arXiv:2301.02220v1 [stat.ML] 5 Jan 2023

通过高效且稳健的信赖域优化实现强化学习的价值增强

Chengchun Shi[a]*、Zhengling Qi[b]*†、Jianing Wang[c] 和 Fan Zhou[c]†[a]
统计系，伦敦政治经济学院[b]决策科学系，乔治·华盛顿大学[c]统计系，上海财经大学

摘要

强化学习（RL）是一种强大的机器学习技术，使智能体能够在序贯决策中学习最大化累积奖励的最优策略。现有文献中的大多数方法都是在在线场景下开发的，在这些场景中，数据易于收集或模拟。受到移动健康研究等高风险领域的启发——这些领域中的数据有限且为预先收集的数据——本文研究离线强化学习方法。为了高效利用这些数据集进行策略优化，我们提出一种新颖的价值增强方法，用于提升由现有最先进的强化学习算法计算得到的给定初始策略的性能。具体而言，当初始策略不一致时，我们的方法将输出一个价值不低于初始策略且通常优于初始策略的策略。当初始策略一致时，在一些温和条件下，我们的方法将以快于初始策略的速率得到一个价值收敛于最优值的策略，从而实现理想的“价值增强”性质。所提出的方法通常适用于属于某个预先指定函数类（例如深度神经网络）的任意参数化策略。我们开展了广泛的数值研究，以证明该方法的优越性能。

关键词：离线强化学习；信赖域优化；半参数效率；移动健康研究。

*前两位作者对本文贡献相同。

†qizhengling@gwu.edu;zhoufan@mail.shufe.edu.cn

1 引言

强化学习（RL，参见 Sutton and Barto，2018年）关注智能体如何在动态环境中采取序贯行动，其主要目标是最大化它们获得的累积奖励。近年来，强化学习在人工智能（AI）领域取得了巨大成就。例如，$AlphaGo$ （Silver 等，2016年）是人工智能领域最成功的应用之一，它利用强化学习和深度学习算法训练机器下围棋，并击败了许多顶尖人类棋手。强化学习的卓越表现也在许多科学领域得到彰显。在医疗应用中，强化学习已被用于帮助临床医生为脓毒症患者制定更好的治疗决策（Komorowski 等，2018年）。在经济学领域，计量经济学家经常研究动态离散选择模型（Rust，1987年），以了解理性智能体的行为，这与逆强化学习问题（Abbeel 和 Ng，2004年）类似。在运筹学领域，强化学习已广泛应用于商业运营，例如供应链管理、金融和物流（Hubbs 等，2020年）。有关强化学习各种应用的概述，请参阅 Li（2017）的第5节。

本文的研究部分受到近期兴起的移动健康（mHealth）研究的启发。移动技术和传感器技术的进步为我们提供了独特的机会，使我们能够随时随地提供健康干预，以促进健康行为（如定期进行体育活动）并预防药物滥用等。例如，OhioT1DM 数据集（Marcolino 等，2018年）旨在促进血糖水平预测，从而改善 1 型糖尿病患者的健康状况和福祉。该数据集包含 6 人为期 8 周的数据。针对每位患者，在胰岛素泵治疗期间，通过连续血糖监测（CGM）收集其治疗信息。此外，还通过定制智能手机应用持续测量并记录血糖水平以及患者自我报告的进餐和锻炼时间。针对不同场景为每位患者寻找最优的胰岛素泵注射策略，可能有助于改善其健康状况（Shi、Zhang、Lu 和 Song，2020年）。这与强化学习算法的目标相符。

我们在此旨在探究的一个根本问题是：如何从移动健康等高风险领域的批量数据中高效地学习最优策略。解决这一问题至少面临两项重大挑战。首先，与标准临床试验数据不同，

移动健康数据通常包含每位患者的大量决策点，但患者人数可能有限（例如，在OhioT1DM 数据集中，只有 6 位患者，却有数千个决策点）。这给最优策略的搜索带来了独特挑战。在统计学领域，已有大量文献研究动态治疗方案（DTR，例如 Murphy，2003；Chakraborty 和 Moodie，2013；Qian 和 Murphy，2011；Zhao 等人，2015；Shi 等人，2018；Wang 等人，2018）。有关 DTR 的综述，参见 Laber 等人（2014）、Kosorok 和 Laber（2019）以及 Tsiatis 等人（2019）。然而，这些方法主要针对数量较少的治疗决策点而设计，而且通常要求观测数据中有大量患者，才能保证一致性。

其次，与电子游戏等在线强化学习领域不同，在这些领域中可以主动与环境交互，且数据易于生成或模拟；而在一些高风险领域，数据通常是根据某种实验设计预先收集的，且十分有限。面对如此有限的数据，研究如何从批量数据中高效地学习最优策略至关重要。值得指出的是，计算机科学文献中的强化学习主要关注在线学习。在现有方法中，Q学习无疑是最受欢迎的无模型强化学习算法之一（Watkins and Dayan，1992）。它通过学习最优Q函数来推导最优策略（参见第 2.2 节中Q函数的定义）。沿着这一研究路线，人们提出了多种Q学习方法的变体，包括拟合Q迭代（FQI，Ernst et al.，2005；Fan et al.，2020a）、深度Q网络（DQN，Mnih et al.，2015）等。基于策略的学习是另一类强化学习算法，它在参数化策略类中搜索最优策略。一些流行的算法包括REINFORCE和演员-评论家方法（例如，参见Sutton and Barto，2018，第13章）。由于这些算法主要受在线电子游戏中开发人工智能这一应用需求的推动，因此将它们推广到移动健康应用等离线场景的研究仍然十分有限。

为应对第一个挑战，我们使用时齐马尔可夫决策过程（MDP；Puterman，1994年）对观测数据进行建模。这一框架尤其适用于对移动健康研究所收集的数据建模，因为这类研究中的决策点总数通常较多（例如，Liao 等，2019年、2020年）。马尔可夫假设和时间齐次性假设使得即使只有少量患者，也能一致地估计最优策略。

为应对第二项挑战，我们开发了一种新颖的方法来推导最优

策略。近期，统计学文献中提出了一些用于移动健康应用中策略优化的算法（Ertefaie 和 Strawderman，2018；Luckett等人，2020；Liao 等人（2020）；Hu 等人，2021）。特别是，Liao 等人（2020）提出了一种在平均奖励马尔可夫决策过程中实现统计效率的批量策略学习方法。然而，由于干扰函数（例如 Q函数和边际密度比）具有依赖于策略的结构，他们提出的算法计算效率较低，因为在其策略梯度下降算法的每次迭代中，都需要更新干扰函数的估计。我们不针对策略优化提出特定算法，而是设计了一种“价值增强”方法，可普遍用于由某种最先进的强化学习算法计算得到的任意初始策略，以提升其性能。具体而言，在采用某种计算高效的强化学习算法并获得初始策略后，我们通过高效估计价值增强分量（定义见第2.3节）并求解一个受约束的优化问题，对该策略进行单步更新。这样既能利用现有最先进的强化学习算法的计算效率，又无需迭代更新干扰函数。更重要的是，所提出的方法保证：当初始策略不一致时，所提出算法输出的策略不会比初始策略差，而且通常会更好。如果初始策略一致，我们的方法将得到一个价值以更快速度收敛至最优值的策略，从而实现预期的“价值增强”特性。近期，计算机科学文献中，Kallus 和 Uehara（2020）提出了一种离线策略梯度算法，该算法考虑了对策略梯度进行统计高效估计。我们的方案与其不同之处在于，我们致力于开发一种通用的价值增强工具，可应用于任何现有的强化学习算法以提升其性能。

我们的方法受到 Kakade 和 Langford（2002）中的引理 6.1 以及 Schulman 等人（2015）提出的信赖域策略优化算法的启发，该算法最初是为在线设定设计的。一个关键观察是，任意两个策略之间的价值差可以分解为一阶分量和高阶余项。基于对高阶项的下界，可以为任意策略的价值函数构造一个下界函数。使用这种下界函数的一大优势在于，它能够从本质上解耦干扰函数中依赖于策略的结构。这确保了所提出算法的计算效率。

关键的“价值增强”性质在很大程度上依赖于对分解中的一阶项进行统计上高效的估计。在在线场景中，Schulman 等人（2015）提出通过模拟数据轨迹来估计这一量。然而，在离线设定中，如何基于观测数据有效评估这一量仍是未知的。借助半参数统计学，我们为一阶项开发了一种三重稳健估计量；与初始策略相比，该估计量被证明能够达到效率界。通过优化所提出的估计量，我们能够提升初始策略的价值。三重稳健性保证，即使某些干扰函数模型设定错误，“价值增强”性质仍然成立。半参数效率则保证价值能够以足够快的速率得到提升。这确保了所提出算法的统计效率，而这对于离线设定是必要的。

在理论方面，我们在对干扰函数估计量提出温和条件的情况下，建立了价值提升性质。具体而言，我们仅要求它们以非参数速率收敛。详情请参见第4节。这一良好性质主要得益于我们组合这些干扰函数估计量的创新方式，从而使三重稳健估计量的一阶项具有参数收敛速率。此外，我们指出，所有与估计相关的理论结果都是基于决策点总数建立的，因此表明，即使轨迹数量较少而每条轨迹的长度较大，所提出的方法仍具有普适性；这种情况在移动健康应用中十分常见。

本文其余部分安排如下。第2节将在时齐马尔可夫决策过程框架下介绍离线强化学习问题，并综述信赖域算法。第3节将介绍我们提出的价值增强策略优化方法及相关估计。第4节将研究我们算法的统计性质。第5节将开展广泛的数值研究，包括一个展示价值提升性质的玩具示例、一项真实数据驱动的模拟研究以及一项真实数据应用，以展示所提方法的优越性能。最后，我们将在第6节总结全文。所有技术证明和详细内容均见在线补充材料。

2 预备知识

本节安排如下。我们首先在第 2.1 节介绍离线数据结构并描述模型设定。在第 2.2 节中，我们介绍推导所提方法所需的一些符号。在第 2.3 节中，我们综述 Schulman 等人（2015）为在线学习提出的信赖域策略优化（TRPO）方法，因为它与我们的方法密切相关。

2.1 价值函数与最优策略

考虑单条轨迹 $\{(S_t, A_t, R_t)\}_{t\geq 0}$，其中 (S_t, A_t, R_t) 表示在时间 t收集的状态—动作—奖励三元组。我们分别用 $\mathcal{S}$ 和 $\mathcal{A}$ 表示状态空间和动作空间。我们假设 $\mathcal{S}$ 和 $\mathcal{A}$ 是离散的，且奖励 R_t 一致有界。假设状态空间为离散空间仅是为了简化表述和理论分析。我们提出的方法同样适用于具有连续状态空间的设定。观测数据由 N 条轨迹组成，对应于 N 个关于 $\{(S_t, A_t, R_t)\}_{t\geq 0}$的独立同分布副本。对于任意 $i = 1, \cdots, N$，从第 i条轨迹收集的数据可概括为 $\{(S_{i,t}, A_{i,t}, R_{i,t}, S_{i,t+1})\}_{0\leq t<T_i}$，其中 T_i 表示第 i条轨迹的终止时间。

策略定义了智能体在每个决策时刻选择动作的方式。历史依赖策略 π 是由一系列决策规则 $\{\pi_t\}_{t\geq 0}$ 构成的，其中每个 π_t 都将观测数据历史 $\overline{S}_t = S_t \cup \{S_j, A_j, R_j\}_{0\leq j<t}$ 映射为 $\mathcal{A}$上的概率质量函数，记作 $\pi_t(\cdot|\overline{S}_t)$。当每个 π_t 输出 $\mathcal{A}$中的一个值时，π 称为确定性策略。在 π下，智能体将在第 t个决策点以概率 $\pi_t(a|\overline{S}_t)$设置 $A_t = a$。假设存在某个函数 $\widetilde{\pi}$，使得对于任意 t，几乎必然有 $\pi_t(\cdot|\overline{S}_t) = \widetilde{\pi}(\cdot|S_t)$，则 π 称为平稳策略。

强化学习的主要目标是找出能够最大化智能体所获得累积奖励的最优策略。为了正式表述这一目标，我们定义价值函数

$$V^{\pi}(s) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^{\pi}(R_t|S_0 = s), \tag{1}$$

其中，$\mathbb{E}^{\pi}$ 表示假设动作按照 π进行选择时的期望，$0 \leq \gamma < 1$ 表示用于平衡即时奖励和未来奖励之间权衡的某个折扣因子。我们的目标是学习一种能够最大化以下积分价值的策略

函数，

$$\mathcal{V}(\pi) = \sum_{s \in \mathcal{S}} V^{\pi}(s)\nu(s), \tag{2}$$

其中，ν 表示 $\mathcal{S}$上的某个已知参考分布函数。在强化学习文献中，参考分布已知是一个典型假设。我们假设，对于任意 $s \in \mathcal{S}$，$\nu(s)$均有一个严格大于零的统一下界。当 N 较大时，也可以将 ν 设为 S_0 的分布函数，并通过数据样本 $\{S_{i,0}\}_{1 \le i \le N}$的经验分布函数对其进行估计。

以下假设使我们能够专注于平稳策略，并构成现有最先进强化学习算法的基础：

（A1）具有平稳转移的马尔可夫假设：存在一个马尔可夫转移函数 p，使得对于任意 $t \ge 0$、$a \in \mathcal{A}$ 和 s、$s' \in \mathcal{S}$，

$$\Pr(S_{t+1} = s' | A_t = a, S_t = s, \{S_j, A_j, R_j\}_{0 \le j < t}) = p(s'|a, s),$$

（A2）具有平稳奖励函数的条件均值独立性假设：存在某个函数 r，使得对于任意 $t \ge 0$、$s \in \mathcal{S}$ 和 $a \in \mathcal{A}$，

$$\mathbb{E}(R_t | S_t = s, A_t = a, \{S_j, A_j, R_j\}_{0 \le j < t}) = \mathbb{E}(R_t | S_t = s, A_t = a) = r(s, a),$$

这两个假设要求在每个决策时刻 t，给定当前状态-动作对，未来状态和即时奖励的条件均值与过去的观测相互独立。在这些假设下，存在一个最优平稳策略 π^{opt}，其价值函数 $V^{\pi^{opt}}(s)$ 不劣于任意依赖历史的策略 π 的价值函数 $V^{\pi}(s)$（Puterman，1994年，第6.2节）。因此，该策略也能使积分价值函数 $\mathcal{V}(\pi)$ 最大化。(A1) 和 (A2) 可以利用观测数据进行检验。参见 Shi、Wan、Song、Lu 和 Leng（2020年）提出的拟合优度检验。在实践中，为确保满足马尔可夫性质，我们可以将多个决策点的测量值拼接起来构造状态，直至满足马尔可夫性质（详情见第5.2节）。为确保转移概率 p 和奖励函数 r 满足时间齐次性，我们可以在状态中加入一些辅助变量（例如一天中的时间）。下文中，我们将注意力限制在属于某个预先指定类别 Π （例如线性模型或神经网络）的平稳策略上。对于任意 $\pi \in \Pi$，我们记作

$\pi(\bullet|s)$，表示当状态价值为 s 时，智能体将遵循的、定义在 $\mathcal{A}$ 上的概率质量函数。我们的目标是基于观测数据学习 $\pi^* \in \text{argmax}_{\pi\in\Pi}\mathcal{V}(\pi)$。

最后，为了处理离线数据，我们采用文献中常见的一项附加假设（Sutton 和 Barto，2018年）：

（A3）数据由某个固定的平稳策略生成，记为 b。此外，随机过程 $\{(A_t, S_t)\}_{t\geq 0}$ 是平稳的。

在（A1）和（A3）成立的条件下，过程 $\{(A_t, S_t)\}_{t\geq 0}$ 构成一条时间齐次马尔可夫链。我们用 p_∞ 表示状态-动作对的平稳分布。假设对于任意 $a \in \mathcal{A}, s \in \mathcal{S}$，$p_\infty(a, s)$ 均一致地有正下界。为方便起见，我们假设 $\{(A_t, S_t)\}_{t\geq 0}$ 是平稳的，因为该马尔可夫链最终会达到平稳状态。为便于表述，本文始终假设（A1）-（A3）成立。

