right\|_\sigma^2$：

$$\sup_{f'\in\mathcal{G}_0}\inf_{f_0\in\mathcal{F}_0}\left\|f_0 - f'\right\|_\sigma^2 \le n^{\alpha^*-1},$$

其中， $\mathcal{F}_0$ 中的神经网络条件与应用 Fan 等人（2020）公式 (4.18) 的条件相匹配。

现在我们可以通过下一个引理来界定项 $\|Z - Z^*\|_\sigma^2$。

引理 C.4. 对于高频向量 Z，我们假设 $E(Z)=0$，且 $Cov(Z)$ 的特征值遵循以下指数衰减趋势 $\lambda_k = O(e^{-\zeta k})$。我们还假设 $Cov(Z)$ 的估计满足 $\left\| \hat{U} - U \right\| = O_p(n^{-\Delta})$。于是我们有

$$\|Z - Z^*\|_\sigma^2 = O_p(e^{-\zeta\kappa} - e^{-\zeta m} + n^{-2\Delta})$$

根据这两个引理，我们知道

$$\sup_{g\in\mathcal{F}_2} \inf_{f\in\mathcal{F}_2} \|f - Tg\|_\sigma^2 = O_p(n^{\alpha^*-1} + L_{\mathcal{F}_0}(e^{-\zeta\kappa} - e^{-\zeta m} + n^{-2\Delta})).$$

然后，当我们取 $\delta = \frac{1}{n}$ 时，需要对方程 (4) 中的 $\log N_{\delta,\ 2}$ 给出一个界。这里我们需要证明另一个命题，即命题 C.5：

命题 C.5. 当 $\delta = \frac{1}{n}$ 时，对于某个常数 C_1，有 $\log N_{\delta,\ 2} \le C_1|\mathcal{A}|(n^{\alpha^*}(\log\ n)^{\xi^*} + d_1^*\kappa)(\log\ n)^{1+\xi^*}$。

然后由方程 (4)，我们知道

$$\left\| T\tilde{Q}_{k-1} - \tilde{Q}_k \right\|_\sigma^2 = O_p(|\mathcal{A}|n^{-1}(n^{\alpha^*}(\log n)^{\xi^*} + d_1^*\kappa)(\log n)^{\xi^*+1} + L_{\mathcal{F}_0}(e^{-\zeta\kappa} - e^{-\zeta m} + n^{-2\Delta})) \tag{5}$$

Then 由方程 (5) 和方程 (3)，定理 4.6 的证明完成 .

C.2 定理 4.8 的证明

定理 4.8：在假设 4.3 至 4.7 以及定理 4.6 中 $\mu, \mathcal{F}_2, n$ 的条件下，我们可以证明：

$$\begin{aligned} E_{\mu_1}[V^{\pi^*}(s) - V^{\pi_K}(s)] = O_p(\frac{1}{1-\gamma}\{&|\mathcal{A}|(n^{\alpha^*}\log^{\xi^*} n + d_1^*\kappa)n^{-1}\log^{\xi^*+1} n \\ &+ L_{\mathcal{F}_0}(e^{-\zeta\kappa} - e^{-\zeta m} + n^{-2\Delta}) + \frac{\gamma^{2K}}{(1-\gamma)^2}R_{max}^2\}^{\frac{\eta+1}{\eta+2}}). \end{aligned}$$

证明。该 $\tilde{}$ 在算法 1 中获得的策略 π_K 是关于 Q_K 的贪心策略。

首先，可以应用 Shi 等人（2022）的一条命题：

命题 C.6（Shi 等人（2022）中的不等式 E.10 和 E.11）。

$$\begin{aligned} E_{\mu_1}[V^{\pi^*}(s) - V^{\pi_K}(s)] \leq & \int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\} d\mu_1(s) \\ & + \frac{c\gamma}{1-\gamma} \int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\} ds. \end{aligned} \tag{6}$$

接下来，只需对这两项 $\int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\} d\mu_1(s)$ 和 $\int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\} ds$ 进行界定。我们先说明如何对下式进行界定：

$$\int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\} d\mu_1(s).$$

定义 $\hat{a}_{\pi^K}(s) = \arg\max_{a\in\mathcal{A}} \tilde{Q}_K(s,a)$。对于 $\epsilon > 0$，令 $A_* = \{0 < \max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s, \hat{a}_{\pi^K}) \leq \epsilon\}$。于是我们知道补集 $A_*^C = \{\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s, \hat{a}_{\pi^K}) > \epsilon\} \cup \{\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s, \hat{a}_{\pi^K}) = 0\}$。由此可得

$$\begin{aligned} & \int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\} d\mu_1(s) = \\ & \int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\} \mathbb{1}(s \in A_*) d\mu_1(s) + \\ & \int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\} \mathbb{1}(s \in A_*^C) d\mu_1(s). \end{aligned} \tag{7}$$

第一步是对公式（7）的第一部分进行界定，

$$\int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\} \mathbb{1}(s \in A_*) d\mu_1(s).$$

已知

$$\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) = \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\pi^*(a|s)$$

and

$$Q^{\pi^*}(s,\hat{a}_{\pi^K}) = \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\pi^K(a|s),$$

因为 π^* 和 π^K 都是贪心策略（π^* 可视为相对于 Q^{π^*}的贪心策略）。于是，当 $s \in A_*$ 时，$\sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\} = \max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s,\hat{a}_{\pi^K}) \leq \epsilon$。也就是说，

$$\int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\}\mathbb{1}(s\in A_*)d\mu_1(s) \leq \epsilon \int_{s\in\mathcal{S}} \mathbb{1}(s\in A_*)d\mu_1(s). \tag{8}$$

此外，当 $s \in A_*$时，根据 A_*的定义，我们知道 $\hat{a}_{\pi^K}$永远不会与 $a \in \arg\max_{a'\in\mathcal{A}} Q^{\pi^*}(s,a')$ 相同，因为在 A_*中，$\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s,\hat{a}_{\pi^K}) > 0$ 始终成立。因此，$\max_{a\in\mathcal{A}-\arg\max_{a'} Q^{\pi^*}(x,a')} Q^{\pi^*}(s,a) \geq Q^{\pi^*}(s,\hat{a}_{\pi^K})$。于是我们知道，$\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s,\hat{a}_{\pi^K}) \leq \epsilon$会导出 $\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - \max_{a\in\mathcal{A}-\arg\max_{a'} Q^{\pi^*}(x,a')} Q^{\pi^*}(s,a) \leq \epsilon$。

因此，可以得到

$$\int_{s\in\mathcal{S}} \mathbb{1}(s\in A_*)d\mu_1(s) \leq \int_{s\in\mathcal{S}} \mathbb{1}\{s : \max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - \max_{a\in\mathcal{A}-\arg\max_{a'} Q^{\pi^*}(x,a')} Q^{\pi^*}(s,a) \leq \epsilon\}d\mu_1(s) = O(\epsilon^\eta).$$

根据假设 4.7，结合公式 (8)，可知

$$\int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\}\mathbb{1}(s\in A_*)d\mu_1(s) = O(\epsilon^{\eta+1}). \tag{9}$$

然后，我们可以对方程（7）的第二部分进行界定，

$$\int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\}\mathbb{1}(s\in A_*^C)d\mu_1(s).$$

回顾 $A_*^C = \{\max_{a\in\mathcal{A}} Q^{\pi^*}(s,\ a) - Q^{\pi^*}(s,\ \hat{}a_{\pi^K}) > \epsilon\} \cup \{\max_{a\in\mathcal{A}} Q^{\pi^*}(s,\ a) - Q^{\pi^*}(s,\ \hat{}a_{\pi^K}) = 0\}$ 的定义。我们知道，$\{\max_{a\in\mathcal{A}} Q^{\pi^*}(s,\ a) - Q^{\pi^*}(s,\ \hat{}a_{\pi^K}) = 0\}$ 是一个平凡情形。容易证明

$$\int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\}\mathbb{1}\{s : \max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s,\hat{a}_{\pi^K}) = 0\} d\mu_1(s) = 0,$$

as $\pi^*(a|s) = \pi^K(a|s)$ 当 $\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s,\hat{}a_{\pi^K}) = 0$ 始终成立 .