2.2 附加符号

对于任意 $a \in \mathcal{A}, s \in \mathcal{S}$，我们定义给定策略 π 对应的动作价值函数为 $Q^\pi(a, s) = \sum_{t=0}^{+\infty} \gamma^t \mathbb{E}^\pi(R_t|A_0 = a, S_0 = s)$，通常称为 Q-函数。根据定义，当初始动作-状态对为 (a, s)，且之后所有动作均遵循 π 时，该函数等于智能体获得的折扣累计奖励。由（1）可得，对于任意 π，有 $V^\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)Q^\pi(a, s)$。

与 π 相关的 $\mathcal{A} \times \mathcal{S}$ 上的优势函数 A^π 定义为Q函数与价值函数之差，即对于任意 $a \in \mathcal{A}$ 和 $s \in \mathcal{S}$，$A^\pi(a, s) = Q^\pi(a, s) - V^\pi(s)$。它表示在初始阶段执行动作 a 而非遵循 π 所带来的期望累计奖励增益。根据定义，我们得到

$$\sum_{a\in\mathcal{A}} \pi(a|s)A^\pi(a, s) = 0, \quad \text{for any } s, \pi. \tag{3}$$

下面介绍折扣访问概率。对于任意 $t \geq 0$，令 $p_t^\pi(s'|a, s)$ 表示 t-步访问概率 $\Pr^\pi(S_t = s'|A_0 = a, S_0 = s)$，其中假设在时间 1、$\cdots$、$t-1$ 时按照 π 选择动作。当 $t = 0$ 时，$p_t^\pi(s'|a, s)$ 变为点质量函数 $\mathbb{I}(s' = s)$，其中 $\mathbb{I}(\cdot)$ 表示示性函数。当 $t = 1$ 时，p_t^π 等于条件（A1）中定义的转移函数 p。我们将条件折扣访问概率函数定义为 $d^\pi(s'|a, s) = (1-\gamma)\sum_{t\geq 0} \gamma^t p_t^\pi(s'|a, s)$。令

$$d^{\pi,\nu}(s') = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} \pi(a|s) d^{\pi}(s'|a,s)\nu(s),$$

其为积分折扣访问概率函数。假设按照 π 选择动作，且初始状态服从分布 ν。则 $d^{\pi,\nu}$ 对应于状态 S^* 的概率质量函数，该状态是随机变量$\{S_t\}_{t\geq 0}$的混合，其对应的混合权重为 $\{(1-\gamma)\gamma^t\}_{t\geq 0}$。

最后，我们引入折扣平稳概率比。对于任意 π，定义

$$\omega^{\pi}(a', s'; a, s) = \frac{(1-\gamma)\{\mathbb{I}(s'=s, a'=a) + \sum_{t\geq 1}\gamma^t \pi(a'|s') p_t^{\pi}(s'|a,s)\}}{p_{\infty}(a', s')}. \tag{4}$$

根据定义，式（4）的分母对应于 (A_t, S_t) 的平稳分布。当系统遵循 π 时，分子对应于状态-动作对 (A^*, S^*) 的概率质量函数，该函数是 $\{(A_t, S_t)\}_{t\geq 0}$ 在事件 $(A_0, S_0) = (a, s)$ 条件下的混合。因此，我们将 ω^{π} 称为条件折扣平稳概率比。类似地，定义 $\omega^{\pi,\nu}(a', s') = \sum_{a,s}\pi(a|s)\nu(s)\omega^{\pi}(a', s'; a, s)$ 为积分折扣平稳概率比。我们指出，这些概率比在构建用于离策略评估的去偏价值函数估计量时发挥着重要作用（Kallus 和 Uehara（2019）；Shi 等（2021））。

2.3 信赖域策略优化

以下公式构成了信赖域策略优化的基础（Kakade 和 Langford，2002，引理 6.1），

$$(1-\gamma)\{\mathcal{V}(\pi_{\text{new}}) - \mathcal{V}(\pi_{\text{old}})\} = \sum_{a\in\mathcal{A}, s\in\mathcal{S}} \pi_{\text{new}}(a|s) A^{\pi_{\text{old}}}(a,s) d^{\pi_{\text{new}},\nu}(s), \tag{5}$$

对于任意两个策略 π_{old} 和 π_{new}，其中 A^{π} 和 $d^{\pi,\nu}$ 在第 2.2 节中定义。根据公式（5），对于给定的初始策略 $\pi_{\text{old}} \in \Pi$，直接搜索能够最大化右端（RHS）估计值的 $\pi_{\text{new}} \in \Pi$似乎是可行的。然而，由于 $d^{\pi_{\text{new}},\nu}$ 对 π_{new} 的依赖关系复杂，这种直接搜索方法在计算上具有挑战性。一般而言，很难构造出一个关于 π_{new} 且对 $d^{\pi_{\text{new}},\nu}$具有显式解的估计量。相应估计量的梯度对于公式（5）而言极难计算，因此难以应用梯度类方法。

为使计算可行且高效，Schulman 等人（2015）考虑将式（5）的右端近似为

$$\eta_1(\pi_{\text{new}}, \pi_{\text{old}}) = \sum_{a\in\mathcal{A}, s\in\mathcal{S}} \pi_{\text{new}}(a|s) A^{\pi_{\text{old}}}(a,s) d^{\pi_{\text{old}},\nu}(s). \tag{6}$$

注意，式（6）中的折扣访问概率现在取决于 π_{old}，而不是 π_{new}。量 $(1-\gamma)\mathcal{V}(\pi_{\text{old}}) + \eta_1(\pi_{\text{new}}, \pi_{\text{old}})$ 可视为 $(1-\gamma)\mathcal{V}(\pi_{\text{new}})$ 的一阶近似。进一步地，注意到 $\eta_1(\pi_{\text{new}}, \pi_{\text{old}})$ 可根据式（3）改写为 $\sum_{a\in\mathcal{A},\ s\in\mathcal{S}} \{\pi_{\text{new}}(a|s) - \pi_{\text{old}}(a|s)\} A^{\pi_{\text{old}}}(a,s) d^{\pi_{\text{old}},\ \nu}(s)$。由式（5）可得

$$\begin{aligned}(1-\gamma)\mathcal{V}(\pi_{\text{new}}) &= (1-\gamma)\mathcal{V}(\pi_{\text{old}}) + \underbrace{\sum_{a\in\mathcal{A}, s\in\mathcal{S}} \{\pi_{\text{new}}(a|s) - \pi_{\text{old}}(a|s)\} A^{\pi_{\text{old}}}(a,s) d^{\pi_{\text{old}},\nu}(s)}_{\eta_1(\pi_{\text{new}}, \pi_{\text{old}})} \\ &+ \underbrace{\sum_{a\in\mathcal{A}, s\in\mathcal{S}} \{\pi_{\text{new}}(a|s) - \pi_{\text{old}}(a|s)\} A^{\pi_{\text{old}}}(a,s) \{d^{\pi_{\text{new}},\nu}(s) - d^{\pi_{\text{old}},\nu}(s)\}}_{\eta_2(\pi_{\text{new}}, \pi_{\text{old}})},\end{aligned}$$

其中，$\eta_2(\pi_{\text{new}}, \pi_{\text{old}})$对应一个高阶余项。为量化该高阶余项，对于 $\mathcal{A}$上的任意两个概率分布 μ_1和 μ_2，我们用 $\mathcal{D}_{\text{TV}}(\mu_1, \mu_2)$表示全变差距离$2^{-1}\sum_{a\in\mathcal{A}} |\mu_1(a) - \mu_2(a)|$。令

$$\mathcal{D}_{\text{KL}}(\mu_1, \mu_2) = \sum_{a\in\mathcal{A}} \mu_1(a) \log\{\mu_1(a)/\mu_2(a)\}$$

表示从 μ_1 到 μ_2的 Kullback–Leibler（KL）散度。在引理 1（见附录A.1）中，我们证明

$$\begin{aligned}|\eta_2(\pi_{\text{new}}, \pi_{\text{old}})| &\le c^* \left[\mathbb{E}_{S^*\sim d^{\pi_{\text{old}},\nu}} \mathcal{D}_{\text{TV}}\left(\pi_{\text{old}}(\bullet|S^*), \pi_{\text{new}}(\bullet|S^*)\right)\right]^2 \\ &\le c^* \mathbb{E}_{S^*\sim d^{\pi_{\text{old}},\nu}} \mathcal{D}_{\text{KL}}\left(\pi_{\text{old}}(\bullet|S^*), \pi_{\text{new}}(\bullet|S^*)\right),\end{aligned} \tag{7}$$

其中某个正常数为 $c^* > 0$。(7)中的第一个不等式表明，$|\eta_2(\pi_{\text{new}}, \pi_{\text{old}})|$ 确实是一个二阶项。基于这一观察，我们考虑通过最大化(5)的一个下界来进行策略优化，该下界为

$$\pi_{\text{new}} \in \text{argmax}_{\pi\in\Pi} \left[\eta_1(\pi, \pi_{\text{old}}) - c^* \mathbb{E}_{S^*\sim d^{\pi_{\text{old}},\nu}} \mathcal{D}_{\text{KL}}\left(\pi_{\text{old}}(\bullet|S^*), \pi(\bullet|S^*)\right)\right]. \tag{8}$$

迭代求解上述优化问题会得到一种称为 minorization-maximization（MM）的算法（Hunter和Lange，2004），正如我们所见，当 $\pi = \pi_{\text{old}}$ 时，式（8）中的目标函数变为 0，而 $(1-\gamma)\mathcal{V}(\pi)$ 则变为 $(1-\gamma)\mathcal{V}(\pi_{\text{old}})$。因此，可以保证优化后式（5）始终非负。由此，这类算法无需计算 $d^{\pi_{\text{new}},\ \nu}$，便能大幅降低计算成本，同时单调提升积分价值函数。不过在实践中，要稳健地

选择式（8）中的惩罚系数 c^* 可能较为困难。为了解决这一问题，可以考虑迭代求解以下带有所谓信赖域约束的等价优化问题：

$$\begin{aligned} &\pi_{\text{new}} \in \operatorname{argmax}_{\pi \in \Pi} \eta_1(\pi, \pi_{\text{old}}) \\ &\text{subject to } \mathbb{E}_{S^* \sim d^{\pi_{\text{old}}, \nu}} \mathcal{D}_{\text{KL}}\left(\pi_{\text{old}}(\bullet|S^*), \pi(\bullet|S^*)\right) \leq \delta, \end{aligned} \tag{9}$$

对于某个常数 $\delta > 0$。由此得到信赖域策略优化算法。

3 价值增强型策略优化

本节首先介绍我们方法的动机。要实现信赖域策略优化，我们需要估计 $\eta_1(\pi, \pi_{\text{old}})$。在在线场景中，Schulman 等人（2015）提出按照策略 π_{old} 生成轨迹，以估计 $\eta_1(\pi, \pi_{\text{old}})$。在离线场景中，如何根据观测数据有效评估这一量仍未可知。根据其定义，我们注意到 η_1 取决于干扰函数 $A^{\pi_{\text{old}}}$ 和 $d^{\pi_{\text{old}}, \nu}$。一种朴素方法是先估计这些量（分别记为 $\widetilde{A}$ 和 $\widetilde{d}^{\nu}$），然后使用相应的插件估计量 $\widetilde{\eta}_1 = \mathbb{E}_{S^* \sim \widetilde{d}^{\nu}} \sum_{a \in \mathcal{A}} \pi(a|S^*) \widetilde{A}(a, S^*)$ 来估计 $\eta_1(\pi, \pi_{\text{old}})$。然而，这一过程存在以下三个主要缺点：

(I) 迭代求解优化问题 (9) 仍可能需要大量计算，因为依赖于策略的干扰函数需要在每次迭代时更新，尤其是在我们没有用于估计这些干扰函数的闭式表达式（例如 d^{ν}）时。因此，直接实现信赖域策略优化方法可能并不可取。

(II) 当 $\widetilde{A}$ 或 $\widetilde{d}^{\nu}$ 不一致时，$\widetilde{\eta}_1$ 可能也不一致。因此，无法保证由此得到的新策略 π_{new} 能够优于 π_{old}。

(III) 为确保 $\widetilde{A}$ 和 $\widetilde{d}^{\nu}$ 都具有一致性，可以考虑对这些函数进行非参数估计。即使两者都具有一致性，代入估计量 $\widetilde{\eta}_1$ 也可能不是速率最优的，即 $(NT)^{-1/2}$，因为非参数估计量 $\widetilde{A}$ 和 $\widetilde{d}^{\nu}$ 的收敛速度通常远慢于 $(NT)^{-1/2}$。因此，与 π_{old} 相比，π_{new} 带来的改进可能微乎其微，从而导致收敛至最优策略的速度较慢。

为解决第一个问题，我们提出先采用现有的一些最先进的离线强化学习方法，以获得良好的初始策略。这里可以采用多种方法，包括保守Q学习（CQL，Kumar 等人，2020年）、拟合Q迭代、V学习等。CQL 和神经网络拟合Q迭代使用神经网络来表示策略类，而 V学习则考虑由有限维向量索引的参数化策略类。值得注意的是，这些方法基本上依赖于对价值函数或 Q函数的估计。它们的计算效率高于（I）中描述的迭代过程。这些算法得到的初始策略下的价值函数若具有一致性，则可能以较慢的速度收敛，这是换取快速计算所付出的代价。在第二步中，我们提出求解（9）以提升其性能。这相当于对初始策略进行一次单步更新。也可以对这一新策略再更新几次，以确保最终估计的策略能够达到较快的收敛速度。为消除初始策略与策略优化之间的依赖关系，我们采用数据分割策略；该策略在统计学和机器学习中十分常见，例如 Chernozhukov等人（2018年）；Kallus 和 Uehara（2019年）。详细过程见第3.2节。

为解决第二和第三个问题，我们开发了一种针对 η_1的高效稳健估计程序，详见第3.1节。具体而言，当输入策略π_{old} 一致时，我们可以保证，通过求解（9）得到的输出策略 π_{new} 具有所期望的“价值增强”性质。我们将这类方法称为“价值增强策略优化（VEPO）”。第3.2节概述了我们的算法，该算法整合了Q学习、折扣平稳概率比估计、转移动态估计和策略搜索。接下来，我们将在本节余下部分讨论各个组成部分。

3.1 针对 η_1的高效多重稳健估计量

对于给定的 $\pi_{\text{old}} \in \Pi$，我们首先概述三种从观测数据中估计 $\eta_1(\pi, \pi_{\text{old}})$的潜在方法（见下文（i）—（iii））。这些方法都需要对某些干扰函数进行一致估计。随后，我们提出一种结合这三种方法的方案，以实现高效且三重稳健的估计。

（i）代入估计量：$\widetilde{\eta}_1^{(1)} = \mathbb{E}_{S^* \sim \widetilde{d}^{\nu}} \sum_{a \in \mathcal{A}} \pi(a|S^*)\, \widetilde{A}(a, S^*)$。这就是前面讨论的插件方法。$\widetilde{\eta}_1^{(1)}$ 的有效性要求对 $d^{\pi_{\text{old}},\ \nu}$ 和 $A^{\pi_{\text{old}}}$进行一致估计。

（ii）重要性采样（IS）估计量 I:

$$\widetilde{\eta}_1^{(2)} = \frac{1}{\sum_i T_i} \sum_{i=1}^{N} \sum_{t=0}^{T_i-1} \mathbb{E}_{S^* \sim \widetilde{d}^{\nu}} \sum_{a \in \mathcal{A}} \{\pi(a|S^*) - \pi_{\text{old}}(a|S^*)\} \widetilde{\omega}(A_{i,t}, S_{i,t}; a, S^*) R_{i,t}, \tag{10}$$