因此,

$$\int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\}\mathbb{1}(s \in A_*^C) d\mu_1(s) =$$

$$\int_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q^{\pi^*}(s,a)\{\pi^*(a|s) - \pi_K(a|s)\}\mathbb{1}\{s : \max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s,\hat{a}_{\pi^K}) > \epsilon\} d\mu_1(s).$$

我们可以证明，当 $\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s,\hat{}a_{\pi^K}) > \epsilon$ 成立时,

$$\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s,\hat{a}_{\pi^K}) \le 2\max_{a\in\mathcal{A}} |Q^{\pi^*}(s,a) - \tilde{Q}_K(s,a)|.$$

为证明这一点，注意到 $\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - Q^{\pi^*}(s,\hat{}a_{\pi^K}) = \max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - \tilde{Q}_K(s,\hat{}a_{\pi^K}) + \tilde{Q}_K(s,\hat{}a_{\pi^K}) - Q^{\pi^*}(s,\hat{}a_{\pi^K})$。显然可以证明

$$\tilde{Q}_K(s,\hat{a}_{\pi^K}) - Q^{\pi^*}(s,\hat{a}_{\pi^K}) \le \max_{a\in\mathcal{A}} |Q^{\pi^*}(s,a) - \tilde{Q}_K(s,a)|.$$

此外，我们可以得到 $\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - \tilde{Q}_K(s,\hat{}a_{\pi^K}) = \max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - \tilde{Q}_K(s, a \in \tilde{}\arg\max_{a\in\mathcal{A}} Q^*(s,a)) + Q_K(s, a \in \arg\max_{a\in\mathcal{A}} Q^*(s,a)) - Q_K(s,\hat{}a_{\pi^K})$。由此可知，$\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - \tilde{Q}_K(s, a \in \arg\max_{a\in\mathcal{A}} Q^*(s,a)) \le \max_{a\in\mathcal{A}} |Q^{\pi^*}(s,a) - \tilde{Q}_K(s,a)|$ and $\tilde{Q}_K(s, a \in \arg\max_{a\in\mathcal{A}} Q^*(s,a)) - \tilde{Q}_K(s,\hat{}a_{\pi^K}) \le 0$，这意味着 $\max_{a\in\mathcal{A}} Q^{\pi^*}(s,a) - \tilde{Q}_K(s,\hat{}a_{\pi^K}) \le \max_{a\in\mathcal{A}} |Q^{\pi^*}(s,a) - \tilde{Q}_K(s,a)|$。

为了界定 $\max_{a\in\mathcal{A}} |Q^{\pi^*}(s,a) - \tilde{Q}_K(s,a)|$，我们知道

$$\max_{a\in\mathcal{A}} |Q^{\pi^*}(s,a) - \tilde{Q}_K(s,a)|^2 \le \sum_{a\in\mathcal{A}} |Q^{\pi^*}(s,a) - \tilde{Q}_K(s,a)|^2 \le$$

$$\frac{1}{\min_{a\in\mathcal{A}} \mu_2(a)} \sum_{a\in\mathcal{A}} |Q^{\pi^*}(s,a) - \tilde{Q}_K(s,a)|^2 \mu_2(a).$$

于是，可以得到

$$\begin{aligned}
&\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}Q^{\pi^*}(s,a)\{\pi^*(a|s)-\pi_K(a|s)\}\mathbb{1}(s\in A_*^C)d\mu_1(s)=\\
&\int_{s\in\mathcal{S}}(\max_{a\in\mathcal{A}}Q^{\pi^*}(s,a)-Q^{\pi^*}(s,\hat{a}_{\pi^K}))\mathbb{1}(s\in A_*^C)d\mu_1(s)\leq\\
&\int_{s\in\mathcal{S}}\frac{\left(2\max_{a\in\mathcal{A}}|Q^{\pi^*}(s,a)-\tilde{Q}_K(s,a)|\right)^2}{(\max_{a\in\mathcal{A}}Q^{\pi^*}(s,a)-Q^{\pi^*}(s,\hat{a}_{\pi^K}))}\mathbb{1}(s\in A_*^C)d\mu_1(s)\leq\\
&\frac{4}{\epsilon\min_{a\in\mathcal{A}}\mu_2(a)}\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|Q^{\pi^*}(s,a)-\tilde{Q}_K(s,a)|^2\mu_2(a)\mathbb{1}(s\in A_*^C)d\mu_1(s)\leq\\
&\frac{4}{\epsilon\min_{a\in\mathcal{A}}\mu_2(a)}\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}|Q^{\pi^*}(s,a)-\tilde{Q}_K(s,a)|^2\mu_2(a)d\mu_1(s)=\\
&\frac{4}{\epsilon\min_{a\in\mathcal{A}}\mu_2(a)}\left\|Q^{\pi^*}(s,a)-\tilde{Q}_K(s,a)\right\|_{\mu,2}^2.
\end{aligned}\tag{10}$$

根据定理 4.6，我们已经得到

$$\begin{aligned}
\left\|Q^{\pi^*}-\tilde{Q}_K\right\|_{2,\mu}^2=O_p(&\gamma^2(1-\gamma)^4\times[|\mathcal{A}|n^{-1}(n^{\alpha^*}(\log n)^{\xi^*}+d_1^*\kappa)(\log n)^{\xi^*+1}\\
&+L_{\mathcal{F}_0}(e^{-\zeta\kappa}-e^{-\zeta m}+n^{-2\Delta})]+\frac{\gamma^{2K}}{(1-\gamma)^2}R_{max}^2).
\end{aligned}\tag{11}$$

然后将方程（11）代入方程（10），可得

$$\begin{aligned}
&\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}Q^{\pi^*}(s,a)\{\pi^*(a|s)-\pi_K(a|s)\}\mathbb{1}(s\in A_*^C)d\mu_1(s)\\
&=O_p(\frac{1}{\epsilon\min_{a\in\mathcal{A}}\mu_2(a)}\gamma^2(1-\gamma)^4\times[|\mathcal{A}|n^{-1}(n^{\alpha^*}(\log n)^{\xi^*}\\
&+d_1^*\kappa)(\log n)^{\xi^*+1}+L_{\mathcal{F}_0}(e^{-\zeta\kappa}-e^{-\zeta m}+n^{-2\Delta})]+\frac{\gamma^{2K}}{(1-\gamma)^2}R_{max}^2).
\end{aligned}\tag{12}$$

然后根据方程（7）、方程（9）和方程（12），可得

$$\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}Q^{\pi^*}(s,a)\{\pi^*(a|s)-\pi_K(a|s)\}d\mu_1(s)$$
$$=O_p(\epsilon^{\eta+1}+\frac{1}{\epsilon\min_{a\in\mathcal{A}}\mu_2(a)}\gamma^2(1-\gamma)^4\times[|\mathcal{A}|n^{-1}(n^{\alpha^*}(\log n)^{\xi^*}$$
$$+d_1^*\kappa)(\log n)^{\xi^*+1}+L_{\mathcal{F}_0}(e^{-\zeta\kappa}-e^{-\zeta m}+n^{-2\Delta})]+\frac{\gamma^{2K}}{(1-\gamma)^2}R_{max}^2).$$

如果我们取

$$\epsilon^{\eta+1}=\frac{1}{\epsilon\min_{a\in\mathcal{A}}\mu_2(a)}\gamma^2(1-\gamma)^4\times[|\mathcal{A}|n^{-1}(n^{\alpha^*}(\log n)^{\xi^*}+d_1^*\kappa)(\log n)^{\xi^*+1}+$$
$$L_{\mathcal{F}_0}(e^{-\zeta\kappa}-e^{-\zeta m}+n^{-2\Delta})]+\frac{\gamma^{2K}}{(1-\gamma)^2}R_{max}^2,$$

则可得到以下结果:

$$\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}Q^{\pi^*}(s,a)\{\pi^*(a|s)-\pi_K(a|s)\}d\mu_1(s)=O_p(\{\gamma^2(1-\gamma)^4\times[|\mathcal{A}|n^{-1}(n^{\alpha^*}(\log n)^{\xi^*}$$
$$+d_1^*\kappa)(\log n)^{\xi^*+1}+L_{\mathcal{F}_0}(e^{-\zeta\kappa}-e^{-\zeta m}+n^{-2\Delta})]+\frac{\gamma^{2K}}{(1-\gamma)^2}R_{max}^2\}^{\frac{\eta+1}{\eta+2}}).\tag{13}$$

然后，采用类似的过程，我们可以得到关于勒贝格测度的结果,

$$\int_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}Q^{\pi^*}(s,a)\{\pi^*(a|s)-\pi_K(a|s)\}ds=O_p(\{\gamma^2(1-\gamma)^4\times[|\mathcal{A}|n^{-1}(n^{\alpha^*}(\log n)^{\xi^*}$$
$$+d_1^*\kappa)(\log n)^{\xi^*+1}+L_{\mathcal{F}_0}(e^{-\zeta\kappa}-e^{-\zeta m}+n^{-2\Delta})]+\frac{\gamma^{2K}}{(1-\gamma)^2}R_{max}^2\}^{\frac{\eta+1}{\eta+2}}).\tag{14}$$

结合公式 (6)、公式 (13) 和公式 (14)，即可得到定理 4.8 的结论。

□

C.3 引理 C.1 的证明

˜引理 C.1 如下：对于算法 1 的第 K Q_K 次迭代中得到的估计的 Q 函数，我们有

$$\left\|Q^* - \tilde{Q}_K\right\|_{2,\mu}^2 \leq 2\gamma^2(1-\gamma)^4\epsilon_{max}^2\phi_{\mu,\sigma}^2 + 8\frac{\gamma^{2K}R_{max}^2}{(1-\gamma)^2},$$

其中 $\epsilon_{max} = \max_{k\in\{1,2...,K\}}\left\|T\tilde{Q}_{k-1} - \tilde{Q}_k\right\|_\sigma$。

˜ ˜证明。首先，记 $\rho_k = TQ_{k-1} - Q_k$，并将算子 P^π 定义如下：

$$(P^\pi Q)(s,a) = E[Q(S',A')|S'\ P(.|s,a), A'\ \pi(.|S')]. \tag{15}$$

我们可以采用 Fan 等人（2020）中的以下引理:

引理 C.7（Fan 等人（2020）中的引理 C.2）。

$$Q^* - \tilde{Q}_l \leq \sum_{i=k} l-1\gamma^{l-1-i}(P^{\pi^*})^{l-1-i}\rho_{i+1} + \gamma^{l-k}(P^{\pi^*})^{l-k}(Q^* - \tilde{Q}_k)$$

对于 $\forall k$、$l \in \mathbb{Z}$、$0 \leq k < l \leq K-1$。

代入 $l = K, k = 0$，可得

$$Q^* - \tilde{Q}_K \leq \sum_{i=0} K-1\gamma^{K-1-i}(P^{\pi^*})^{K-1-i}\rho_{i+1} + \gamma^K(P^{\pi^*})^K(Q^* - \tilde{Q}_0)$$

我们可以将 $U_i = (P^{\pi^*})^{K-1-i}$ 记作 $i = 0, 1, ..., K-1$，并将 $U_K = (P^{\pi^*})^K$ 记作。于是可得

$$|Q^*(s,a) - \tilde{Q}_K(s,a)| \leq \sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a) + \gamma^K(U_K|Q^* - \tilde{Q}_0|)(s,a).$$

两边平方可得

$$\begin{aligned}|Q^*(s,a) - \tilde{Q}_K(s,a)|^2 \leq [\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)]^2 + 2\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)\\ \times\gamma^K(U_K|Q^* - \tilde{Q}_0|)(s,a) + \gamma^{2K}[(U_K|Q^* - \tilde{Q}_0|)(s,a)]^2.\end{aligned}$$

也就是说，

$$E_\mu[|Q^*(s,a)-\tilde{Q}_K(s,a)|^2] \le E_\mu\{[\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)]^2\}$$
$$+E_\mu\{2\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)\times\gamma^K(U_K|Q^*-\tilde{Q}_0|)(s,a)\} \tag{16}$$
$$+E_\mu\{\gamma^{2K}[(U_K|Q^*-\tilde{Q}_0|)(s,a)]^2\}.$$

对于第三项 $E_\mu\{\gamma^{2K}[(U_K|Q^*-\tilde{Q}_0|)(s,a)]^2\}$，我们可以使用以下界

$$E_\mu\{\gamma^{2K}[(U_K|Q^*-\tilde{Q}_0|)(s,a)]^2\} \le 4\gamma^{2K}V_{max}^2 = \frac{4\gamma^{2K}R_{max}^2}{(1-\gamma)^2}. \tag{17}$$

˜等式（17）成立的原因如下：我们知道 $|Q^*(s,a)-Q_0(s,a)| \le 2V_{max}, \forall(s,a)$，由此可得 $|(U_K|Q^*-\tilde{Q}_0|)(s,a)| = |((P^{\pi^*})^K|Q^*-\tilde{Q}_0|)(s,a)| \le 2V_{max}, \forall(s,a)$。

对于（16）中的交叉项，我们有

$$E_\mu\{2\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)\times\gamma^K(U_K|Q^*-\tilde{Q}_0|)(s,a)\} \le$$
$$E_\mu\{[\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)]^2\} + E_\mu\{\gamma^{2K}[(U_K|Q^*-\tilde{Q}_0|)(s,a)]^2\}. \tag{18}$$

因此，主要需要关注的是对 $E_\mu\{[\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)]^2\}$ 的界。我们有

$$E_\mu\{[\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)]^2\} = \sum_{0\le i,j\le K-1}\gamma^{2K-2-i-j}E_\mu[(U_i|\rho_{i+1}|)(s,a)(U_j|\rho_{j+1}|)(s,a)].$$

然后，我们需要处理 $E_\mu[(U_i|\rho_{i+1}|)(s,a)(U_j|\rho_{j+1}|)(s,a)]$，其中 $0 \le i,j \le K-1$。

我们有

$$E_\mu[(U_i|\rho_{i+1}|)(s,a)(U_j|\rho_{j+1}|)(s,a)] = E_\mu[(P^{\pi^*})^{K-1-i}|\rho_{i+1}|(s,a)(P^{\pi^*})^{K-1-j}|\rho_{j+1}|(s,a)].$$

回顾一下，$(P^\pi Q)(s,\ a)\ = E[Q(S',\ A')|S' \sim P(.|s,\ a)\ A' \sim \pi(.|S')]$。有一个关键观察

$$\begin{aligned}[(P^\pi Q)(s,a)]^2 &= E[Q(S',A')|S' \sim P(.|s,a), A' \sim \pi(.|S')] \\ &\times E[Q(S',A')|S' \sim P(.|s,a), A' \sim \pi(.|S')] \\ &\leq E[Q^2(S',A')|S' \sim P(.|s,a), A' \sim \pi(.|S')].\end{aligned}$$

这里我们滥用记号 $P^\pi\mu$ 来表示下一阶段的分布（$S',\ A'$），使得 $S' \sim P(.|s,\ a)$ $A' \sim \pi(.|S')$（当记号 P^π 后接定义在 $\mathcal{S} \times \mathcal{A}$ 上的函数时，它表示方程（15）中定义的算子；当其后接定义在 $\mathcal{S} \times \mathcal{A}$ 上的分布时，它表示此处讨论的分布变换）。

于是可以证明

$$E_\mu\{[(P^\pi Q)(s,a)]^2\} \leq E_\mu\{E[Q^2(S',A')|S' \sim P(.|s,a), A' \sim \pi(.|S')]\} = E_{P^\pi\mu}[Q^2(s,a)].$$

我们还可以得到

$$E_\mu\{[(P^{\pi_1}P^{\pi_2}...P^{\pi_l}Q)(s,a)]^2\} \leq E_{\pi_1 P^{\pi_2}...P^{\pi_l}\mu}[Q^2(s,a)].$$

为简便起见，记 $\pi_1 P^{\pi_2}...P^{\pi_l}\mu$ 为 $\tilde{\mu}_l$。

对于定义在 $\mathcal{S} \times \mathcal{A}$ 上的两个函数 $f_1,\ f_2$，我们知道

$$E_\mu[(P^{\pi^*})^{K-1-i}f_1(P^{\pi^*})^{K-1-j}f_2] \leq \{E_\mu\{[(P^{\pi^*})^{K-1-i}f_1]^2\}E_\mu\{[(P^{\pi^*})^{K-1-j}f_2]^2\}\}^{\frac{1}{2}}.$$

可以得到

$$\begin{aligned}E_\mu\{[(P^{\pi^*})^{K-1-i}f_1]^2\} \leq E_{(P^{\pi^*})^{K-1-i}\mu}[f_1^2] &= \int_{\mathcal{S}\times\mathcal{A}} f_1^2(s,a)d\tilde{\mu}_{K-1-i}(s,a) \\ = \int_{\mathcal{S}\times\mathcal{A}} |f_1^2(s,a)\frac{d\tilde{\mu}_{K-1-i}(s,a)}{d\sigma(s,a)}|d\sigma(s,a) &\leq \int_{\mathcal{S}\times\mathcal{A}} |f_1^2(s,a)|d\sigma(s,a) \times \left\|\frac{d\tilde{\mu}_{K-1-i}(s,a)}{d\sigma(s,a)}\right\|_{\infty,\sigma} \\ &= \|f_1\|_{2,\sigma}^2 \times \omega_\infty(K-1-i;\mu,\sigma).\end{aligned}$$