其中，$\widetilde{\omega}$ 表示条件折扣概率比 $\omega^{\pi_{\text{old}}}$的某个估计量。详细定义见式（4）。$\widetilde{\eta}_1^{(2)}$ 的有效性要求对$d^{\pi_{\text{old}},\nu}$ 和 $\omega^{\pi_{\text{old}}}$进行一致估计。这种 IS 估计量的构思源自 Liu 等人（2018）关于离策略价值评估的研究。一个关键观察是，在（A2）和（A3）的条件下，Q^{π}（a，s）可表示为

$$\sum_{a',s'} \{\mathbb{I}(s'=s, a'=a) + \sum_{t \geq 1} \gamma^t \pi(a'|s') p_t^{\pi}(s'|a,s)\} r(s',a') = \frac{1}{1-\gamma} \mathbb{E}\omega^{\pi}(A_t, S_t; a, s) R_t,$$

对于任意 t、s 和 a。由此得到 $A^{\pi}(a, s)$ 的以下重要性抽样估计量：

$$\frac{1}{\sum_i T_i} \sum_{i=1}^{N} \sum_{t=0}^{T_i-1} \sum_{a' \in \mathcal{A}} \{\mathbb{I}(a'=a) - \pi(a'|s)\} \widetilde{\omega}(A_{i,t}, S_{i,t}; a', s) R_{i,t}.$$

~将上述 $A^{\pi_{\text{old}}}(a, s)$ 的估计量和 $d^{\pi_{\text{old}}}$ 的估计量代入 d^{ν}，ν即可得到 (10)。

(iii) 重要性抽样估计量 II：

$$\widetilde{\eta}_1^{(3)} = \frac{1}{\sum_i T_i} \sum_{i=1}^{N} \sum_{t=0}^{T_i-1} \sum_{a \in \mathcal{A}} \pi(a|S_{i,t}) \widetilde{A}(a, S_{i,t}) \widetilde{\omega}^{\nu}(A_{i,t}, S_{i,t}), \tag{11}$$

其中，$\widetilde{\omega}^{\nu}$ 表示积分概率比 $\omega^{\pi_{\text{old}}, \nu}$ 的某个估计量。$\widetilde{\eta}_1^{(3)}$ 的有效性要求对 $d^{\pi_{\text{old}}, \nu}$ 和积分概率比 $\omega^{\pi_{\text{old}}, \nu}$进行一致估计。为说明该估计量的合理性，我们注意到，期望 $\mathbb{E}_{S \sim d^{\pi, \nu}} f(S)$ 可改写为 $\mathbb{E} f(S_t)\omega^{\pi, \nu}(A_t, S_t)$，其中函数 f、策略 π和决策点 t 均可任取。因此，我们可以将 η_1 表示为

$$\mathbb{E} \sum_{a \in \mathcal{A}} \pi(a|S_t) A^{\pi_{\text{old}}}(a, S_t) \omega^{\pi_{\text{old}}, \nu}(A_t, S_t).$$

由此得到式 (11) 中的重要性抽样估计量。

我们注意到，当相应的估计干扰函数不一致时，上述每个估计量都可能存在严重偏差。为此，我们通过谨慎结合 (i)–(iii) 中采用的估计策略，构造了一个多重稳健估计量。与此同时，该估计量只需满足更弱的假设即可实现一致性。令 o 作为数据元组 (s, a, r, s') 的简写。构造该估计量的关键在于

以下估计函数，

$$\psi(o;\pi,\pi_{\text{old}},\widetilde{V},\widetilde{A},\widetilde{\omega},\widetilde{d})=\psi_1(\pi,\pi_{\text{old}},\widetilde{A},\widetilde{d})+\psi_2(o;\pi,\pi_{\text{old}},\widetilde{V},\widetilde{A},\widetilde{\omega},\widetilde{d})+\psi_3(o;\pi,\pi_{\text{old}},\widetilde{A},\widetilde{\omega},\widetilde{d}),$$

其中，给定干扰函数 $\widetilde{V}$、$\widetilde{A}$、$\widetilde{\omega}$ 和 $\widetilde{d}$，

$$\psi_1(\pi,\pi_{\text{old}},\widetilde{A},\widetilde{d})=\mathbb{E}_{S^*\sim\widetilde{d}^\nu}\sum_{a\in\mathcal{A}}\pi(a|S^*)\widetilde{A}(a,S^*),$$

$$\psi_2(o;\pi,\pi_{\text{old}},\widetilde{V},\widetilde{A},\widetilde{\omega},\widetilde{d})=\frac{1}{1-\gamma}\mathbb{E}_{S^*\sim\widetilde{d}^\nu}\sum_{a^*\in\mathcal{A}}\{\pi(a^*|S^*)-\pi_{\text{old}}(a^*|S^*)\}\widetilde{\omega}(a,s;a^*,S^*)$$
$$\times\{r+\gamma\widetilde{V}(s')-\widetilde{V}(s)-\widetilde{A}(a,s)\}$$

$$\psi_3(o;\pi,\pi_{\text{old}},\widetilde{A},\widetilde{\omega},\widetilde{d})=\sum_{a^*\in\mathcal{A}}\frac{\widetilde{\omega}^\nu(a,s)}{1-\gamma}\Big[\gamma\mathbb{E}_{\substack{a'\sim\pi_{\text{old}}(\bullet|s')\\S^*\sim\widetilde{d}(\bullet|a',s')}}\widetilde{A}(a^*,S^*)\pi(a^*|S^*)$$
$$-\mathbb{E}_{S^*\sim\widetilde{d}(\bullet|a,s)}\widetilde{A}(a^*,S^*)\pi(a^*|S^*)+(1-\gamma)\pi(a^*|s)\widetilde{A}(a^*,s)\Big],$$

其中，干扰函数 d^ν 和 $\widetilde{\omega}^\nu$ 由 d和 $\widetilde{\omega}$确定，具体如下：$d^\nu(\bullet)=\sum_{a,s}\pi_{\text{old}}(a|s)\nu(s)d(\bullet|a,s)$ 和 $\widetilde{\omega}^\nu(\bullet,\bullet)=\sum_{a,s}\pi_{\text{old}}(a|s)\nu(s)\widetilde{\omega}(\bullet,\bullet;a,s)$。

根据定义，ψ 由三项组成。第一项 ψ_1 本质上是仅依赖于 A 和 d的代入估计量 $\widetilde{}$。第二项和第三项，即 ψ_2 和 ψ_3，是增广项。对任意 t，令 $O_t=(S_t,A_t,R_t,S_{t+1})$，则当 $A=A^{\pi_{\text{old}}}$时，我们有 $\mathbb{E}\psi_2(O_t;\pi,\pi_{\text{old}},V,A,\widetilde{\omega},d)=0$；当 $d=d^{\pi_{\text{old}}}$时，$\mathbb{E}\psi_3(O_t;\pi,\pi_{\text{old}},A,\widetilde{\omega},d)=0$。详情见附录 A.2。加入这两项的目的是进一步防范因 A 和 d的偏差而导致 ψ_1 可能产生的偏差。因此，我们有以下命题。

命题 1 假设 $\sum_a\pi_{\text{old}}(a|s)A(a,s)=0$ 对任意 s均成立。那么，只要满足以下三个假设之一，$\psi(O_t;\pi,\pi_{\text{old}},V,A,\widetilde{\omega},d)$ 对 η_1 而言就是无偏的：(B1) $A=A^{\pi_{\text{old}}}$，$V=V^{\pi_{\text{old}}}$和 $d=d^{\pi_{\text{old}}}$；(B2) $\widetilde{\omega}=\omega^{\pi_{\text{old}}}$和 $d=d^{\pi_{\text{old}}}$；(B3) $A=A^{\pi_{\text{old}}}$和$\widetilde{\omega}=\omega^{\pi_{\text{old}}}$。

如果我们设定 $A(a,s)=A^*(a,s)-\sum_a\pi_{\text{old}}(a|s)A^*(a,s)=0$，那么对于任意初始优势估计量 A^*，条件 $\sum_a\pi_{\text{old}}(a|s)A(a,s)Q^{\pi_{\text{old}}}$ 都会自动满足。我们指出，如果 $Q^{\pi_{\text{old}}}$ 被正确设定，那么 $A^{\pi_{\text{old}}}$ 和 $V^{\pi_{\text{old}}}$ 也会被正确设定。基于该估计函数，η_1 的三重稳健估计量为

$$\frac{1}{\sum_i T_i}\sum_{i=1}^{N}\sum_{t=0}^{T_i-1}\psi(O_{i,t};\pi,\pi_{\text{old}},\widetilde{V},\widetilde{A},\widetilde{\omega},\widetilde{d}),\tag{12}$$

其中 $O_{i,\ t} = (S_{i,\ t}, A_{i,\ t}, R_{i,\ t}, S_{i,\ t+1})$。接下来需要说明如何估计干扰函数。我们将在下一节介绍具体细节。在第4.2节中，我们将证明所得估计量是有效的。

3.2 完整算法

我们的主要思路是构造一个高效且稳健的估计量 η_1，以提升初始策略 π_{old}的性能。为实现这一目标，我们需要估计四个关键干扰函数：(a) 初始策略 π_{old}；(b) 价值函数和优势函数$V^{\pi_{\text{old}}}$ 与 $A^{\pi_{\text{old}}}$；(c) 条件折扣平稳概率比 $\omega^{\pi_{\text{old}}}$；(d) 条件折扣访问概率函数 $d^{\pi_{\text{old}}}$。

相应地，我们的估计程序包含四个关键步骤，分别在第3.2.1至3.2.4节中介绍。除这四个主要估计环节外，我们还提出将式（12）中的估计量与数据拆分和交叉拟合策略相结合。具体而言，不失一般性，我们将所有轨迹的索引随机划分为$\{1, 2, \cdots, N\}$ 个 $\mathbb{L}$ 子集$\cup_{\ell=1}^{\mathbb{L}} \{O_{i,t}\}_{i\in\mathcal{I}_\ell,\ 0\le t<T_i}$每个子集大小相同，其中 $\mathcal{I}_\ell$ 表示第 ℓ个数据子集中所含轨迹的索引。接下来，我们对 $\mathcal{I}_\ell^c = \{1, \cdots, N\}\backslash\mathcal{I}_\ell$ 中的数据子集应用(a)至(d)中的学习环节，$\ell = 1,\ \cdots,\ \mathbb{L}$ 并通过交叉拟合构造估计量 $\widehat{\eta}_1$。交叉拟合实质上保证了用于学习(a)至(d)的数据集与用于构造 $\widehat{\eta}_1$的数据集相互独立。这使我们无需施加Donsker型条件；此类条件会限制(a)至(d)中估计器的VC维数增长速率（Chernozhukov等人，2018年），从而使我们的程序具备理想性质。然后，我们提出搜索 π_{new}，使其最大化 $\widehat{\eta}_1$，并满足式（9）中的信赖域约束，其中 $d^{\pi_{\text{old}}}, \nu$ 由相应的估计量替代。这相当于对初始策略进行一次单步更新。计算出 π_{new}后，我们可以重复上述步骤若干次，以确保最终估计策略达到快速收敛速率。详情见第3.2.5节。算法1给出了本方法的伪代码。

我们指出，无需针对信赖域约束中的积分KL散度开发稳健且高效的估计程序，因为对于价值差而言，它对应的是一个高阶余项（见附录A.1中的引理 1）。只要 $d^{\pi_{\text{old}}}, \nu$ 及其估计量具有共同支撑，我们提出的方法就适用。当参考分布 ν 为

算法1 价值增强策略优化 输入：策略类 Π 和观测数据。输出：更新后的策略 π_{new}。

步骤 0. 将轨迹随机划分为 $\mathbb{L}$ 个不相交子集，$\cup_{\ell=1}^{\mathbb{L}}\mathcal{I}_\ell$。令 $\mathcal{I}_\ell^c = \{1,\cdots,N\} - \mathcal{I}_\ell$，其中 $\ell = 1、\cdots、\mathbb{L}$。

步骤 1. 对于 $\ell = 1、\cdots、\mathbb{L}$：应用某种现有的先进离线强化学习算法，利用数据子集 $\mathcal{I}_\ell^c$ 得到输入初始策略 π_{old}。将得到的策略记为 $\pi^{(\ell)}{}_{\text{旧}}$。

步骤 2. 对于 $\ell = 1、\cdots、\mathbb{L}$:

(2a) 基于 $\mathcal{I}_\ell^c$ 中的数据子集，应用拟合Q评估（见 (14)）来估计 Q函数 $Q^{\pi^{(\ell)}}{}_{\text{旧}}$。将得到的估计量记为 $\widehat{Q}^{(\ell)}$。(2b) 对任意 s，设 $\widehat{V}_{(\ell)}(s) = \sum_a \pi^{(\ell)}{}_{\text{旧}}(a|s)\,\widehat{Q}^{(\ell)}(a,s)$，且对任意 a 和 s，$\widehat{A}^{(\ell)}(a,s) = \widehat{Q}^{(\ell)} - \widehat{V}_{(\ell)}(s)$。

步骤 3. 对于 $\ell = 1, 2, \cdots, \mathbb{L}$，根据数据子集 $\mathcal{I}_\ell^c$，应用第 3.2.3 节中详述的方法，学习条件概率比 $\widehat{\omega}^{(\ell)}$。

步骤 4. 对于 $\ell = 1, 2, \cdots, \mathbb{L}$，使用数据子集 $\mathcal{I}_\ell^c$，应用机器学习方法近似 S_{t+1} 给定 A_t 和 S_t 时的条件分布，以生成伪样本。随后可利用该伪样本近似分布函数 $d^{\pi^{(\ell)}}{}_{\text{旧}}$ 和 $d^{\pi^{(\ell)}}{}_{\text{旧}}{}^{\nu}$（参见算法 2 和 3）。

步骤 5. 构造 η_1 的估计量并更新相应的策略:

(5a) 应用交叉拟合构造价值差估计量 $\widehat{\eta}_1$（参见（19））。(5b) 使用估计得到的动态量构造信赖域约束（参见（20））。(5c) 搜索使 $\widehat{\eta}_1$ 最大化且满足（20）的 $\pi_{\text{new}} \in \Pi$。

步骤 6. 将所有 $\pi^{(\ell)}{}_{\text{旧}}$ 设置为 π_{new}，针对 $\ell = 1、\cdots、\mathbb{L}$，并重复步骤 2–5 几次。

在 $\mathcal{S}$上均一致地有正下界。

3.2.1 步骤 1：初始策略优化

首先，为了初始化我们的 VEPO 算法，我们需要估计一个记作 π_{old} 的初始策略。我们提出在数据子集 $\mathcal{I}_\ell^c$ 上应用一些现有的最先进的离线强化学习算法，并得到相应的估计策略，记作 $\hat{}\pi^{(\ell)}{}_{\text{旧策略}}$，用于 $\ell = 1、\cdots、\mathbb{L}$。为便于表述，在不引起混淆时，我们常将 $\hat{}\pi^{(\ell)}{}_{\text{旧策略}}$写作 $\pi^{(\ell)}{}_{\text{旧策略}}$。具体而言，在数值研究中，我们实现了三种离线强化学习算法来获得初始策略。第一种是拟合Q迭代（FQI），采用价值迭代结合函数近似的思想（Sutton and Barto，2018）。它依赖于最优贝尔曼方程（Bertsekas and Tsitsiklis, 1996）。第二种是 Luckett 等人提出的 V学习。

（2020），该方法考虑了结合函数近似的策略迭代。最后一种是 Kumar 等人（2020）提出的保守Q学习（CQL），其在策略迭代过程中学习 Q 函数的下界。该方法旨在解决由于批次数据集中的行为策略与学习得到的策略之间存在分布不匹配而导致的价值函数高估问题。我们指出，只要满足定理 2 中的假设，任何有效的离线强化学习方法都可以用于此处。详情请参见第 4.1 节。

3.2.2 步骤2：Q学习

其次，为估计 (b) 中的干扰函数，我们采用 Q学习型算法，基于 $\mathcal{I}_\ell^c$ 中的数据子集来学习 Q函数 $Q^{\pi^{(\ell)}}$old。将相应的估计量记为 $\widehat{Q}^{(\ell)}$。随后，根据以下关系 $V^{\pi^{(\ell)}}$old(s) $=\sum_a \pi^{(\ell)}$old$(a|s)Q^{\pi^{(\ell)}}$old (a,s)，$A^{\pi^{(\ell)}}$old(a,s) $=Q^{\pi^{(\ell)}}$old (a,s) $-V^{\pi^{(\ell)}}$old(s)，我们构造相应的价值函数和优势函数估计量：$\widehat{V}(\ell)(s)$ $=\sum_a \pi^{(\ell)}$old$(a|s)$ $\widehat{Q}^{(\ell)}(a,s)$ 和 $\widehat{A}^{(\ell)}(a,s)$ $=\widehat{Q}^{(\ell)}(a,s)$ $-\widehat{V}(\ell)(s)$，其中 s 和 a 任意。因此，优势函数估计量自动满足条件 $\sum_a \pi^{(\ell)}$ old$(a|s)$ $\widehat{A}^{(\ell)}(a,s)$ $=0$。