回顾方程 (2) 中 $\omega_\infty(i;\mu,\ \sigma)$ 的定义。类似地，我们可以有 e

$$E_\mu\{[(P^{\pi^*})^{K-1-j}f_2]^2\} \le \|f_2\|_{2,\sigma}^2 \times \omega_\infty(K-1-j;\mu,\sigma).$$

代入 $f_1 = |\rho_{i+1}|$ 和 $f_2 = |\rho_{j+1}|$，可得

$$\begin{aligned} &E_\mu[(P^{\pi^*})^{K-1-i}|\rho_{i+1}|(s,a)(P^{\pi^*})^{K-1-j}|\rho_{j+1}|(s,a)] \\ \le &\{E_\mu\{[(P^{\pi^*})^{K-1-i}|\rho_{i+1}|]^2\}E_\mu\{[(P^{\pi^*})^{K-1-j}|\rho_{j+1}|]^2\}\}^{\frac{1}{2}} \\ \le &\|\rho_{i+1}\|_{2,\sigma}\|\rho_{j+1}\|_{2,\sigma}[\omega_\infty(K-1-i;\mu,\sigma)\omega_\infty(K-1-j;\mu,\sigma)]^{\frac{1}{2}}. \end{aligned}$$

于是，

$$\begin{aligned} E_\mu\{[\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)]^2\} \le &\sum_{0\le i,j\le K-1}\gamma^{2K-2-i-j}\|\rho_{i+1}\|_{2,\sigma}\|\rho_{j+1}\|_{2,\sigma} \\ &\times[\omega_\infty(K-1-i;\mu,\sigma)\omega_\infty(K-1-j;\mu,\sigma)]^{\frac{1}{2}}. \end{aligned}$$

我们可以定义

$$\epsilon_{max} = \max_{i\in\{1,2...,K\}}\|\rho_i\|_{2,\sigma} = \max_{i\in\{1,2...,K\}}\left\|T\tilde{Q}_{i-1}-\tilde{Q}_i\right\|_{2,\sigma},$$

使得 $\epsilon_{max}^2 = max_{i\in\{1,2...,K\}}\|\rho_i^2\|_{2,\ \sigma}$。因此，可以证明

$$E_\mu\{[\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)]^2\} \le \epsilon_{max}^2\{\sum_{i=0}^{K-1}\gamma^i[\omega_\infty(i;\mu,\sigma)]^{\frac{1}{2}}\}^2.$$

我们已经证明了 $^1{}_{(1-\gamma)^2}\sum_{m=0}^{\infty}\gamma^{m-1}(m+1)[\omega_\infty(m;\ \mu,\ \sigma)]^{\frac{1}{2}} \le \phi_{\mu,\text{-}\sigma}$。于是，我们知道

$$\sum_{i=0}^{K-1}\gamma^i[\omega_\infty(i;\mu,\sigma)]^{\frac{1}{2}} < (1-\gamma)^2\gamma\times\frac{1}{(1-\gamma)^2}\sum_{i=0}^{\infty}\gamma^{i-1}(i+1)[\omega_\infty i;\mu,\sigma]^{\frac{1}{2}} \le (1-\gamma)^2\gamma\phi_{\mu,\sigma}.$$

因此，

$$E_\mu\{[\sum_{i=0}^{K-1}\gamma^{K-1-i}(U_i|\rho_{i+1}|)(s,a)]^2\} < \epsilon_{max}^2(1-\gamma)^4\gamma^2\phi_{\mu,\sigma}^2. \tag{19}$$

然后结合公式 (16)、公式 (17)、公式 (18) 和公式 (19)，即可得到引理 C.1 的结论。

□

C.4 命题 C.2 的证明

命题 C.2 如下：函数类

$$\mathcal{F}_1(L, \{\{d_i\}_{i=1}^{L+1}, d_0 = m + m_0\}, s, V_{max}, \hat{G}, \kappa)$$

是.....的子集

$$\mathcal{F}_2(L, \{\{d_i\}_{i=1}^{L+1}, d_0 = \kappa + m_0\}, s + d_1\kappa, V_{max}, \hat{G}, \kappa).$$

证明。注意，由 z 的前 κ 个主成分分析值恢复得到的向量 z^* 与包含相应主成分分析值的向量之间存在如下关系： $z^* = \hat{U}\hat{\Sigma}_2^1 \begin{pmatrix} I_\kappa \\ 0_{(m-\kappa)\times\kappa} \end{pmatrix} \hat{v}$。根据 A 节中的定义 A.1， $\forall f_1 \in \mathcal{F}_1(L, \{\{d_i\}_{i=1}^{L+1}, d_0 = m + m_0\}, s, V_{max}, \hat{G}, \kappa)$，我们有 $f_1(x, z, a) = f_0^1(x, z^*, a)$，其中 $f_0^1(., a) \in \mathcal{F}_{SReLU}(L, \{d_i\}_{i=0}^{L+1}, s, V_{max})$。我们将稀疏 ReLU 网络 $f_0^1(., a)$ 的第一层记为 $\sigma(W_1^1 \begin{pmatrix} x \\ z^* \end{pmatrix} + b_1^1)$。我们知道 $W_1^1 \begin{pmatrix} x \\ z^* \end{pmatrix} = \begin{pmatrix} W_{11}^1 & W_{12}^1 \end{pmatrix} \begin{pmatrix} x \\ z^* \end{pmatrix} = W_{11}^1 x + W_{12}^1 z^*$，其中

$$W_{12}^1 z^* = W_{12}^1 \hat{U}\hat{\Sigma}_2^1 \begin{pmatrix} I_\kappa \\ 0_{(m-\kappa)\times\kappa} \end{pmatrix} \hat{v}。$$

那么，我们想知道 $\exists f_2 \in \mathcal{F}_2(L, \{\{d_i\}_{j=1}^{L+1}, d_0 = \kappa + m_0\}, s + d_1\kappa, V_{max})$，其中

$$f_0^2(., a) \in \mathcal{F}_{SReLU}(L, \{d_i\}_{i=0}^{L+1}, s + \kappa d_1, V_{max})$$

根据定义。 f_0^2的第一层可以用类似的方式表示： $\sigma(W_1^2 \begin{pmatrix} x \\ \hat{v} \end{pmatrix} + b_1^2)$,

其中 $W_1^2 \begin{pmatrix} x \\ \hat{v} \end{pmatrix} = \begin{pmatrix} W_{11}^2 & W_{12}^2 \end{pmatrix} \begin{pmatrix} x \\ \hat{v} \end{pmatrix} = W_{11}^2 x + W_{12}^2 \hat{v}$。我们可以设定 $W_{11}^2 = W_{11}^1$, $b_1^1 = b_1^2$和$W_{12}^2 = W_{12}^1 \hat{U} \hat{\Sigma}_2^1 \begin{pmatrix} I_\kappa \\ \bar{0}_{(m-\kappa)\times\kappa} \end{pmatrix}$，从而得到 $\sigma(W_1^1 \begin{pmatrix} x \\ z^* \end{pmatrix} + b_1^1) = \sigma(W_1^2 \begin{pmatrix} x \\ \hat{v} \end{pmatrix} + b_1^2)$。$f_0^2$中其他层的参数将设为与 f_0^1相同。

那么我们知道 $\forall f_1 \in \mathcal{F}_1(L, \{\{d_j\}_{j=1}^{L+1}, d_0 = m+m_0\}, s, V_{max}, \hat{G}, \kappa)$、$\exists f_2 \in \mathcal{F}_2(L, \{\{d_j\}_{j=1}^{L+1}, d_0 = \hat{\kappa} + m_0\}, s + d_1\kappa, V_{max}, G, \kappa)$，使得

$$f_1(x, z, a) = f_0^1(x, z^*, a) = f_0^2(x, \hat{v}, a) = f_2(x, z, a).$$

f_0^2中的稀疏度之所以增加到 $s + \kappa d_1$，是因为权重 $W_{12}^2 = W_{12}^1 \hat{U} \hat{\Sigma}_2^1 \begin{pmatrix} I_\kappa \\ \bar{0}_{(m-\kappa)\times\kappa} \end{pmatrix} \in \mathbb{R}^{d_1 \times \kappa}$可能不再稀疏，尽管 W_{12}^1是稀疏的。

□

C.5 引理 C.3 的证明

引理 C.3 如下：在与定理 4.6 相同的假设下,

$$\sup_{g\in\mathcal{F}_2} \inf_{f\in\mathcal{F}_2} \|f - Tg\|_\sigma^2 \le 2 \sup_{f'\in\mathcal{G}_0} \inf_{f_0\in\mathcal{F}_0} \left\| f_0 - f' \right\|_\sigma^2 + 2L_{\mathcal{F}_0} \|Z - Z^*\|_\sigma^2,$$

其中 $L_{\mathcal{F}_0} = \sup_{f\in\mathcal{F}_0} \sup_{x\neq y} \frac{|f_{(y-f\,x\,|^2})()}{\|y-x\|^2}$。$Z$ 是分布σ 中的高频向量，而 $Z^* = (\sum_{k=1}^{\kappa_j} \hat{U}_k \hat{U}_k^T)Z$ 是根据 Z的 κ 个主成分分析值恢复得到的向量。

证明。我们有 $\sup_{g\in\mathcal{F}_2} \inf_{f\in\mathcal{F}_2} \|f - Tg\|_\sigma^2 \le \sup_{f'\in\mathcal{G}_0} \inf_{f_2\in\mathcal{F}_2} \|f_2 - f'\|_\sigma^2$，这依赖于以下假设：$Tg \in \mathcal{G}_0$, $\forall g \in \mathcal{F}_2$。根据命题 C.2，我们有 $\mathcal{F}_1 \subset \mathcal{F}_2$。于是我们有 $\sup_{f'\in\mathcal{G}_0} \inf_{f_2\in\mathcal{F}_2} \|f_2 - f'\|_\sigma^2 \le \sup_{f'\in\mathcal{G}_0} \inf_{f_1\in\mathcal{F}_1} \|f_1 - f'\|_\sigma^2$。

于是我们知道

$$\sup_{f'\in\mathcal{G}_0}\inf_{f_1\in\mathcal{F}_1}\|f_1-f'\|_\sigma^2=\sup_{f'\in\mathcal{G}_0}\inf_{f_1\in\mathcal{F}_1}E_\sigma\{[f_1(x,z,a)-f'(x,z,a)]^2\}$$
$$=\sup_{f'\in\mathcal{G}_0}\inf_{f_0\in\mathcal{F}_0}E_\sigma\{[f_0(x,z^*,a)-f'(x,z,a)]^2\}$$

根据 $\mathcal{F}_1$ 的定义。

我们知道

$$E_\sigma\{[f_0(x,z^*,a)-f'(x,z,a)]^2\}=E_\sigma\{[f_0(x,z^*,a)-f_0(x,z,a)+f_0(x,z,a)-f'(x,z,a)]^2\}$$
$$\le 2E_\sigma\{[f_0(x,z^*,a)-f_0(x,z,a)]^2\}+2E_\sigma\{[f_0(x,z,a)-f'(x,z,a)]^2\}.$$

此外，项 $E_\sigma\{[f_0(x,z^*,a)-f_0(x,z,a)]^2\}$ 满足

$$E_\sigma\{[f_0(x,z^*,a)-f_0(x,z,a)]^2\}\le L_{\mathcal{F}_0}\|Z-Z^*\|_\sigma^2.$$

注意，$L_{\mathcal{F}_0}$ 是类 $\mathcal{F}_0$ 中 ReLU 网络的 Lipschitz 常数。根据链式法则以及对于具有 m_W 列的矩阵 W，$\|W\|_2^2\le m_W\|W\|_\infty^2$ 这一事实，我们知道 $L_{\mathcal{F}_0}$ 总是有界的：$L_{\mathcal{F}_0}\le\prod_l^L d_{l-1}$。

因此我们知道

$$E_\sigma\{[f_0(x,z^*,a)-f'(x,z,a)]^2\}\le 2L_{\mathcal{F}_0}\|Z-Z^*\|_\sigma^2+2E_\sigma\{[f_0(x,z,a)-f'(x,z,a)]^2\}$$

$\forall f_0\in\mathcal{F}_0, f'\in\mathcal{G}_0$。然后，我们可以对两边分别取上确界和下确界，得到

$$\sup_{f'\in\mathcal{G}_0}\inf_{f_0\in\mathcal{F}_0}E_\sigma\{[f_0(x,z^*,a)-f'(x,z,a)]^2\}\le 2L_{\mathcal{F}_0}\|Z-Z^*\|_\sigma^2+$$
$$2\sup_{f'\in\mathcal{G}_0}\inf_{f_0\in\mathcal{F}_0}E_\sigma\{[f_0(x,z,a)-f'(x,z,a)]^2\},$$

其中 $E_\sigma\{[f_0(x,z,a)-f'(x,z,a)]^2\}=\|f_0-f'\|_\sigma^2$。

因此最终得到

$$\sup_{g\in\mathcal{F}_2}\inf_{f\in\mathcal{F}_2}\|f-Tg\|_\sigma^2\le 2L_{\mathcal{F}_0}\|Z-Z^*\|_\sigma^2+2\sup_{f'\in\mathcal{G}_0}\inf_{f_0\in\mathcal{F}_0}\|f_0-f'\|_\sigma^2.$$

□

C.6 引理 C.4 的证明

引理 C.4 如下：对于高频向量 Z，我们假设 $E(Z) = 0$，并且 $Cov(Z)$ 的特征值遵循以下指数衰减趋势 $\lambda_j = O(e^{-\zeta j})$。我们还假设 $Cov(Z)$ 的估计满足 $\left\|\hat{U}_j - U_j\right\| = O_p(n^{-\Delta})$。于是我们有

$$\|Z - Z^*\|_\sigma^2 = O_p(e^{-\zeta\kappa} - e^{-\zeta m} + n^{-2\Delta})$$

证明。 注意到

$$\|Z - Z^*\|_\sigma^2 = E[(Z - Z^*)^T(Z - Z^*)] = E[Z^T(\sum_{j=\kappa+1}^{m} \hat{U}_j\hat{U}_j^T)^2 Z] = \sum_{j=\kappa+1}^{m} E[tr(Z^T\hat{U}_j\hat{U}_j^T Z)]$$
$$= \sum_{j=\kappa+1}^{m} tr[E(ZZ^T)\hat{U}_j\hat{U}_j^T] = tr[E(ZZ^T)\sum_{j=\kappa+1}^{m} \hat{U}_j\hat{U}_j^T].$$

我们知道 $E(ZZ^T) = Cov(Z) = \sum_{i=1}^{m} \lambda_i U_i {U_i}^T$。因此我们有

$$\|Z - Z^*\|_\sigma^2 = tr[\sum_{i=1}^{m} \lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m} \hat{U}_j\hat{U}_j^T] = tr[\sum_{i=1}^{m} \lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m} (\hat{U}_j - U_j + U_j)(\hat{U}_j - U_j + U_j)^T]$$
$$= tr[\sum_{i=1}^{m} \lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m} U_j {U_j}^T] + tr[\sum_{i=1}^{m} \lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m} (\hat{U}_j - U_j){U_j}^T]$$
$$+tr[\sum_{i=1}^{m} \lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m} U_j(\hat{U}_j - U_j)^T] + tr[\sum_{i=1}^{m} \lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m} (\hat{U}_j - U_j)(\hat{U}_j - U_j)^T].$$

也就是说，我们需要对这里的四项进行界定。

第一项

$$tr[\sum_{i=1}^{m} \lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m} U_j {U_j}^T] = tr[\sum_{j=\kappa+1}^{m} \lambda_j U_j {U_j}^T U_j {U_j}^T] = \sum_{j=\kappa+1}^{m} tr(\lambda_j U_j {U_j}^T) = \sum_{j=\kappa+1}^{m} \lambda_j.$$

由于我们已经知道特征值 λ_j 满足 $\lambda_j = O(e^{-\zeta j})$，因此有 $\sum_{j=\kappa+1}^{m} \lambda_j = O(e^{-\zeta(\kappa+1)}\frac{1-e^{-\zeta(m-\kappa)}}{1-e^{-\zeta}}) = O(e^{-\zeta\kappa} - e^{-\zeta m})$。于是我们知道第一项

$$tr[\sum_{i=1}^{m} \lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m} U_j {U_j}^T] = O(e^{-\zeta\kappa} - e^{-\zeta m}).$$

第二项为

$$tr[\sum_{i=1}^{m}\lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m}(\hat{U}_j - U_j){U_j}^T] = \sum_{i=1}^{m}\sum_{j=\kappa+1}^{m} tr[\lambda_i U_i {U_i}^T(\hat{U}_j - U_j){U_j}^T]$$
$$= \sum_{i=1}^{m}\sum_{j=\kappa+1}^{m} tr[\lambda_i {U_j}^T U_i {U_i}^T(\hat{U}_j - U_j)] = \sum_{j=\kappa+1}^{m}\lambda_j {U_j}^T(\hat{U}_j - U_j).$$