此处可使用多种算法来估计 Q 函数 $Q^{\pi^{(\ell)}}$旧。我们采用 Le 等人（2019）提出的拟合 Q 评估（ FQE ）方法。以下贝尔曼方程是所有 Q 学习型算法的基础：对于 $t \geq 0$,

$$Q^{\pi_{\text{old}}^{(\ell)}}(A_t, S_t) = \mathbb{E}\left\{ R_t + \gamma \sum_a \pi_{\text{old}}^{(\ell)}(a|S_{t+1}) Q^{\pi_{\text{old}}^{(\ell)}}(a, S_{t+1}) \middle| A_t, S_t \right\}. \tag{13}$$

基于这一恒等式，我们通过迭代计算来估计 $Q^{\pi^{(\ell)}}$旧

$$\widehat{Q}_k^{(\ell)} = \text{argmin}_{Q_k} \sum_{i,t} \left\{ R_{i,t} + \sum_a \pi_{\text{old}}^{(\ell)}(a|S_{i,t+1}) \widehat{Q}_{k-1}^{(\ell)}(a, S_{i,t+1}) - Q_k(A_{i,t}, S_{i,t}) \right\}^2, \tag{14}$$

对于 $k = 1, 2, \cdots$，任意初始 $\widehat{Q}_0^{(\ell)}$ 均可。由于（14）本质上是一个回归问题，因此可以采用多种监督学习方法。在我们的实现中，我们使用深度学习（LeCun 等人，2015年）在每次迭代中计算 $\widehat{Q}_k^{(\ell)}$。

3.2.3 步骤 3：折扣平稳概率比估计

我们采用 Shi 等人（2021年）提出的算法来估计 (c)。该过程基于以下观察：对于任意两对 (i,t) 和 (i',t')，满足 $O_{i,t}$

且 $O_{i',t'}$ 相互独立，对于任意满足 $\mathbb{E}\Delta(\omega^\pi, f, \pi;\ i, t, i', t') = 0$ 的函数 f，都有

$$\Delta(\omega^\pi, f, \pi; i, t, i', t') = \omega^\pi(S_{i',t'}, A_{i',t'}; S_{i,t}, A_{i,t})\Big\{\gamma \sum_a \pi(a|S_{i',t'+1}) f(S_{i',t'+1}, a; S_{i,t}, A_{i,t}) \\ - f(S_{i',t'}, A_{i',t'}; S_{i,t}, A_{i,t})\Big\} + (1-\gamma) f(S_{i,t}, A_{i,t}, S_{i,t}, A_{i,t}). \tag{15}$$

反之，对于任意满足 $\mathbb{E}\Delta(\omega、\ f、\ \pi;\ i、\ t、\ i'、\ t') = 0$ 的 ω，对于任意 f，我们都有 $\omega = \omega^\pi$。详情请参见 Shi 等人（2021）的公式（5）。

对于每个函数 f，都可以基于（15）为 $\omega^{\pi^{(\ell)}_{\text{旧}}}$ 构造一个估计方程。与 Liu 等人（2018）的提议类似，f 可以作为判别器，用于构造如下极小极大损失函数

$$\text{argmin}_{\omega\in\Omega} \sup_{f\in\mathcal{F}} \left|\mathbb{E}\Delta(\omega^{\pi^{(\ell)}_{\text{old}}}, f, \pi^{(\ell)}_{\text{old}}; i, t, i', t')\right|^2, \tag{16}$$

对于某些函数类 Ω 和 $\mathcal{F}$。为简化计算，将 $\mathcal{F}$ 设为再生核希尔伯特空间的单位球。这样便可得到（16）中内层最大化问题的闭式表达式。随后，利用 $\mathcal{I}^c_\ell$ 中数据子集的经验分布近似（16）中的期望。ω 中涉及的参数通过随机梯度下降算法进行更新。为节省篇幅，具体细节请参见 B.1 节。

3.2.4 步骤 4：底层动力学的估计

当状态空间维度很高时，(d) 中的条件折扣访问概率 $d^{\pi^{(\ell)}_{\text{old}}}$ 极难估计，因为它对应于不同决策点处状态变量的混合分布。一个关键的观察是，$d^{\pi^{(\ell)}_{\text{old}}}$ 完全由系统动力学决定。只要能够一致地估计转移核 p，就可以很好地近似 $d^{\pi^{(\ell)}_{\text{old}}}$。

为了计算 $d^{\pi^{(\ell)}_{\text{旧}}}$，在连续状态空间中，我们提出使用高斯概率模型来估计转移核 p，并使用机器学习模型近似均值和协方差矩阵。具体而言，我们假设给定当前动作-状态 S_{t+1}（S_t，A_t）的下一个状态服从多元高斯分布$\mathcal{N}$（μ（S_t，A_t）Σ（S_t，A_t）），其中 μ和 Σ分别是对应的均值函数和协方差矩阵函数。请注意，估计 μ 本质上是一个回归问题，并且有许多监督学习方法可用。在我们的实现中，我们使用深度

算法2 生成伪样本以近似 $d^{\pi^{(\ell)}} \text{旧}(\bullet|s,a)$

输入：通过监督学习计算得到的估计器 $(\widehat{\mu}^{(\ell)}, \widehat{\Sigma}^{(\ell)})$ 。

对于 $m=1$ 到 M: 执行

(a) Set $\widetilde{S}_0^{(m)} = s$ and $\widetilde{A}_0^{(m)} = a$;

(b) For $t' = 1$ to T', generate $\widetilde{S}_{t'}^{(m)}$ by $\mathcal{N}(\widehat{\mu}^{(\ell)}(\widetilde{A}_{t'-1}^{(m)}, \widetilde{S}_{t'-1}^{(m)}), \widehat{\Sigma}^{(\ell)}(\widetilde{A}_{t'-1}^{(m)}, \widetilde{S}_{t'-1}^{(m)}))$ and randomly sample $\widetilde{A}_{t'}^{(m)}$ from $\pi_{\text{old}}^{(\ell)}(\bullet|\widetilde{S}_{t'}^{(m)})$.

输出 $\{\widetilde{S}_{t'}^{(m)}\}_{1\le m\le M,\ 0\le t\le T'}$。

算法3 生成伪样本以近似 $d^{\pi^{(\ell)}} \text{旧} \nu$

输入：通过监督学习计算得到的估计器 $(\widehat{\mu}^{(\ell)}, \widehat{\Sigma}^{(\ell)})$ 。

对于 $m=1$ 到 M: 执行

(a) Sample $\widetilde{S}_0^{(m),\nu}$ from ν;

(b) For $t' = 0$ to $T'-1$, randomly sample $\widetilde{A}_{t'}^{(m),\nu}$ from $\pi_{\text{old}}^{(\ell)}(\bullet|\widetilde{S}_{t'}^{(m),\nu})$ and generate generate $\widetilde{S}_{t'+1}^{(m),\nu}$ by $\mathcal{N}(\widehat{\mu}^{(\ell)}(\widetilde{A}_{t'}^{(m)}, \widetilde{S}_{t'}^{(m),\nu}), \widehat{\Sigma}^{(\ell)}(\widetilde{A}_{t'}^{(m)}, \widetilde{S}_{t'}^{(m),\nu}))$.

输出 $\{\widetilde{S}_{t'}^{(m),\nu}\}_{1\le m\le M,\ 0\le t\le T'}$。

神经网络来计算 $\widehat{\mu}^{(\ell)}$，通过

$$\widehat{\mu}_j^{(\ell)} = \arg\min_{\mu_j} \sum_{i\in\mathcal{I}_\ell^c} \sum_{t=0}^{T_i-1} [S_{i,t+1,j} - \mu_j(S_{i,t}, A_{i,t})]^2,$$

其中，$\widehat{\mu}_j^{(\ell)}$ 和 $S_{i,\ t+1,\ j}$ 分别表示 $\widehat{\mu}^{(\ell)}$ 和 $S_{i,\ t+1}$ 的第 j 个元素。接下来，令 $\varepsilon_{i,\ t,\ j}$表示残差 $S_{i,\ t+1,\ j} - \widehat{\mu}_j^{(\ell)}$ $(S_{i,\ t},\ A_{i,\ t})$ 。我们再次采用深度学习来计算 $\widehat{\Sigma}^{(\ell)}$，通过

$$\widehat{\Sigma}_{j_1,j_2}^{(\ell)} = \arg\min_{\Sigma_{j_1,j_2}} \sum_{i\in\mathcal{I}_\ell^c} \sum_{t=0}^{T_i-1} [\varepsilon_{i,t,j_1}\varepsilon_{i,t,j_2} - \Sigma_{j_1,j_2}(S_{i,t}, A_{i,t})]^2,$$

其中 $\widehat{\Sigma}_{j_1 j_2}^{(\ell)}$ 表示 $\widehat{\Sigma}^{(\ell)}$ 的第 $(j_1,\ j_2)$ 个条目，它是 $\Sigma^{(\ell)}$ 的最终估计量。

为了近似任意 a 和 s 下的条件分布 $d^{\pi^{(\ell)}} \text{旧}(\bullet|a,s)$，我们可以采用蒙特卡罗方法，并基于 $\widehat{\mu}^{(\ell)}$ 和 $\widehat{\Sigma}^{(\ell)}$生成一系列伪样本 $\{\widetilde{S}_{t'}^{(m)}\}_{1\le m\le M,\ 0\le t\le T'}$。具体细节见算法2。对于任意函数 f，积分 $\mathbb{E}_{S^*\sim d^{\pi^{(\ell)}} \text{旧}(\bullet|a,\ s)} f(S^*)$ 可以近似为

$$(1-\gamma)M^{-1} \sum_{m=1}^{M} \sum_{t'=0}^{T'} \gamma^t f(\widetilde{S}_{t'}^{(m)}), \tag{17}$$

我们将其记为 $\mathbb{E}_{S^* \sim \widehat{d}^{(\ell)}(\bullet|a,\ s)} f(S^*)$，其中 $\widehat{d}^{(\ell)}$ 表示估计的概率密度/质量函数。我们使用聚合平方总变差距离

$$\mathbb{E}_{(S,A)\sim p_\infty} \mathcal{D}_{\mathrm{TV}}^2(\widehat{d}^{(\ell)}(\bullet|A,S), d^{\pi_{\mathrm{old}}^{(\ell)}}(\bullet|A,S)) \tag{18}$$

用于衡量其拟合优度。详情请参见第4节中的条件（C3）。在补充材料的 B.2 节中，我们证明，当条件高斯模型设定正确时，对于任意 a、s，$\Sigma(a,s)$ 的最小特征值都一致地远离 ^零；对于任意 a、s，$\Sigma(a,s)$ 都是正定的，因此 (18) 的上界为

$$3\gamma^{2T'} + \frac{3}{M} + O(1)\mathbb{E}_{(A^*,S^*)\sim p_\infty}[\|\mu(S^*,A^*) - \widehat{\mu}^{(\ell)}(S^*,A^*)\|_2 + \|\Sigma(S^*,A^*) - \widehat{\Sigma}^{(\ell)}(S^*,A^*)\|_F]^2,$$

其中 $O(1)$ 表示某个正常数。因此，只要 T'、$M \to \infty$，且估计得到的条件均值和协方差函数一致，$\widehat{d}^{(\ell)}$ 就是一致的。我们还要指出，条件高斯模型在强化学习文献中被广泛用于学习转移函数（例如，参见 Yu 等，2020）。或者，也可以采用条件高斯混合模型来缓解模型设定错误（Bishop，1994）。

3.2.5 步骤 5：策略优化

获得四个关键组成部分后，我们接下来讨论计算新策略 π_{new} 的流程。我们提出通过交叉拟合构造估计量 $\widehat{\eta}_1$。具体而言，我们通过以下方式对其进行估计

$$\widehat{\eta}_1(\pi) = \frac{1}{\sum_i T_i} \sum_{\ell=1}^{\mathbb{L}} \left\{ \sum_{i\in\mathcal{I}_\ell} \sum_{t=0}^{T_i-1} \psi(O_{i,t}; \pi, \pi_{\mathrm{old}}^{(\ell)}, \widehat{V}^{(\ell)}, \widehat{A}^{(\ell)}, \widehat{\omega}^{(\ell)}, \widehat{d}^{(\ell)}) \right\}. \tag{19}$$

请注意，干扰函数 $\pi^{(\ell)}$ 旧、$\widehat{A}^{(\ell)}$、$\widehat{V}^{(\ell)}$、$\widehat{\omega}^{(\ell)}$ 和 $\widehat{d}^{(\ell)}$ 是基于 $\mathcal{I}_\ell^c$ 中的数据子集计算得到的，并且与用于构造估计函数 ψ 的 $\mathcal{I}_\ell$ 中的观测值相互独立。该估计量的理论性质将在第4.2节中研究。

我　们接着提出通过求解以下约束优化来学习 π_{new} ，

$$\begin{aligned} \pi_{\mathrm{new}} \in & \mathrm{argmax}_{\pi\in\Pi} \widehat{\eta}_1(\pi), \\ & \text{subject to } \frac{1}{\mathbb{L}} \sum_{\ell=1}^{\mathbb{L}} \mathbb{E}_{S^*\sim\widehat{d}^{(\ell),\nu}} \mathcal{D}_{\mathrm{KL}} \left(\pi_{\mathrm{old}}^{(\ell)}(\bullet|S^*), \pi(\bullet|S^*) \right) \le \delta, \end{aligned} \tag{20}$$

其中 $\widehat{d}^{(\ell),\ \nu}$表示伪样本 $\{\widetilde{S}_{t'}^{(m),\nu}\}_{1\le m\le M,\ 0\le t\le T'}$的分布，这些伪样本由.....生成

根据算法3。

我们指出，当初始策略 $\pi^{(\ell)}$ 旧策略 设为行为策略 b 时，式（20）中的约束要求学习得到的策略在批次数据集中接近行为策略，这也是近期开发的强化学习算法中常用的方法，例如 Wu 等人（2019）。正如 Levine 等人（2020）所指出的，离线强化学习的根本挑战之一是由于行为策略与目标策略不匹配而导致的分布外问题。分布外问题会导致价值函数被高估，进而降低策略学习的性能。限制学习得到的策略，使其保持接近行为策略，有可能缓解这一局限。实践中，我们可以通过将 $\ell = 1$、$\cdots$、$\mathbb{L}$ 的初始策略 $\pi^{(\ell)}$ 旧策略 设为上一迭代得到的 π_{new}，来多次重复式（20）中的约束优化。这能保证最终估计得到的策略实现快速收敛。最后，我们指出，与现有最先进的强化学习算法相比，我们的方法并不复杂，而且相当灵活。尽管在下面的数值实验中，我们需要学习多个组成部分并使用深度学习模型，但也可以改用简单得多的方法来估计这些组成部分（例如参数模型、筛选方法或核函数）。在这种情况下，我们提出的算法更易于使用。

4 理论

本节系统地研究我们算法的理论性质。在第 4.1 节中，我们确立了估计的最优策略的性质。在第4.2节中，我们证明所提出的一阶价值差估计量 $\widehat{\eta}_1$ 具有高效性。为简化理论分析，我们假设 $T_1 = \cdots = T_N = T$。当轨迹数量 N或决策点数量 T趋于无穷时，我们推导所有渐近结果。此类结果为强化学习中的各种应用提供了有用的理论保证。我们将此类理论称为双向理论。我们还允许状态-动作空间、转移矩阵 p、奖励函数r 和策略类 Π 依赖于 N 和 T。因此，最优策略、Q函数和折扣访问概率可以随 N 或 T变化。