我们知道， U_j的每一项都是 $O_p(1)$，因此 $|{U_j}^T(\hat{U}_j - U_j)| \leq \left\|{U_j}^T\right\| \left\|\hat{U}_j - U_j\right\| = O_p(n^{-\Delta})$，且 $\sum_{j=\kappa+1}^{m}\lambda_j {U_j}^T(\hat{U}_j - U_j) = O_p(n^{-\Delta}(e^{-\zeta\kappa} - e^{-\zeta m}))$。

第三项与第二项类似，因此有

$$tr[\sum_{i=1}^{m}\lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m} U_j\hat{U}_j - {U_j}^T] = \sum_{i=1}^{m}\sum_{j=\kappa+1}^{m} tr[\lambda_i U_i {U_i}^T U_j(\hat{U}_j - U_j)^T]$$
$$= \sum_{j=\kappa+1}^{m}\lambda_j(\hat{U}_j - U_j)^T U_j = O_p(n^{-\Delta}(e^{-\zeta\kappa} - e^{-\zeta m})).$$

那么第四项为

$$tr[\sum_{i=1}^{m}\lambda_i U_i {U_i}^T \sum_{j=\kappa+1}^{m}(\hat{U}_j - U_j)(\hat{U}_j - U_j)^T] = \sum_{i=1}^{m}\sum_{j=\kappa+1}^{m} tr[\lambda_i U_i {U_i}^T(\hat{U}_j - U_j)(\hat{U}_j - U_j)^T]$$
$$= \sum_{i=1}^{m}\sum_{j=\kappa+1}^{m} tr[\lambda_i {U_i}^T(\hat{U}_j - U_j)(\hat{U}_j - U_j)^T U_i] = \sum_{i=1}^{m}\sum_{j=\kappa+1}^{m}\lambda_i[{U_i}^T(\hat{U}_j - U_j)]^2$$
$$= O_p(n^{-2\Delta}e^{-\zeta}) = O_p(n^{-2\Delta}).$$

因此，可以证明

$$\|Z - Z^*\|_\sigma^2 = O_p((e^{-\zeta\kappa} - e^{-\zeta m}) + n^{-\Delta}(e^{-\zeta\kappa} - e^{-\zeta m}) + n^{-2\Delta}) = O_p(e^{-\zeta\kappa} - e^{-\zeta m} + n^{-2\Delta}).$$

□

C.7 命题 C.5 的证明

命题 C.5 为：当 $\delta = \frac{1}{n}$ 时，$\log N_{\delta,\ 2} \leq C_1|\mathcal{A}|(n^{\alpha^*}(\log\ n)^{\xi^*} + d_1^*\kappa)(\log\ n)^{1+\xi^*}$ 对于某个常数 C_1。

证明。为证明这一结果，我们尝试证明对于函数类 $\mathcal{F}_3 = \{f_3 :\ S \to \mathbb{R} : f_3(x,z) = f(x, \left(I_\kappa \quad 0_{\kappa\times(m-\kappa)}\right)\hat{\Sigma}^{-\frac{1}{2}}\hat{U}^T z) = f(x,\hat{v})\ f \in \mathcal{F}_{SReLU}(L^*, \{d_j^*\}_{j=0}^{L+1}, s^*, V_{max})\ d_0 = \kappa + m_0\}$。回顾 A 节中 $\mathcal{F}_2$ 的定义，我们知道 $\forall f \in \mathcal{F}_2, f(.,a) \in \mathcal{F}_3$。也就是说，$\mathcal{F}_3$ 是用于对 $Q(s,a)$ 的动作分量 a 建模的函数类。$\mathcal{F}_2, N_{\frac{1}{n}},2$ 的最小 σ- 覆盖集的基数与 $\mathcal{F}_3, N_{\frac{1}{n}},3$ 的最小 σ- 覆盖集的基数之间存在关系。采用与 Fan 等人（2020）方程 (6.26) 中步骤 $|\mathcal{N}(\delta, \mathcal{F}_0, \|.\|_\infty)| \leq |\mathcal{N}_\delta|^{|\mathcal{A}|}$ 类似的论证，我们可以证明

$$\log N_{\frac{1}{n},2} \leq |\mathcal{A}| \log N_{\frac{1}{n},3}. \tag{20}$$

这里可以说明，$\mathcal{F}_3$ 也是一种稀疏 ReLU 网络，因此可以直接使用 Fan 等人（2020）中的引理 6.4。我们可以证明 $\mathcal{F}_3 \subset \mathcal{F}_{SReLU,\ 3}$，其中 $\mathcal{F}_{SReLU,\ 3}$ 是 $\mathcal{F}_{SReLU}(L^*+1, \{\{d_i\}_{i=0}^{L+2} :\ d_0 = m,\ d_{i+1} = d_i^*,\ i \in \{0,1,2,...,L+1\},\ s^* + m\kappa,\ V_{max}\}$ 的简写。也就是说，$\mathcal{F}_3$ 多出的这一层实际上是线性变换 $\hat{v} = \left(I_\kappa \quad 0_{\kappa\times(m-\kappa)}\right)\hat{\Sigma}^{-\frac{1}{2}}\hat{U}^T z$。我们不知道权重矩阵 $\left(I_\kappa \quad 0_{\kappa\times(m-\kappa)}\right)\hat{\Sigma}^{-\frac{1}{2}}\hat{U}^T$是否稀疏，因此需要将原始稀疏度提高到 $s^* + m\kappa$。

稀疏 ReLU 网络的 δ-覆盖集的基数已由 Fan 等人（2020）的引理 6.4 给出界，因此我们可以在此应用该结果。我们知道 $\mathcal{F}_3, N_{\frac{1}{n},3}$ 的 δ-覆盖集的基数满足

$$\log N_{\frac{1}{n},3} \leq \log|\mathcal{N}(\delta = \frac{1}{n}, \mathcal{F}_{SReLU,3}, \|.\|_\infty)| \leq (s^* + m\kappa + 1)\log\left[2n(L^*+2)\prod_{l=0}^{L+1} d_l^* + 1\right]$$

.

这里 $L^* = O((\log n)^{\xi^*}), m + m_0 \leq min_{j\in\{1,2,\ldots,L\}} d_j^* \leq max_{j\in\{1,2,\ldots,L\}} d_j^* = O(n^{\xi^*})$ 且 $s^* \asymp n^{\alpha^*}(\log n)^{\xi^*} + d_1^*\kappa$，其中 $\xi^* > 1 + 2\xi$ 为某个常数。因此

$$\begin{aligned}
\log N_{\frac{1}{n},3} &\leq C(n^{\alpha^*}(\log n)^{\xi^*} + d_1^*\kappa + m\kappa)\log\left[2n(L^*+2)\prod_{l=0}^{L+1} d_l^* + 1\right] \\
&\leq C'(n^{\alpha^*}(\log n)^{\xi^*} + d_1^*\kappa)\log\left[2n(L^*+2)\prod_{l=0}^{L+1} d_l^* + 1\right] \\
&\leq C''(n^{\alpha^*}(\log n)^{\xi^*} + d_1^*\kappa)[\log n + \log L^* + L^*\log(n^{\xi^*})] \\
&\leq C_1(n^{\alpha^*}(\log n)^{\xi^*} + d_1^*\kappa)(\log n)^{\xi^*+1},
\end{aligned}$$

其中 C，C'，C''，C_1 均为常数。

然后，根据公式（20），我们知道 $\log N_{\frac{1}{n},2} \leq C_1|\mathcal{A}|(n^{\alpha^*}(\log n)^{\xi^*} + d_1^*\kappa)(\log n)^{\xi^*+1}$。 □

D 选择主成分数的详细说明

使偏差和方差达到平衡的最优主成分数 κ^* 的阶为 $\kappa^* \asymp \log(n)$。

证明。我们知道，定理 4.8 中涉及 κ 的项为 $\kappa|\mathcal{A}|d_1^* n^{-1}\log^{\xi^*+1} n + CL_{\mathcal{F}_0}e^{-\zeta\kappa}$，其中 $C > 0$ 为某个常数。我们可以对该项关于 κ 求导。由此可知，当 $\kappa^* = \frac{1}{\zeta}[\log(C\zeta L_{\mathcal{F}_0}|\mathcal{A}|^{-1}d_1^{*-1}) + \log n - (\xi^*+1)\log(\log n)]$, 时，该项将达到最小值。因此，使偏差和方差达到平衡的最优 κ^* 必须满足$\kappa^* \asymp \log(n)$。

当 n 较大时，适用此 $\log n$ 指南，因为它是基于渐近趋势的定理推导得出的。另一方面，$\kappa \leq m$ 必须成立，因为主成分的数量不能超过原始维度。因此，当 n 趋于无穷大时，