4.1 估计的最优策略的性质

我们首先在定理 1 中证明，新策略与旧策略之间的价值差为 $O(\sqrt{\delta})$，其中 δ 对应于信赖域约束（20）中的阈值。因此，通过设定 $\delta \to 0$，我们可以保证新策略在平均意义上渐近地不劣于旧策略。

定理 1 $|\mathcal{V}(\pi_{\text{new}}) - \mathbb{L}^{-1}\sum_{\ell=1}^{\mathbb{L}} \mathcal{V}(\pi^{(\ell)}{}_{\text{旧}})| \leq O(1)\sqrt{\delta}$，其中 $O(1)$ 表示某个正常数。

我们指出，定理 1 不要求对估计的干扰函数 $\widehat{Q}^{(\ell)}$、$\widehat{\omega}^{(\ell)}$ 和 $\widehat{d}^{(\ell)}$ 作出任何条件限制。它也不要求 $\pi^{(\ell)}{}_{\text{旧}}$ 收敛到最优策略 π^{opt}。此外，即使策略 $\pi^{(\ell)}{}_{\text{旧}}$ 依赖于数据，定理 1 仍然确定性地成立。更多细节请参见附录A.1。

注意，$N \times T$ 对应于决策点的总数。接下来，我们考虑输入策略 $\pi^{(\ell)}{}_{\text{旧}}$ 接近 π^{opt} 的情形，即满足 $\mathcal{V}(\pi^{(\ell)}{}_{\text{旧}}) = \mathcal{V}(\pi^{opt}) + O\{(NT)^{-\kappa_0}\}$，其中 $\kappa_0 > 0$ 为某个常数。这意味着，无论是 N 还是 T 趋于无穷大，$\pi^{(\ell)}{}_{\text{旧}}$ 都与 π^{opt} 一致。我们进一步假设，当 $NT \to \infty$ 时，$\mathcal{V}(\pi^*) = \mathcal{V}(\pi^{opt}) + o(1)$。回顾一下，$\pi^*$ 被定义为在 Π 中价值最大的类内最优策略。换言之，随着样本量增加，类内最优策略下的价值趋近于最优价值函数（因为策略类的规模也随之增大）。为简化理论分析，我们假设 $\pi^{opt} \in \Pi$，使得 $\pi^* = \pi^{opt}$。同时，我们的理论同样适用于 $\pi^{opt} \notin \Pi$ 的情形，只要价值差 $\mathcal{V}(\pi^*) - \mathcal{V}(\pi^{opt})$ 以足够快的速率收敛。在实践中，当我们拥有关于 π^{opt} 的参数形式的领域知识，或使用具有通用逼近能力的函数类（例如神经网络）来参数化 Π 时，这一假设是合理的。为建立价值提升性质，我们需要以下一组条件。

(C1) 假设对于某个常数 $\kappa_1 \geq 0$，$\mathbb{E}_{(A,S)\sim p_\infty}|\widehat{Q}^{(\ell)}(A,S) - Q^{\pi^{(\ell)}{}_{\text{旧}}}(A,S)|^2 = O_p\{(NT)^{-2\kappa_1}\}$。此外，$\widehat{Q}^{(\ell)}$ 几乎必然一致有界。

(C2) 假设 $\mathbb{E}_{(A,S),\ (\widetilde{A},\widetilde{S})\sim p_\infty}|\widehat{\omega}^{(\ell)}(\widetilde{A},\widetilde{S};\ A,S) - \omega^{\pi^{(\ell)}{}_{\text{旧}}}(\widetilde{A},\widetilde{S};\ A,S)| = O_p\{(NT)^{-2\kappa_2}\}$ 对于某个常数 $\kappa_2 \geq 0$，其中 (A,S) 和 $(\widetilde{A},\widetilde{S})$ 表示根据 p_∞ 生成的两个独立状态-动作对。此外，$\widehat{\omega}^{(\ell)}$ 几乎必然一致有界。

(C3) 假设 $\mathbb{E}_{(A,S)\sim p_\infty}|\mathcal{D}_{\text{TV}}(d^{\pi^{(\ell)}{}_{\text{旧的}}}(\bullet|A,\ S),\ \widehat{d}^{(\ell)}(\bullet|A,\ S))|^2 = O_p\{(NT)^{-2\kappa_3}\}$ 对于某个常数 $\kappa_3 \geq 0$。

(C4) 假设 Π对应于某个 VC 型函数类（Chernozhukov等人，2014），其 VC 指数的上界为 $O\{(NT)^{\kappa_4}\}$，其中 α定义如下。

(C5) 最优策略是唯一的。此外，存在一些正常数 $\alpha, \bar{c}, \bar{\epsilon}$，使得对于任意 $a \in \mathcal{A}$ 和 $0 < \epsilon \leq \bar{\epsilon}$，$\Pr(-\epsilon \leq A^{\pi^{opt}}(a, S^*) < 0) \leq \bar{c}\epsilon^{\alpha}$；其中，随机变量 S^* 服从分布 $d^{\pi^{opt}}, \nu$。

(C6) 过程 $\{(S_t, A_t, R_t)\}_{t \geq 0}$ 是指数 β-混合的。

条件（C1）-（C3）分别刻画了对（a）-（c）中学习器的理论要求。具体而言，（C1）-（C2）要求估计得到的 Q函数和条件概率比的平方预测损失满足一定的收敛速率，而条件（C3）则假设转移函数与其估计量之间的平方全变差范数满足一定的收敛速率。如果采用某些参数模型来学习 $Q^{\pi_{\text{old}}}$, $\omega^{\pi_{\text{old}}}$ 和转移矩阵 p，则有 $\kappa_1 = \kappa_2 = \kappa_3 = 1/2$。在我们的设定中，对于任意互不相交的 $i_1, i_2 \in \{1, 2, 3\}$，我们只要求 $\kappa_{i_1} + \kappa_{i_2} > 1/(2 + 2\alpha)$。见下文定理 2 的陈述。当 $\min_{i \in \{1,2,3\}} \kappa_i > 1/(4 + 4\alpha)$ 时，该条件成立，因此许多非参数估计量都能满足这一条件。该条件也严格弱于近期文献中针对离策略价值评估所提出的要求；那些要求规定干扰函数的收敛速率快于 $(NT)^{-1/4}$（例如，Kallus 和 Uehara（2019））。例如，使用核平滑器（Feng 等，2020）、筛法（Shi、Zhang、Lu 和 Song，2020；Chen 和 Qi，2022）或深度神经网络（Fan 等，2020b）来近似 Q函数时，可以证明，在某些技术条件下，（C2）以 $\kappa_1 = \beta_1/(2\beta_1 + d_S)$ 和 $\beta_1 > d_S/2$ 成立，其中 d_S 表示状态空间的维度，β_1 表示刻画 Q函数光滑性的 H¨older 指数。对于条件概率比函数，也可以得到类似的结果（即最优非参数收敛速率）。如第 3.2.4 节所述，当 T' 和 M 足够大时，（C3）实质上要求条件高斯模型中的估计均值函数和协方差函数以 $(NT)^{-\kappa_3}$ 的速率收敛。在一些正则条件下，同样可以达到最优非参数收敛速率。

条件（C4）较为宽松，因为策略类 Π 是预先指定的。当采用线性策略类时，我们有 $\kappa_4 = \#s$，其中 $\#s$ 表示用于索引该策略类的参数数量。当 Π设为某些深度神经网络时，相应的

VC 维数也可从文献中获得（例如，Harvey 等人，2017）。

文献中通常假设最优策略具有唯一性（C5）（Ertefaie 和 Strawderman，2018；Luckett等人，2020）。(C5) 的第二部分与间隔型条件密切相关，这类条件通常用于界定超额误分类错误（Tsybakov 等人，2004；Audibert 等人，2007），以及点治疗研究中个体化治疗方案的遗憾（Qian 和 Murphy，2011；Luedtke 和 Van Der Laan，2016；Shi、Lu 和 Song，2020）。

为了更好地理解 (C5) 中的间隔条件，我们首先注意到，对于任意 a 和 s，$A^{\pi^{opt}}(a, s) \leq 0$。具体来说，我们注意到，对于任意 π，π^{opt} 在 $V^{\pi}(s)$ 处取得最大值。考虑以下依赖历史的策略 $\pi^{opt}(a)$：它在初始决策点采取 a，并在后续步骤中遵循 π^{opt}。该策略下的价值为 $Q^{\pi^{opt}}(a, s)$。由此可得，$Q^{\pi^{opt}}(a, s) \leq V^{\pi^{opt}}(s)$。或者等价地，对于任意 a 和 s，$A^{\pi^{opt}}(a, s) \leq 0$。仅当 $a = \text{argmax}_{a'} Q^{\pi^{opt}}(a', s)$ 时等号成立。根据最优策略的唯一性，该 argmax 有明确定义。当 $a \neq \text{argmax}_{a'} Q^{\pi^{opt}}(a', s)$ 时，优势函数对应于 $\pi^{opt}(a)$ 与 π^{opt} 之间的价值差。差值越小，识别最优策略就越困难。因此，为确保能够一致地识别 π^{opt}，合理的做法是进一步假设 $\Pr(0 < |A^{opt}(a, S^*)| \leq \epsilon)$ 也随 ϵ 趋于零。(C5) 明确刻画了这种依赖关系。例如，当动作空间为二元时，令 $\tau(s)$ 表示对比函数，即 $\tau(s) = Q^{\pi^{opt}}(1, s) - Q^{\pi^{opt}}(0, s)$。显然，$A^{\pi^{opt}}(0, s) = \min(-\tau(s), 0)$，且 $A^{\pi^{opt}}(1, s) = \min(\tau(s), 0)$。因此，(C5) 的第二部分实质上要求 $\Pr(0 < |\tau(S^*)| \leq \epsilon) \leq \bar{c}\epsilon^{\alpha}$；当 $\tau(S^*)$ 具有有界的概率密度函数时，该条件在 $\alpha = 1$ 下自动满足。更一般地，当 $|\tau(S^*)|^{\alpha}$ 具有有界的概率密度函数时，该条件也成立。例如，假设初始参考分布 ν 和马尔可夫转移函数在 $(0, +\infty)$ 上均具有有界密度函数。那么，S^* 的分布，即以 $\{(1-\gamma)\gamma^t\}_{t \geq 0}$ 为权重的 $\{S_t\}_{t \geq 0}$ 混合分布，也具有有界的概率密度函数。假设 $\tau(S^*) = (S^*)^{1/\alpha}$。显然，$|\tau(S^*)|^{\alpha}$ 具有有界的概率密度函数。最后，当 $|\tau(S^*)|$ 一致地有正下界时，(C5) 在 $\alpha = +\infty$ 下成立。我们将在下文的定理 2 中看到，π_{new} 的收敛速率在很大程度上取决于间隔参数 α。

假设（C6）刻画了数据观测值随时间的依赖关系。它实质上要求 β-混合系数（详细定义参见 Bradley，2005）在滞后 β 时（

定义）以指数速率随 q 衰减至零。该系数衡量变量集 $\{(S_j, A_j, R_j)\}_{j\le t}$ 与 $\{(S_j, A_j, R_j)\}_{j\ge t+q}$ 之间的时间依赖关系。当 $\{(S_t, A_t, R_t)\}_{t\ge 0}$ 构成几何遍历马尔可夫链时，该假设自动成立。几何遍历性的限制性低于现有强化学习文献中的假设；后者要求观测值相互独立（例如，Degris 等人，2012；Farahmand 等人，2016），或服从一致遍历马尔可夫链（例如，Bhandari 等人，2018）。

定理 2（价值提升性质）假设 (C1)-(C6) 成立。若常数$\kappa_1, \kappa_2, \kappa_3$ 满足 $\kappa_{i_1} + \kappa_{i_2} > 1/(2+2\alpha)$, 对于任意不相交的 $i_1, i_2 \in \{1, 2, 3\}$，且对于任意 ℓ， $\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi^{(\ell)}_{\text{old}}) = O_p\{(NT)^{-\kappa_0}\}$，则有 $\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{new}}) = E_1 + E_2$，其中 $E_1 = O_p\{(NT)^{-\frac{\kappa_0}{(2\alpha+1)\alpha+1}}\}$, $E_2 = o_p\{(NT)^{-1/2}\}$。

定理 2 表明，价值差 $\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{new}})$ 可分解为两项。第一项 E_1 描述输入策略 π_{old} 如何发挥作用。这是由于对价值差 $\mathcal{V}(\pi) - \mathcal{V}(\pi_{\text{old}})$ 进行一阶近似时产生了高阶余项 $\eta_2(\pi, \pi_{\text{old}})$。第二项 E_2 源于 $\eta_1(\pi, \pi_{\text{old}})$ 的估计误差。在典型的多重稳健设定中，通常要求 $\kappa_{i_1} + \kappa_{i_2} > 1/2$，以使对 $\eta_1(\pi, \pi_{\text{old}})$ 的估计偏差为 $o_p\{(NT)^{-1/2}\}$。在定理 2 中，由于我们对干扰参数要求较慢的收敛速率，所提出的 $\eta_1(\pi, \pi_{\text{old}})$ 价值差估计量可能以慢于 $1/2$ 阶的速率收敛。然而，即使在这种较慢的速率要求下，仍可建立价值提升性质。这是因为定理 2 关注的是估计的最优策略 π_{new} 在价值方面的收敛速率，而非所提出的价值差估计量（记为 $\widehat{\eta}_1(\pi_{\text{new}})$）的收敛速率。具体而言， π_{new} 在价值方面的收敛速率主要由 $\widehat{\eta}_1(\pi_{\text{new}})$ 与 $\widehat{\eta}_1(\pi^{opt})$ 之间的差异决定（见补充材料第 12 页），其收敛速率快于 $\widehat{\eta}_1(\pi_{\text{new}})$ 本身。这是因为，根据条件 $\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi^{(\ell)}\text{old}) = O_p\{(NT)^{-\kappa_0}\}$， π_{new} 与最优策略一致。

当 $\kappa_0 \le 1/2$ 时，可以看出，输出策略下的价值比输入策略收敛得更快，从而实现所需的“价值提升性质”。可以多次重复单步更新，以确保估计的最优策略的价值以 $o_p\{(NT)^{-1/2}\}$ 的速率收敛。当初始策略

的收敛速度已经快于参数速率（例如 $\kappa_0 > 1/2$）时，我们的方案在理论上无法保证产生更好的策略。然而，正如实证研究所示（参见第5节），所提出的策略的价值通常高于通过最先进的强化学习算法计算得到的价值。这表明，尽管这些初始策略是一致的，但它们可能以次优速率收敛，因此仍有改进空间。

4.2 价值差估计量的效率

在本小节中，我们证明，给定 $\pi^{(\ell)}$ 旧，针对 $\eta_1(\pi, \pi^{(\ell)}$ 旧$)$ 的所提出的估计量，即，

$$\widehat{\eta}_1(\pi, \pi_{\text{old}}^{(\ell)}) = \frac{1}{T|\mathcal{I}_\ell|} \left\{ \sum_{i \in \mathcal{I}_\ell} \sum_{t=0}^{T-1} \psi(O_{i,t}; \pi, \pi_{\text{old}}^{(\ell)}, \widehat{V}^{(\ell)}, \widehat{A}^{(\ell)}, \widehat{\omega}^{(\ell)}, \widehat{d}^{(\ell)}) \right\}.$$

是 $\eta_1(\pi, \pi^{(\ell)}$ 旧$)$ 的近似无偏估计量，且其渐近方差与该效率界相匹配。效率界的概念见《补充材料》第 A.4 节。因此， $\widehat{\eta}_1$ 是有效的。