最终我们会取 $\kappa^* = m$。也就是说，当训练样本量极大时，我们不再需要对高频部分进行降维，此时算法 1 将等同于原始的神经网络拟合 Q 迭代。

□

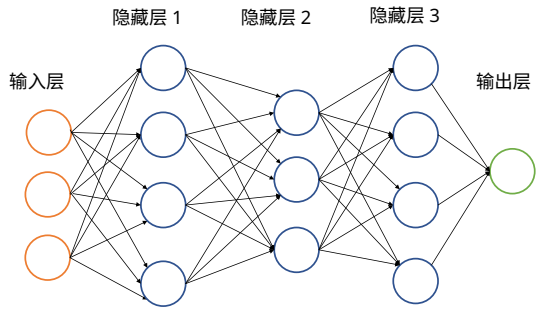

图 1：ReLU 网络示意图。

表 1：四种策略的价值差异估计。

Difference	$\pi_k^{PCA} - \pi_k^{ALL}$	$\pi_k^{PCA} - \pi_k^{AVE}$	$\pi_k^{PCA} - \pi_k^{BOTTLE}$
Mean	1.163	1.399	1.153
Margin of Error	0.181	0.224	0.145

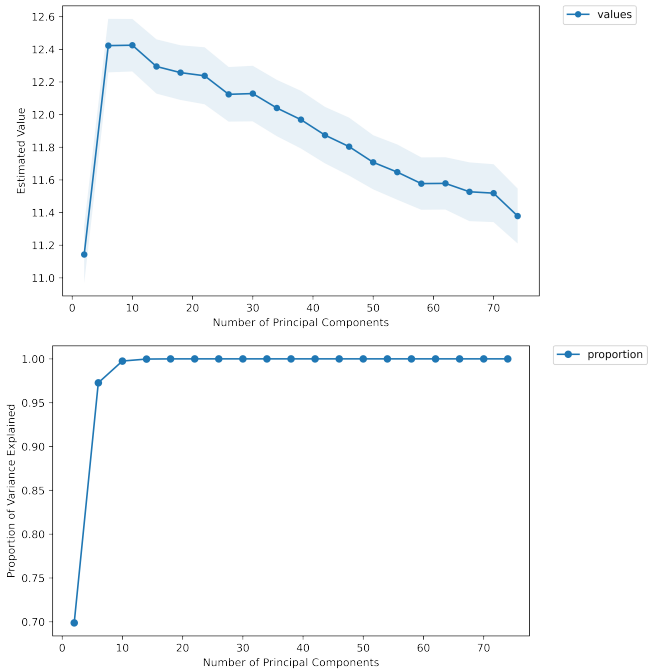

图2：**左图**：当 κ 在 $\{2, 6, 10, 14..., 74\}$中变化时，算法 1 得出的策略价值估计（阴影区域为95%置信区间）；**右图**：前 κ 个主成分解释的方差比例

表2：四种策略的价值估计。

Value Estimate	π_k^{PCA}	π_k^{ALL}	π_k^{AVE}	π_k^{BOTTLE}
Mean	-8.762	-9.926	-10.162	-9.916
Standard Error	0.055	0.096	0.121	0.086

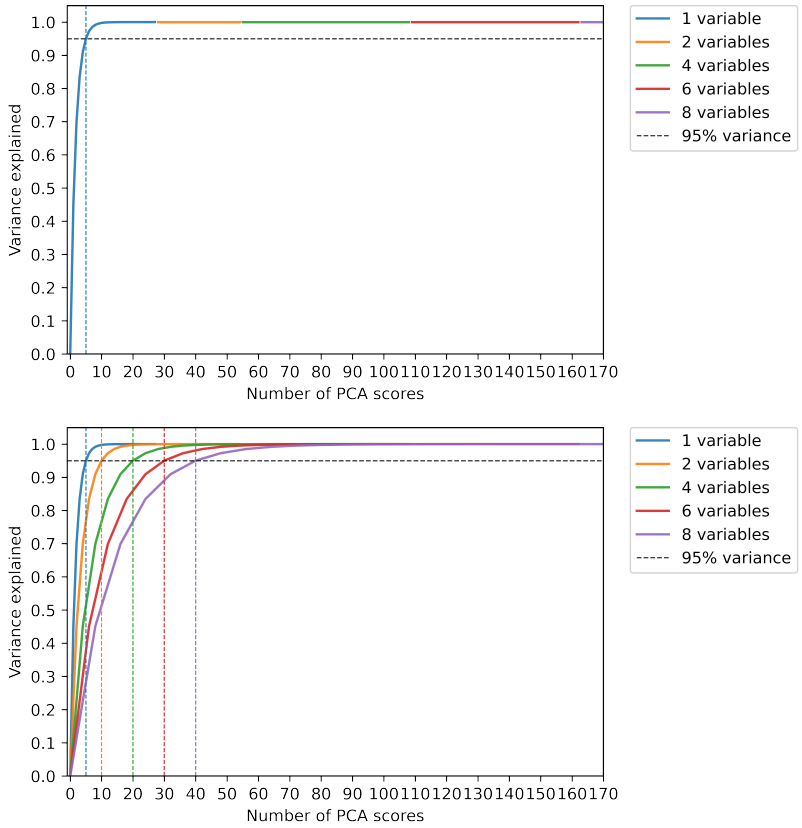

图3：**上图**：模拟的**第一种设定** 中，当高频变量数量为$J = 1$、2、4、6、8时，前 κ 个主成分得分解释的方差。水平虚线表示解释方差的95%。 $\kappa = 5$ 在第一种设定的所有 $J = 1$、2、4、6、8种情况下均可解释95%的方差；**下图**：模拟的**第二种设定** 中，当高频变量数量为 $J = 1$、2、4、 6、8时，前 κ 个主成分得分解释的方差（对应图5）。这里，五种情况 $J = 1$、2、4、6、 8分别对应的主成分数量为 $\kappa = 5$、 10、 20、30、40，以确保解释95%的方差。这里，拼接后的高频 Z 的特征值按指数阶 $\lambda_k = O(e^{-\zeta k})$衰减，其中 $\zeta = 0.6$。