令 $\bar{p}$ 表示给定 (A_t, S_t) 时 (R_t, S_{t+1}) 的条件分布。对于任意给定的 π 和 π_{old}，注意到 $\eta_1(\pi, \pi_{\text{old}})$ 完全由转移函数 $\bar{p}$ 决定。令 $\{\bar{p}_{\theta_1} : \theta_1 \in \Theta_1\}$ 为 $\bar{p}$ 的一个正则参数子模型。这要求对于任意 θ_1 ，$\bar{p}_{\theta_1}$ 都是一个 $^-$ 转移矩阵，并且存在某个 $\theta_1^* \in \Theta_1$，使得 $\bar{p} = \bar{p}_{\theta_1^*}$。类似地，令 $\{b_{\theta_2} : \theta_2 \in \Theta_2\}$ 和 $\{\bar{\nu}_{\theta_3} : \theta_3 \in \Theta_3\}$ 分别为行为策略和初始状态分布的正则参数子模型。令 $\theta = (\theta_1, \theta_2, \theta_3)$ 和 $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*)$，其中 θ_2^* 和 θ_3^* 分别对应于 Θ_2 和 Θ_3 中的真实参数。在由 θ 索引的给定子模型下，单条数据轨迹的对数似然函数可写为

$$\ell_T(\{O_t\}_t; \theta) = \log \left[\bar{\nu}_{\theta_3}(S_0) \prod_{t=0}^{T} \{\bar{p}_{\theta_1}(R_t, S_{t+1} | A_t, S_t) \bar{b}_{\theta_2}(A_t | S_t)\} \right].$$

注意， η_1 也可以定义为 θ 的函数。我们将效率界定义为

$$\text{EB}(|\mathcal{I}_\ell|, T) = |\mathcal{I}_\ell| \sup \nabla_\theta \eta_1(\theta^*) \left\{ \mathbb{E} \nabla_\theta \ell_T(\{O_t\}_t; \theta^*) \nabla_\theta^\top \ell_T(\{O_t\}_t; \theta^*) \right\}^{-1} \nabla_\theta^\top \eta_1(\theta^*),$$

其中，上确界是对所有正则参数子模型取的，且 $\nabla_\theta g(\theta')$表示函数 g 关于 θ的导数，在 $\theta = \theta'$处取值。如前所述，η_1 仅通过 θ_1依赖于 θ 。

定理 3 假设定理 2 中的条件成立，且 $\kappa_{i_1}+\kappa_{i_2}>1/2$ 对任意不相交的i_1、 $i_2\in\{1,2,3\}$均成立。那么，在给定 $\pi^{(\ell)}$ 旧的条件下，对任意 π，我们有

$$\frac{\widehat{\eta}_1(\pi,\pi^{(\ell)}_{\text{old}})-\eta_1(\pi,\pi^{(\ell)}_{\text{old}})}{\sqrt{\text{EB}(|\mathcal{I}_\ell|,T)}}\xrightarrow{d}N(0,1).$$

定理 3 表明， $\widehat{\eta}_1$ 是渐近无偏的，其渐近方差为 EB($|\mathcal{I}_\ell|$,T)。这体现了所提出的估计量的效率。

5 数值示例

本节通过一个玩具示例和真实数据相关研究，展示我们方法的优越性能。具体而言，在第5.1节中，我们使用一个玩具示例，展示估计量的多重稳健性和价值提升性质。随后，我们在第5.2节中展示所提出的方法在 OhioT1DM 相关数据集上的表现。附录D中，我们还开展了另一项模拟研究，以说明与若干现有方法相比，我们的算法在有限样本下的性能。

5.1 玩具示例

我们设计了一个玩具示例，以说明估计量的多重稳健性和所期望的价值提升性质。考虑二元状态空间 $\mathcal{S}=\{0,1\}$，其中 S_0以 0.4 的概率取值为 0，否则取其他值。动作空间 $\mathcal{A}$ 的取值为 $\{0,1\}$。奖励函数定义为 $r(a,s)=\mathbb{I}(s=a)$，随后按照 $R_t=r(A_t,S_t)+e_t$生成奖励，其中 $\{e_t\}_{0\le t<T}$是一列独立同分布的 $N(0,2)$ 随机误差。 $p(S'|A,S)$ 的转移矩阵和行为策略可见补充材料的 D 节。

在这个表格情形中，可以使用蒙特卡洛方法模拟 Q^π、 ω^π 和 p 的预言机值。为展示三重稳健性，我们将在 Q^π、 ω^π 或 p 上加入一些随机误差，使其产生偏差。随后，我们使用这些干扰函数计算 η_1 ，并通过所提出的算法得到相应的估计最优策略。具体而言，我们考虑以下五种干扰函数估计量组合：（i）“原点”：所有干扰函数均设为其预言机值。（ii）“模块1”： ω^π 设为有偏值，其他干扰函数均为预言机值；（iii）“模块2”： Q^π 设为有偏值，其他干扰函数均为预言机值。（iv）“模块3”：转移 p 设为有偏

值，其他干扰函数均为预言机值。（vi）“模块4”：所有干扰函数均设为偏差值。有关这些场景的详细信息，请参见补充材料的 D节

总而言之，场景（i）对应于所有干扰函数均正确设定的预言机设定。在场景（ii）—（iv）中，有一个干扰函数设定错误。在最后一个场景中，所有干扰函数均设定错误。我们还改变初始策略，以考察价值提升性质。具体而言，我们用一个 2×2 矩阵表示初始策略 π_{old}，并考虑以下参数化形式，

$$\pi_{\text{old}} = \begin{array}{c} \\ S=0 \\ S=1 \end{array} \begin{array}{c} \begin{array}{cc} A=0 & A=1 \end{array} \\ \begin{pmatrix} \kappa & 1-\kappa \\ 1-\kappa & \kappa \end{pmatrix} \end{array},$$

其中 $\kappa \in [0,1]$。根据我们的数据生成机制，$\kappa = 0$ 对应于最优策略。κ 越接近 1，初始策略就越差。我们考虑 κ的三种取值，分别为 0.2、0.5 和 0.8，由此得到三种不同的初始策略。我们进一步考虑两种样本量，$N = 30$、$T = 30$ 和 $N = 50$、$T = 50$，共得到 $5 \times 3 \times 2 = 30$ 种设定。γ 设为 0.9。可以证明，最优价值 $\mathcal{V}$（π^{opt}）等于（$1-\gamma$）$^{-1} = 10$。最后，我们考虑 δ 的三种取值（见公式（20）），分别为 0.05、0.1 和 0.2。

结果如图1、图4和图5所示（参见补充文章中的附录D）。所有估计策略的值均通过蒙特卡洛模拟计算得到。可以观察到，在场景(i)–(iv)中，我们提出的算法使用(i)–(iv)中的模型，显著提升了初始策略的表现，展示了预期的价值提升性质。特别是，当两个干扰函数模型中的任意一个存在错误设定时，所提出的方法仍然有效。这一结果从经验上验证了三重稳健性。

此外，当所有模型都存在设定错误时，无法保证所提出的方法一定能提升价值。具体而言，在前两列中，所提出的方法在“模块4”设定下经过几次迭代后提升了初始策略。然而，在最后一列中，估计策略的价值低于初始策略。我们推测，这是因为在模型设定错误的情况下，我们的方法可能会收敛到一个次优策略，其价值介于第二列和最后一列中的初始策略价值之间。因此，当其他方法给出的现有策略

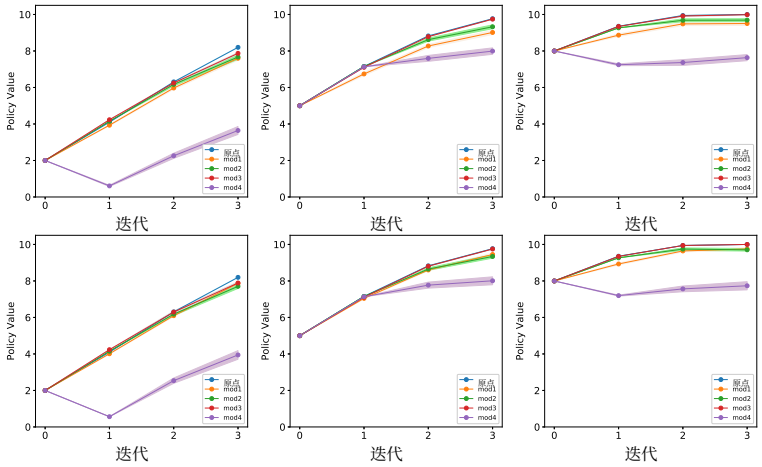

图1：玩具示例中估计策略的价值。第一行表示使用（T, N）对为（30, 30）时的结果，第二行则使用（50, 50）。三列分别表示初始策略因子 κ 取值为 0.8、0.5 和 0.2 时的结果。横轴表示价值增强过程中使用的迭代次数。当迭代次数为零时，我们绘制初始策略的评估值。最优价值为 10，且 δ 固定为 0.1。置信带根据 100 次重复试验计算得出。

已经接近最优策略时，使用不一致的干扰函数估计器可能会降低性能。最后，在初始策略与最优策略相差很大的设定下（即附录D中图1、图4和图5的第一列），我们的方法需要更多迭代次数和更大的 δ，才能实现更高的价值。相反，当初始策略接近最优策略时，所需的迭代次数较少，较小的 δ 更为合适。例如，从附录D中图5的第三列可以看出，当 $\delta = 0.2$ 时，使用模型(ii)和(iii)得到的估计策略的价值在第三次迭代时下降。

最后，为了进一步展示所提方法的优势，我们使用查找表（例如，带有查表特征的线性模型）而非深度学习模型，对所有干扰函数（包括 Q函数、概率比和转移核）进行参数化，并将所提方法应用于该玩具示例。结果见补充材料中的图S3。可以看出，所提方法仍能提升初始策略的表现。

5.2 在 OhioT1DM 相关数据集上的应用

将强化学习算法应用于移动健康（mHealth）领域的兴趣日益浓厚。本节基于 OhioT1DM 数据集进行了两项分析。

在第一项分析中，我们生成合成数据来模拟该真实数据集，并将我们的方法应用于合成数据集。具体而言，我们使用 Shi、Wan、Song、Lu 和 Leng（2020）第5.2.2节设计的仿真环境生成数据。在第二项分析中，我们将我们的方法应用于真实数据集。

OhioT1DM 数据集包含六名 1 型糖尿病患者连续八周的测量数据。状态 S_t 由三个状态组成，分别对应平均血糖水平、餐食的碳水化合物估算值和运动强度。动作 A表示胰岛素剂量。我们对动作空间进行离散化，并考虑五种动作，即 $\mathcal{A} = \{0, 1, 2, 3, 4\}$ 从不注射到高剂量注射。Shi、Wan、Song、Lu 和 Leng（2020年）提出的马尔可夫检验表明，这些数据很可能满足四阶马尔可夫性质，因此我们通过拼接过去的测量值来重构状态变量$\tilde{S}_t = (S_{t-3}, A_{t-3}, S_{t-2}, A_{t-2}, S_{t-1}, A_{t-1}, S_t)$，以满足马尔可夫假设。由此得到一个 15 维状态向量。奖励 R_t 定义为血糖控制指数，它是时间区间 $[t, t+1$内平均血糖水平的确定性函数。

我们首先将所提出的方法应用于合成数据集。我们使用拟合Q迭代（FQI）和保守Q学习（CQL）来计算初始策略。此处未实现V学习（VL），因为当数据集较大时，该方法需要较高的计算成本。如图2所示，经过3次所提出算法的迭代后，我们能够获得近似最优策略。在所有情况下，估计的最优策略所取得的值都高于初始策略。当 π_{old} 与最优策略相差较远时，改进尤为显著。

接下来，我们将所提方法应用于真实数据集。为评估估计的最优策略，我们将数据划分为训练集和测试集。在训练数据上得到估计的最优策略后，我们在测试数据上应用 FQE，计算所有这些估计策略的策略价值。图 3 展示了这些价值。结果表明，经过 2 至 3 次迭代后，所提算法会得到价值更高的策略。最后，我们将基于所提算法得到的估计最优策略应用于整个数据集，初始策略则由保守Q学习计算。我们的估计策略对各个动作（$A = 0$、1、$\cdots 4$）的推荐总体占比分别为 15.2%、0.5%、2%、6% 和 76%。结果表明，在大多数场景下，我们的估计策略会推荐最大剂量，同时也有一定比例建议不使用胰岛素剂量。

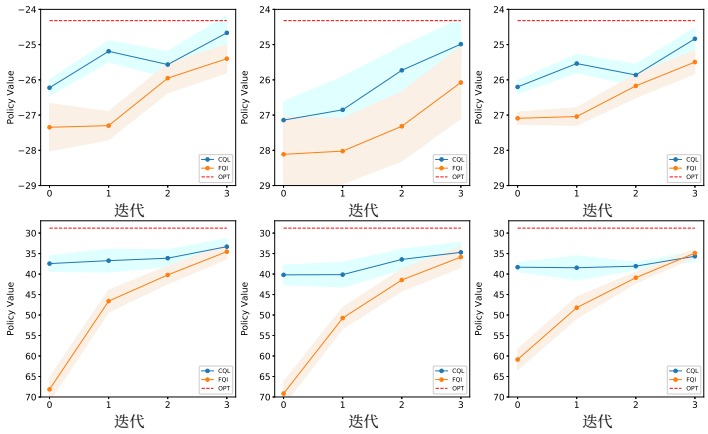

图 2：基于真实数据的模拟研究中各种策略的价值。初始策略由保守Q学习和拟合Q迭代计算得出。第一行展示使用 $\gamma = 0.9$ 的结果，第二行展示使用 $\gamma = 0.95$的结果。最优价值分别为 -24.32 和 -28.79 （如红色虚线所示）。三列分别展示使用（T, N）组合作为（50, 100）、（25, 200）和（100, 50）的结果。置信带根据 100 次重复试验计算得出。

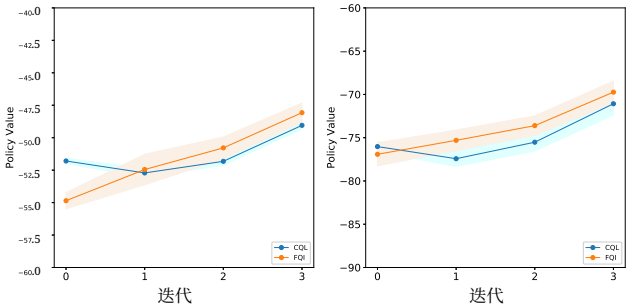

图 3：基于真实数据集计算的各策略价值。初始策略由保守Q学习和拟合Q迭代计算得到。左图对应 $\gamma = 0.9$ 的结果，右图对应 $\gamma = 0.95$。策略价值通过交叉验证过程计算，置信带则基于 100 次重复试验计算。

6 讨论

本文针对离线强化学习问题提出了一种价值增强策略优化方法。所提方法的一个关键要点，是为一阶线性项 η_1 构造一个三重稳健估计量，该项用于衡量任意两个策略之间的差异。因果推断领域已有大量关于多重稳健估计量的研究。例如，Tchetgen Tchetgen 和 Shpitser（2012年）针对因果中介分析，提出了边际自然间接效应和直接效应的三重稳健估计量。Wang 和 Tchetgen Tchetgen（2018年），以及 Shi、Miao、Nelson 和 Tchetgen Tchetgen（2020年），利用工具变量开发了用于估计平均处理效应的三重稳健估计量

工具变量和双重负对照变量。Jiang 等（2020年）提出了用于估计主层内因果效应的三重稳健估计量。我们所提出的估计量与这些估计量具有相似的统计性质，因为其一致性只要求三个干扰函数中有两个设定正确。此外，当所有函数设定正确且满足一定的收敛速率时，该估计量具有较高效率。

基于三重稳健估计量，我们提出在信赖域约束下搜索使价值差最大的策略，并通过迭代这一过程来实现价值增强。实践中，可通过交叉验证以数据自适应的方式确定迭代次数。具体而言，可以先将所有数据轨迹划分为 K 个不相交子集 $\cup_{k=1}^{K}\mathcal{I}_k$。接下来，对于每个 k，可以将一部分数据 $\mathcal{I}_k^c$ 用于应用我们提出的方法，计算每次迭代得到的最优策略；再将现有最先进的离策略评估方法（例如，Jiang 和 Li（2016年）；Kallus 和 Uehara（2019年）；Liao 等（2019年）；Shi 等（2021年））应用于剩余部分 $\mathcal{I}_k$，以评估这些策略的价值。随后，我们对不同k 的价值估计量进行汇总，以实现完全效率，并选择使估计价值最大的迭代次数。

此外，我们在数值研究中表明，与其他基线策略相比，所提出的策略能够取得更大的价值。不过，我们想指出，实现我们的方法比实现那些基线方法耗时更长。具体而言，在补充材料 D.2 节所述的设定下，实现所提出的方法进行一次迭代约需 14 分钟。相比之下，V学习的运行时间约为 17 分钟，而保守Q学习约为 4 分钟。我们还想指出，在离线强化学习领域，策略是基于预先收集的数据集计算得出的。我们的主要目标是学习价值尽可能大的最优策略。只要整个过程能在合理的时间内完成，计算时间就不是大问题。这与在线强化学习领域截然不同，在后者中，通常每当有新的观测到来时，策略就会立即更新。不过，研究如何在我们的价值提升算法框架下提高计算效率，将会很有意义。

最后，所提出的方法可以用作独立的策略迭代算法，从完全随机的初始策略开始，并通过迭代更新该策略来提升其性能。然而，所得算法在实际应用中可能会非常

耗费计算资源，因为它可能需要大量迭代次数才能达到近乎最优的策略。

A 一些补充技术细节

A.1 关于信赖域策略优化的更多内容

引理 1 假设 ν 一致地远离零。则存在某个正常数 $c^* > 0$，使得 $|\eta_2(\pi_{\text{new}}, \pi_{\text{old}})|$ 的上界为

$$\frac{c^*\gamma}{1-\gamma}\left[\mathbb{E}^{d^{\pi_{\text{old}},\nu}}\mathcal{D}_{\text{TV}}\{\pi_{\text{old}}(\bullet|S), \pi_{\text{new}}(\bullet|S)\}\right]^2 \leq \frac{c^*\gamma}{1-\gamma}\mathbb{E}^{d^{\pi_{\text{old}},\nu}}\mathcal{D}_{\text{KL}}\{\pi_{\text{old}}(\bullet|S), \pi_{\text{new}}(\bullet|S)\}. \quad (21)$$

我们指出，式 (21) 右端的上界比 Schulman 等人（2015）的定理 1 中的上界更紧。

证明：注意到

$$|\eta_2(\pi_{\text{new}}, \pi_{\text{old}})| \leq \sum_{a\in\mathcal{A}, s\in\mathcal{S}} |\pi_{\text{new}}(a|s) - \pi_{\text{old}}(a|s)||A^{\pi_{\text{old}}}(a,s)||d^{\pi_{\text{old}},\nu}(s) - d^{\pi_{\text{new}},\nu}(s)|. \quad (22)$$

接下来，我们给出 $|d^{\pi_{\text{old}},\ \nu}(s) - d^{\pi_{\text{new}},\ \nu}(s)|$的一个上界。根据定义，我们有

$$|d^{\pi_{\text{old}},\nu}(s) - d^{\pi_{\text{new}},\nu}(s)| \leq (1-\gamma)\gamma\sum_{t=0}^{+\infty}\gamma^t|p_{t+1}^{\pi_{\text{old}}}(s) - p_{t+1}^{\pi_{\text{new}}}(s)|, \quad (23)$$

对于任意 t，我们可以定义一个时变策略 $\pi(t)$，使智能体在最初的 t 个时间点遵循 π_{old}，之后遵循 π_{new}。由此可得

$$\sum_{t=0}^{+\infty}\gamma^t|p_{t+1}^{\pi_{\text{old}}}(s) - p_{t+1}^{\pi_{\text{new}}}(s)| \leq \sum_{t=0}^{+\infty}\sum_{j=0}^{t}\gamma^t|p_{t+1}^{\pi(j+1)}(s) - p_{t+1}^{\pi(j)}(s)|$$

$$\leq \sum_{t=0}^{+\infty}\sum_{j=0}^{t}\gamma^t\left|\sum_{s_-\in\mathcal{S}} p_j^{\pi_{\text{old}}}(s_-)\sum_a p(s|a,s_-)\{\pi_{\text{old}}(a|s_-) - \pi_{\text{new}}(a|s_-)\}p_{t-j}^{\pi_{\text{new}}}(s)\right|.$$

根据全变差距离的定义，我们有

$$\begin{aligned}\sum_{t=0}^{+\infty}\gamma^t|p_{t+1}^{\pi_{\text{old}}}(s)-p_{t+1}^{\pi_{\text{new}}}(s)| &\le 2\sum_{s_-\in\mathcal{S}}\|\pi_{\text{new}}(\bullet|s_-)-\pi_{\text{old}}(\bullet|s_-)\|_{\text{TV}}\sum_{t=0}^{+\infty}\sum_{j=0}^{t}\gamma^t p_{t-j}^{\pi_{\text{new}}}(s)p_j^{\pi_{\text{old}}}(s_-)\\ &= 2\sum_{s_-\in\mathcal{S}}\|\pi_{\text{new}}(\bullet|s_-)-\pi_{\text{old}}(\bullet|s_-)\|_{\text{TV}}\sum_{j=0}^{+\infty}\sum_{t=j}^{+\infty}\gamma^t p_{t-j}^{\pi_{\text{new}}}(s)p_j^{\pi_{\text{old}}}(s_-)\\ &\le \frac{2}{(1-\gamma)^2}\sum_{s_-\in\mathcal{S}}\|\pi_{\text{new}}(\bullet|s_-)-\pi_{\text{old}}(\bullet|s_-)\|_{\text{TV}}d^{\pi_{\text{old}},\nu}(s_-)d^{\pi_{\text{new}},\nu}(s).\end{aligned}$$

在给定条件下，对于某个常数 $C>0$ 和任意 $s\in\mathcal{S}$，我们有 $\nu(s)\ge C$。因此，我们有 $d^{\pi_{\text{old}}}(s)\ge C(1-\gamma)$，进而有 $d^{\pi_{\text{new}}}(s)/d^{\pi_{\text{old}}}(s)\le 1/d^{\pi_{\text{old}}}(s)\le C^{-1}(1-\gamma)^{-1}$。由此可得

$$\sum_{t=0}^{+\infty}\gamma^t|p_{t+1}^{\pi_{\text{old}}}(s)-p_{t+1}^{\pi_{\text{new}}}(s)|\le\frac{2}{C(1-\gamma)^2}\sum_{s_-\in\mathcal{S}}\|\pi_{\text{old}}(\bullet|s_-)-\pi_{\text{new}}(\bullet|s_-)\|_{\text{TV}}d^{\pi_{\text{old}},\nu}(s_-)d^{\pi_{\text{old}},\nu}(s),$$

对于某个常数 $C>0$。将其与 (22) 及上述不等式结合，可得 $|\eta_2(\pi_{\text{new}},\pi_{\text{old}})|$ 的上界为

$$\begin{aligned}\frac{2C\gamma}{1-\gamma}\sum_{a\in\mathcal{A}}\sum_{s,s_-\in\mathcal{S}}&|\pi_{\text{new}}(a|s)-\pi_{\text{old}}(a|s)||A^{\pi_{\text{old}}}(a,s)|\|\pi_{\text{old}}(\bullet|s_-)-\pi_{\text{new}}(\bullet|s_-)\|_{\text{TV}}\\ &\times d^{\pi_{\text{old}},\nu}(s_-)d^{\pi_{\text{old}},\nu}(s).\end{aligned}\tag{24}$$

注意，Q-函数和价值函数对应于某些期望折扣累积奖励。在即时奖励有界的假设下，这两个函数均有界。因此，对于某个正常数 $O(1)$，我们有 $\sup_{a,\,s}|A^{\pi_{\text{old}}}(a,s)|\le c$。由此可得

$$|\eta_2(\pi_{\text{new}},\pi_{\text{old}})|\le\frac{4c\gamma}{C(1-\gamma)}\left\{\sum_{s\in\mathcal{S}}\|\pi_{\text{old}}(\bullet|s)-\pi_{\text{new}}(\bullet|s)\|_{\text{TV}}d^{\pi_{\text{old}},\nu}(s)\right\}^2.$$

这给出了 (21) 左侧（LHS）的上界。

根据柯西-施瓦茨不等式，可得

$$|\eta_2(\pi_{\text{new}},\pi_{\text{old}})|\le\frac{4c\gamma}{C(1-\gamma)}\sum_{s\in\mathcal{S}}\|\pi_{\text{old}}(\bullet|s)-\pi_{\text{new}}(\bullet|s)\|_{\text{TV}}^2d^{\pi_{\text{old}},\nu}(s).$$

因此，根据 Pinsker不等式可得 (21) 右端的上界。

A.2 关于命题 1 的一些补充细节

我们首先对该命题作一些直观解释：具体而言，假设 A 满足 $\sum_a \pi_{\text{old}}(a|s)A(a,s) = 0$，对任意 s 均成立。我们有以下结果：

- 当 $A = A^{\pi_{\text{old}}}$、$V = V^{\pi_{\text{old}}}$ 和 $d = d^{\pi_{\text{old}}}$ 时，期望 $\mathbb{E}\psi_2(O_t;\ \pi, \pi_{\text{old}} V, A, \widetilde{\omega}, d)$ 以及 $\mathbb{E}\psi_3(O_t;\ \pi, \pi_{\text{old}} A, \widetilde{\omega}, d)$ 均为零。因此，$\mathbb{E}\psi(O_t;\ \pi, \pi_{\text{old}} V, A, \widetilde{\omega}, d)$ 等于代入估计量的期望，从而 $\psi(O_t;\ \pi, \pi_{\text{old}} V, A, \widetilde{\omega}, d)$ 对于 η_1 是无偏的。

- 当 $\widetilde{\omega} = \omega^{\pi_{\text{old}}}$ 和 $d = d^{\pi_{\text{old}}}$ 时，期望 $\mathbb{E}\psi_3(O_t;\ \pi, \pi_{\text{old}} A, \widetilde{\omega}, d)$ 等于零。ψ_2 中存在 $\widetilde{\omega}$，保证了估计函数对 A 和 V 的错误设定具有稳健性。在假设 $\sum_a \pi_{\text{old}}(a|s)A(a,s) = 0$ 对任意 s 均成立的情况下，$\psi_1(\pi, \pi_{\text{old}} A, d) + \psi_2(O_t;\ \pi, \pi_{\text{old}} V, A, \widetilde{\omega}, d)$ 的期望等于重要性抽样估计量 (10) 的期望。由于重要性抽样估计量在 $\widetilde{\omega} = \omega^{\pi_{\text{old}}}$ 且 $d = d^{\pi_{\text{old}}}$ 时是无偏的，因此 $\psi(O_t;\ \pi, \pi_{\text{old}} V, A, \widetilde{\omega}, d)$ 也是无偏的。

- 当 $A = A^{\pi_{\text{old}}}$ 且 $\widetilde{\omega} = \omega^{\pi_{\text{old}}}$ 时，期望 $\mathbb{E}\psi_2(O_t;\ \pi, \pi_{\text{old}} V, A, \widetilde{\omega}, d)$ 等于零。ψ_3 中包含 $\widetilde{\omega}^{\nu}$（或 $\widetilde{\omega}$，因为它完全由 $\widetilde{\omega}$ 决定）保证了该估计方程对 d 的错误设定具有稳健性。更具体地说，$\psi_1(\pi, \pi_{\text{old}} A, d) + \psi_3(O_t;\ \pi, \pi_{\text{old}} A, \widetilde{\omega}, d)$ 的期望与重要性抽样估计量 (11) 的期望相同，其中 $\widetilde{\omega}^{\nu} = \omega^{\pi_{\text{old}}\ \nu}$。由于当 $A = A^{\pi_{\text{old}}}$ 且 $\widetilde{\omega} = \omega^{\pi_{\text{old}}}$ 时，重要性抽样估计量是无偏的，因此 $\psi(O_t;\ \pi, \pi_{\text{old}} V, A, \widetilde{\omega}, d)$ 也是无偏的。

现在，我们正式证明，当某些干扰函数被正确设定时，两个论证项 $\psi_2(O_t;\ V, A, \widetilde{\omega}, d)$ 和 $\psi_3(O_t;\ A, \widetilde{\omega}, d)$ 的期望等于零。首先考虑 ψ_2。当 $A = A^{\pi_{\text{old}}}$ 且 $V = V^{\pi_{\text{old}}}$ 时，根据贝尔曼方程可得

$$\mathbb{E}\{R_t + \gamma \widetilde{V}(S_{t+1}) - \widetilde{V}(S_t) - \widetilde{A}(A_t, S_t) | A_t, S_t\} = 0. \tag{25}$$

因此，$\mathbb{E}\{\psi_2(O_t;\ \pi, \pi_{\text{old}}, V, A, \widetilde{\omega}, d) | A_t,\ S_t\} = 0$ 从而 $\mathbb{E}\psi_2(O_t;\ \pi, \pi_{\text{old}}, V, A, \widetilde{\omega}, d) = 0$。至于 ψ_3，注意到当 $d = d^{\pi_{\text{old}}}$ 时，我们有

$$
\begin{aligned}
&\mathbb{E}\left\{ \mathbb{E}_{\substack{a^*\sim\pi_{\rm old}(\bullet|S_{t+1})\\ S^*\sim\widetilde{d}(\bullet|a^*,S_{t+1})}} \pi_{\rm new}(a'|S^*)\widetilde{A}(a',S^*) \,\middle|\, A_t, S_t \right\} \\
=\;& \sum_{s,a} \pi_{\rm new}(a'|s)\widetilde{A}(a',s)\mathbb{E}\{d^{\pi_{\rm old}}(s|a,S_{t+1})\pi_{\rm old}(a|S_{t+1})|A_t,S_t\} \\
=\;& \sum_{s,a} \pi_{\rm new}(a'|s)\widetilde{A}(a',s)(1-\gamma)\sum_{t\geq 1}\gamma^{t-1}p_t^{\pi_{\rm old}}(s;A_t,S_t) \\
=\;& \frac{1}{\gamma}\left\{ \mathbb{E}_{S^*\sim\widetilde{d}(\bullet|A_t,S_t)}\pi_{\rm new}(a'|S^*)\widetilde{A}(a',S^*) - (1-\gamma)\pi_{\rm new}(a'|S_t)\widetilde{A}(a',S_t) \right\},
\end{aligned}
$$

其中，最后一个等式由 d 的定义可得。由此得到 $\mathbb{E}\{\psi_3(O_t;\,\pi,\pi_{\rm old},V,A,\widetilde{\omega},d)|A_t,\;S_t\}=0$ 从而 $\mathbb{E}\psi_3(O_t;\,\pi,\pi_{\rm old},V,A,\widetilde{\omega},d)=0$。

接下来，假设对于任意 s，$\sum_a \pi_{\rm old}(a|s)A(a,s)=0$。我们证明

$$
\widetilde{A}(a,s) - \frac{1}{1-\gamma}\mathbb{E}\omega^{\pi_{\rm old}}(A_t,S_t;a,s)\widetilde{A}(A_t,S_t) = 0, \tag{26}
$$

and

$$
\sum_a \{\pi_{\rm new}(a|s)-\pi_{\rm old}(a|s)\}\mathbb{E}\omega^{\pi_{\rm old}}(A_t,S_t;a,s)\{\gamma\widetilde{V}(S_{t+1})-\widetilde{V}(S_t)\} = 0. \tag{27}
$$

结合这两个等式可得，当 $\widetilde{\omega}$ 设定正确时，期望 $\mathbb{E}\{\psi_1(\pi,\pi_{\rm old},A,\widetilde{\omega})+\psi_2(O_t;\,V,A,\widetilde{\omega},d)\}$等于式 (10) 中重要性抽样估计量的期望。因此，在 (B2) 的假设下，$\mathbb{E}\psi(O_t;\,V,A,\widetilde{\omega},d)$ 是 η_1 的无偏估计量。

我们先证明 (26)。注意到

$$
\frac{1}{1-\gamma}\mathbb{E}\omega^{\pi_{\rm old}}(A_t,S_t;a,s)\widetilde{A}(A_t,S_t) = \widetilde{A}(a,s) + \sum_{t\geq 1}\gamma^t\sum_{a',s'}p_t(s'|a,s)\pi_{\rm old}(a'|s')\widetilde{A}(a',s').
$$