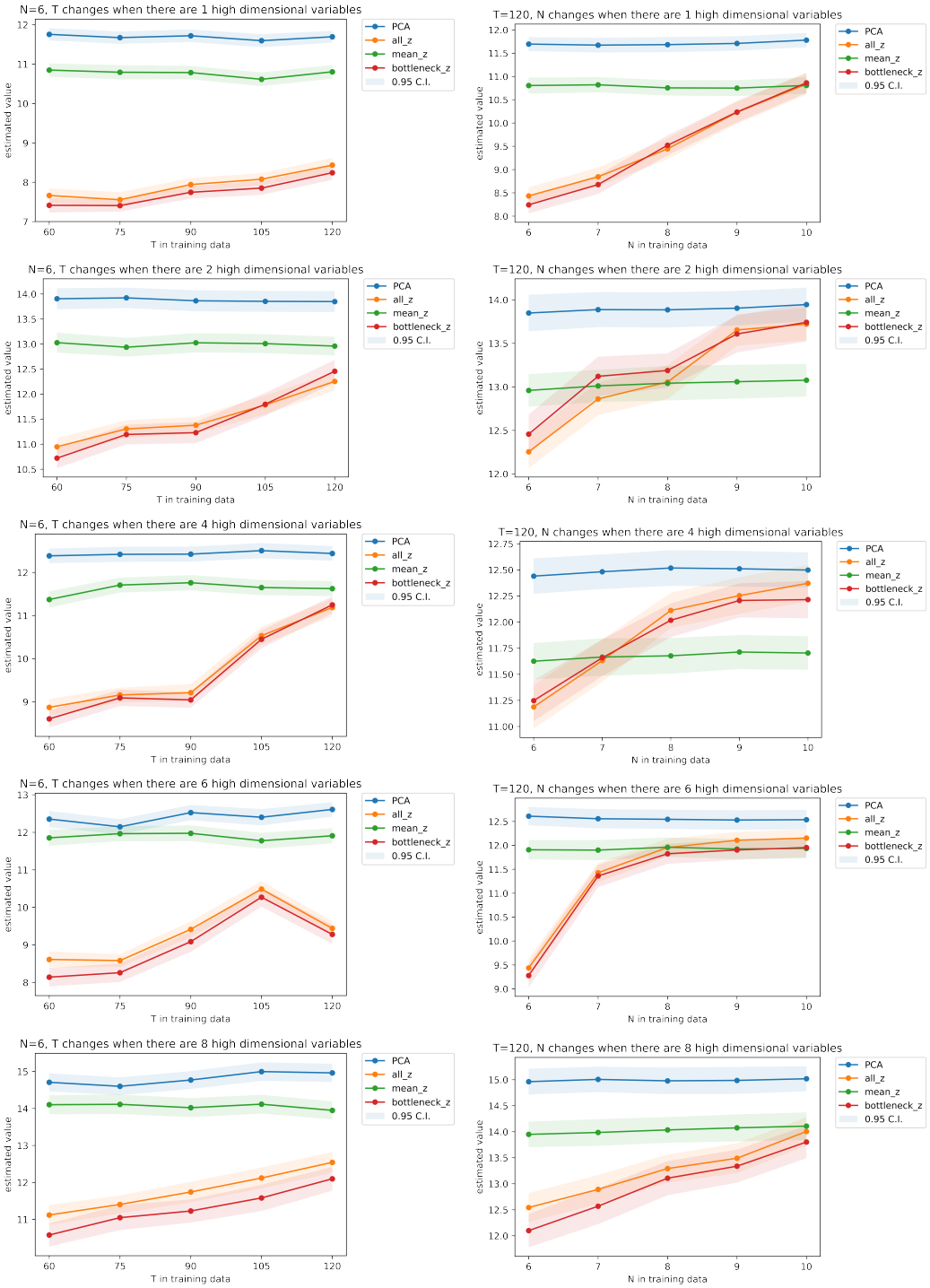

图 4：**模拟中的第一种设定。左图**：训练数据为 $N = 6$ 和 T =60、75、90、105、120；**右图**：当维度为 27 时，变量数为 1、2、4、6、8，训练数据为 $T = 120$ 和 $N = 6$、7、8、9、10（阴影区域为 95%置信区间）。图例中，“PCA”指 π_K^{PCA}；“全部 z”指 π_K^{ALL}；“瓶颈 z”指 π_K^{BOTTLE}；“均值 z”指 π_K^{AVE}。

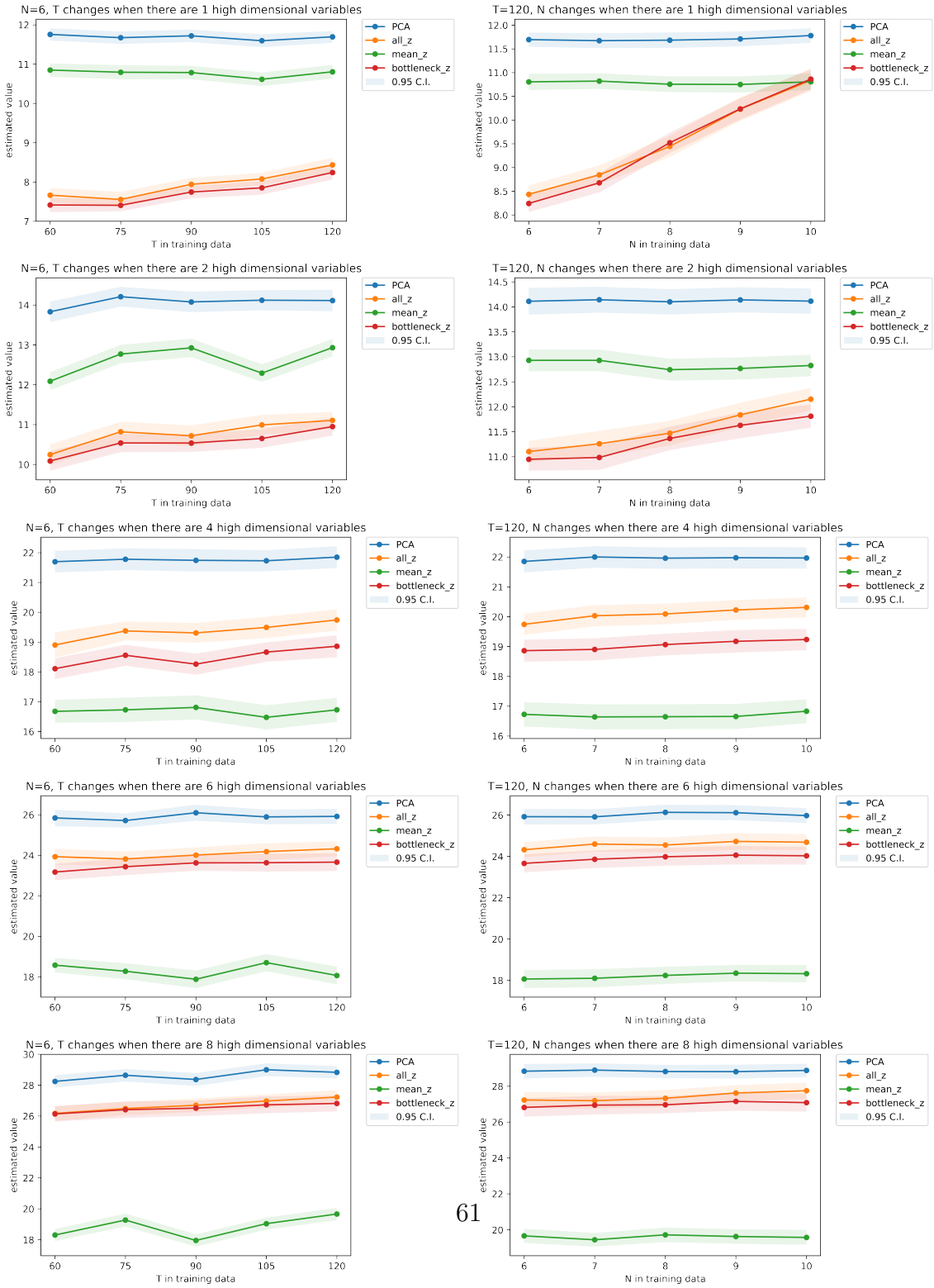

图5：**模拟中的第二种设定。左图**：训练数据为 $N = 6$和 T =60、75、90、105、120；**右图**：训练数据为 $T = 120$ 和 $N = 6$、7、8、9、10，变量数量为 1、2、4、6、8，维度为 27（阴影区域为 95%置信区间）。图例中，“主成分分析”指 π_K^{PCA}；“全部 z”指 π_K^{ALL}；“瓶颈 z”指 π_K^{BOTTLE}；“均值 z”指 π_K^{AVE}。

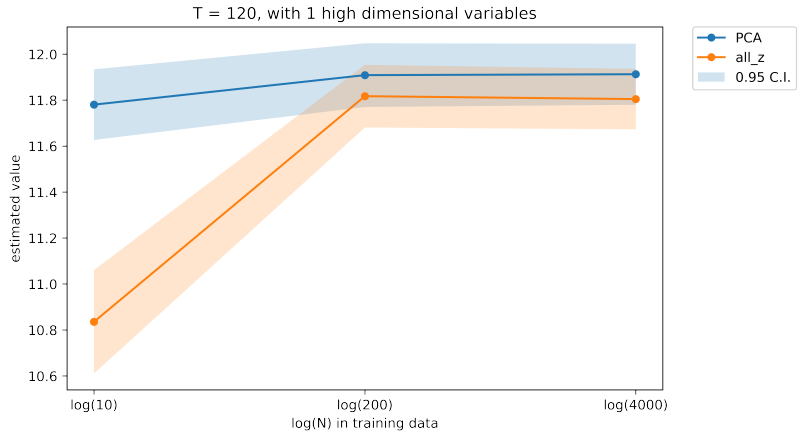

图6：随着训练数据大小 n 增至极大值时， π^{PCA} 和 π^{ALL} 的表现（T固定为 120， N 从 10 到 200 和 4000）。横轴表示 log N。图中阴影区域为 95%置信区间。图例中，“主成分分析”指 π_K^{PCA}；“全部 z”指 π_K^{ALL}。

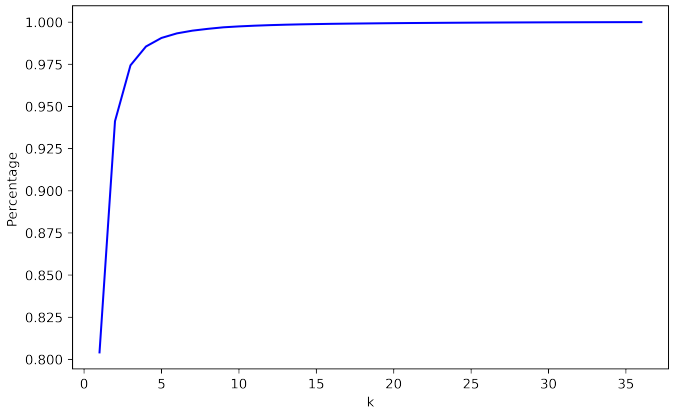

图 7：OhioT1DM 数据中连续血糖监测血糖水平的前 κ 个主成分所解释的方差比例。