在满足对于任意 s，$\sum_a \pi_{\rm old}(a|s)A(a,s)=0$ 的条件下，右端的第二项等于零。由此得到 (26)。接下来证明 (27)。经过一些计算，

$$
\begin{aligned}
\mathbb{E}\omega^{\pi_{\rm old}}(A_t,S_t;a,s)\{\gamma\widetilde{V}(S_{t+1})-\widetilde{V}(S_t)\} &= (1-\gamma)\sum_{t\geq 1}\gamma^t\sum_{s'}p_t(s'|a,s)\widetilde{V}(s') \\
&\quad -(1-\gamma)\sum_{t\geq 0}\gamma^t\sum_{s'}p_t(s'|a,s)\widetilde{V}(s') = -(1-\gamma)\widetilde{V}(s).
\end{aligned}
$$

注意，右端与 a 无关。因此，我们有 $\sum_a\{\pi(a|s)-\pi_{\rm old}(a|s)\}V(s)=0$。由此得到 (27)。

公式 (27) 表明，当 $\widetilde{\omega}=\omega^{\pi_{\rm old}}$ 时，$\mathbb{E}\psi_2(O_t;\,V,A,\widetilde{\omega},d)=\mathbb{E}\psi_2(O_t;\,V^{\pi_{\rm old}},A,\widetilde{\omega},d)$。如果进一步正确指定了 A，则可得 $\mathbb{E}\psi_2(O_t;\,V,A,\widetilde{\omega},d)=0$。

最后，我们展示

$$\mathbb{E}\sum_{a^*\in\mathcal{A}}\frac{\omega^{\pi_{\text{old}},\nu}(A_t,S_t)}{1-\gamma}\left[\gamma\mathbb{E}_{\substack{a'\sim\pi_{\text{old}}(\bullet|S_{t+1})\\ S^*\sim\widetilde{d}(\bullet|a',S_{t+1})}}\widetilde{A}(a^*,S^*)\pi(a^*|S^*)-\mathbb{E}_{S^*\sim\widetilde{d}(\bullet|A_t,S_t)}\widetilde{A}(a^*,S^*)\pi(a^*|S^*)\right]$$
$$=-\psi_1(\pi,\pi_{\text{old}},\widetilde{A},\widetilde{d}).$$

由此进一步可得，$\psi_1(\pi,\pi_{\text{old}},A,d)+\psi_3(O_t;\,A,\widetilde{\omega},d)$ 的期望等于在 $\widetilde{\omega}^\nu=\omega^{\pi_{\text{old}},\,\nu}$ 时重要性抽样估计量 (11) 的期望。因此，当 A 和 $\widetilde{\omega}$ 被正确指定时，$\mathbb{E}\psi(O_t;\,V,A,\widetilde{\omega},d)$ 是无偏的。

经过一些计算，左边等于

$$\mathbb{E}\sum_{a^*,s^*}\frac{\omega^{\pi_{\text{old}},\nu}(A_t,S_t)}{1-\gamma}\left[\left\{\gamma\sum_{a'}\pi_{\text{old}}(a'|S_{t+1})\widetilde{d}(s^*|a',S_{t+1})-\widetilde{d}(s^*|A_t,S_t)\right\}\widetilde{A}(a^*,s^*)\pi_{\text{new}}(a^*|s^*)\right]$$
$$=\frac{1}{1-\gamma}\sum_{a^*,s^*,a',s'}\pi_{\text{new}}(a^*|s^*)\pi_{\text{old}}(a'|s')\widetilde{d}(s^*|a',s')\widetilde{A}(a^*,s^*)\{d^{\pi_{\text{old}}}(s')-(1-\gamma)\nu(s')\}$$
$$-\frac{1}{1-\gamma}\sum_{a^*,s^*,a',s'}\pi_{\text{new}}(a^*|s^*)\pi_{\text{old}}(a'|s')\widetilde{d}(s^*|a',s')\widetilde{A}(a^*,s^*)d^{\pi_{\text{old}}}(s')$$
$$=-\sum_{a^*,s^*,a',s'}\pi_{\text{new}}(a^*|s^*)\pi_{\text{old}}(a'|s')\widetilde{d}(s^*|a',s')\widetilde{A}(a^*,s^*)\nu(s')=-\sum_{a^*,s^*}\pi_{\text{new}}(a^*|s^*)\widetilde{d}^{\nu}(s^*)\widetilde{A}(a^*,s^*),$$

其中最后一个等式由 d^ν 的定义得出。证明至此完成。

A.3 VC 类型类

本节介绍 VC 类型类的概念。具体而言，设 $\mathcal{F}$ 表示一类可测函数，其可测包络函数为 F，满足 $\sup_{f\in\mathcal{F}}|f|\le F$。对于任意概率测度 Q，设 e_Q 表示 $\mathcal{F}$ 上的半度量，满足 $e_Q(f_1,f_2)=\|f_1-f_2\|_{Q,2}=\sqrt{Q|f_1-f_2|^2}$。空间 $(\mathcal{F},e_Q)$ 的 ϵ-网是 $\mathcal{F}$ 的一个子集 $\mathcal{F}_\epsilon$，使得对每个 $f\in\mathcal{F}$，都存在某个 $f_\epsilon\in\mathcal{F}_\epsilon$ 满足 $e_Q(f,f_\epsilon)<\epsilon$。若存在常数 $c_0>0,c_1\ge 1$，使得对于所有 $0<\epsilon\le 1$，都有 $\sup_Q\mathbb{N}(\mathcal{F},e_Q,\epsilon\|F\|_{Q,2})\le(c_0/\epsilon)^{c_1}$，则称 $\mathcal{F}$ 为包络函数为 F 的 VC 类型类。其中，上确界取遍 $\mathcal{F}$ 的支撑集上的所有有限离散概率测度，而 $\mathbb{N}(\mathcal{F},e_Q,\epsilon\|F\|_{Q,2})$ 是 $\mathcal{F}$ 的 $\epsilon\|F\|_Q$ 个 2-网的基数的下确界。我们称 c_1 为 $\mathcal{F}$ 的 VC 指数。

A.4 半参数效率

在独立同分布情形下，对于参数模型，任何无偏估计量的方差都必须大于或等于 Cr´amer-Rao 下界（Casella and Berger，2002，第 7.3 节），而极大似然估计量已知在满足某些正则

性条件时具有有效性。在半参数理论中，效率界定义为所有正则参数子模型的 Cr´amer-Rao 下界的上确界，从而超越参数设定（例如，参见 Tsiatis，2007）。在我们的设定中，观测值依赖于时间。我们采用 Komunjer 和 Vuong（2010）以及 Kallus 和 Uehara（2019）中的定义，该定义将经典半参数效率界推广至非独立同分布设定。

B 关于算法的更多内容

B.1 关于条件折扣平稳概率比的更多内容

考虑以下优化问题

$$\operatorname{argmin}_{\omega\in\Omega}\sup_{f\in\mathcal{F}}\left|\sum_{(i,t)\neq(i',t')}\Delta(\omega,f,\pi_{\text{old}};i,t,i',t')\right|^2. \tag{28}$$

我们将 $\mathcal{F}$ 设为再生核希尔伯特空间（RFHS）的单位球，即 $\mathcal{F}=\{f\in\mathcal{H}:\|f\|_{\mathcal{H}}=1\}$，其中

$$\mathcal{H}=\left\{f(\cdot)=\sum_{(i,t)\neq(i',t')}b_{i,t,i',t'}\kappa(X_{i',t'},X_{i,t};\cdot):b_{i,t,i',t'}\in\mathbb{R}\right\},$$

对于某个正定核 $\kappa(\cdot;\cdot)$，其中 $X_{i,t}$ 是状态-动作对（$S_{i,t}$,$A_{i,t}$）的简写。类似于 Liu 等人（2018）的定理 2，我们可以证明，式（28）中的优化问题随后可化简为

$$\operatorname{argmin}_{\omega\in\Omega}\sum_{(i_1,t_1)\neq(i_1',t_1')}\sum_{(i_2,t_2)\neq(i_2',t_2')}D(\omega,\pi_{\text{old}};i_1,t_1,i_1',t_1',i_2,t_2,i_2',t_2'),$$

其中，$D(\omega,\pi;\ i_1,t_1,i_1',t_1',i_2,t_2,i_2',t_2')$由下式给出

$$\begin{aligned}
&\frac{\omega(X_{i_1',t_1'};X_{i_1,t_1})}{(1-\gamma)^{-1}}\Big\{\gamma\mathbb{E}_{a\sim\pi(\bullet|S_{i_1',t_1'+1})}\kappa(S_{i_1',t_1'+1},a,X_{i_1,t_1};X_{i_2,t_2},X_{i_2,t_2})-\kappa(X_{i_1',t_1'},X_{i_1,t_1};X_{i_2,t_2},X_{i_2,t_2})\Big\}\\
&+\frac{\omega(X_{i_2',t_2'};X_{i_2,t_2})}{(1-\gamma)^{-1}}\Big\{\gamma\mathbb{E}_{a\sim\pi(\bullet|S_{i_2',t_2'+1})}\kappa(S_{i_2',t_2'+1},a,X_{i_2,t_2};X_{i_1,t_1},X_{i_1,t_1})-\kappa(X_{i_2',t_2'},X_{i_2,t_2};X_{i_1,t_1},X_{i_1,t_1})\Big\}\\
&+\omega(X_{i_1',t_1'};X_{i_1,t_1})\omega(X_{i_2',t_2'};X_{i_2,t_2})\Big\{\gamma^2\mathbb{E}_{\substack{a_1\sim\pi(\bullet|S_{i_1',t_1'+1})\\a_2\sim\pi(\bullet|S_{i_2',t_2'+1})}}\kappa(S_{i_2',t_2'+1},a,X_{i_2,t_2};S_{i_1',t_1'+1},a_1,X_{i_1,t_1})\\
&-\gamma\mathbb{E}_{a_1\sim\pi(\bullet|S_{i_1',t_1'+1})}\kappa(S_{i_1',t_1'+1},a,X_{i_1,t_1};X_{i_2',t_2'},X_{i_2,t_2})-\gamma\mathbb{E}_{a_2\sim\pi(\bullet|S_{i_2',t_2'+1})}\kappa(S_{i_2',t_2'+1},a,X_{i_2,t_2};X_{i_1',t_1'},X_{i_1,t_1})\\
&+\kappa(X_{i_2',t_2'},X_{i_2,t_2};X_{i_1',t_1'},X_{i_1,t_1})\Big\}+(1-\gamma)^2\kappa(X_{i_1,t_1},X_{i_1,t_1};X_{i_2,t_2},X_{i_2,t_2}).
\end{aligned}$$

算法4 密度比的估计。

输入： $\mathcal{I}_\ell$中的数据子集。

初始：I nitial 将密度比 $\omega = \omega_\beta$ 设为由 β参数化的神经网络 .

对于迭代 $= 1 \quad , 2, \cdots$执行

a 从数据转移中随机采样批次 $\mathcal{M}, \mathcal{M}^*$ 。b 通过以下方式更新参数 β

$$\beta \leftarrow \beta - \epsilon \binom{|\mathcal{M}|}{2}^{-2} \sum_{\substack{(i_1,t_1),(i_1',t_1')\in\mathcal{M} \\ (i_1,t_1)\neq(i_1',t_1')}} \sum_{\substack{(i_2,t_2),(i_2',t_2')\in\mathcal{M} \\ (i_2,t_2)\neq(i_2',t_2')}} \nabla_\beta D(\frac{\omega_\beta}{z_{\omega_\beta}}, \pi_{\text{old}}; i_1, t_1, i_1', t_1', i_2, t_2, i_2', t_2'),$$

其中， z_{ω_β} 是一个归一化常数

$$z_{\omega_\beta}(\cdot; S_{i,t}, A_{i,t}) = \frac{1}{|\mathcal{M}^*|} \sum_{(i',t')\in\mathcal{M}^*} \omega_\beta(X_{i',t'}; X_{i,t}).$$

输出 $\omega_{\beta\circ}$

在我们的实现中，我们将Ω设为神经网络类。详细的估计程序见算法4。

B.2 关于条件折扣访问概率的更多内容

我们首先针对给定的一对(a, s)，给出 $\mathcal{D}_{\text{TV}}(\widehat{d}^{(\ell)}(\bullet|a, s)\ d^{\pi^{(\ell)}}{}_{\text{old}}(\bullet|a, s))$的一个上界。注意，全变差距离对应于 f-散度的一个特殊情形。因此，我们可以将 $\mathcal{D}_{\text{TV}}(\widehat{d}^{(\ell)}(\bullet|a, s)\ d^{\pi^{(\ell)}}{}_{\text{old}}(\bullet|a, s))$表示为

$$\sup_{\|f\|_\infty \le 1/2} |\mathbb{E}_{S^*\sim d^{\pi^{(\ell)}_{\text{old}}}(\bullet|a,s)} f(S^*) - \mathbb{E}_{S^*\sim\widehat{d}^{(\ell)}(\bullet|a,s)} f(S^*)|,$$

或者等价地，

$$\sup_{\|f\|_\infty \le 1/2} \left|\mathbb{E}_{S^*\sim d^{\pi^{(\ell)}_{\text{old}}}(\bullet|a,s)} f(S^*) - \frac{1-\gamma}{M} \sum_{m=1}^{M} \sum_{t'=0}^{T'} \gamma^t f(\widetilde{S}_{t'}^{(m)})\right|.$$

根据定义，对于给定的 f，我们可以将差值分解为

$$\left|\mathbb{E}_{S^*\sim d^{\pi^{(\ell)}_{\text{old}}}(\bullet|a,s)} f(S^*) - \frac{1-\gamma}{M}\sum_{m=1}^{M}\sum_{t'=0}^{T'}\gamma^t f(\widetilde{S}_{t'}^{(m)})\right|$$

$$\leq \sum_{t'=0}^{T'}\gamma^{t'}\left|(1-\gamma)\mathbb{E}_{S^*\sim p_{t'}^{\pi^{(\ell)}_{\text{old}}}(\bullet|a,s)} f(S^*) - \frac{1-\gamma}{M}\sum_{m=1}^{M} f(\widetilde{S}_{t'}^{(m)})\right|$$

$$+(1-\gamma)\sum_{t'=T'+1}^{+\infty}\gamma^{t'}|\mathbb{E}_{S^*\sim d^{\pi^{(\ell)}_{\text{old}}}(\bullet|a,s)} f(S^*)|.$$

由于 f 被 $1/2$ 一致有界，右端的第二项被 $\gamma^{T'}$ 所界定，并且当 $T'\to\infty$ 时收敛于零。

至于第一项，可以进一步将其上界界定为

$$\sum_{t'=0}^{T'}(1-\gamma)\gamma^{t'}\left|\mathbb{E}_{S^*\sim p_{t'}^{\pi^{(\ell)}_{\text{old}}}(\bullet|a,s)} f(S^*) - \mathbb{E}f(\widetilde{S}_{t'})\right| + \sum_{t'=0}^{T'}(1-\gamma)\gamma^{t'}\left|\frac{1}{M}\sum_{m=1}^{M} f(\widetilde{S}_{t'}^{(m)}) - \mathbb{E}f(\widetilde{S}_{t'})\right|. \quad (29)$$

(29) 中第二项的期望可以上界界定为

$$\sum_{t'=0}^{T'}(1-\gamma)\gamma^{t'}\sqrt{\text{Var}\left(\frac{1}{M}\sum_{m=1}^{M} f(\widetilde{S}_{t'}^{(m)}) - \mathbb{E}f(\widetilde{S}_{t'})\right)} \leq M^{-1/2},$$

根据柯西-施瓦茨不等式。因此，(29) 中的第二项会随着$M\to\infty$趋于零。

最后，考虑 (29) 中的第一项。我们用 $(S_t^{(j)}, A_t^{(j)})$ 表示在时间 t 测得的状态-动作对：它在前 j 步中以 $(A_0, S_0) = (a, s)$ 为条件，随后遵循 $\pi^{(\ell)}$ 旧和转移函数 p；之后的步骤中则遵循 $\pi^{(\ell)}$ 旧和转移函数$\mathcal{N}(\widehat{\mu}^{(\ell)}, \widehat{\Sigma}^{(\ell)})$。对于每个 t'，我们有

$$\left|\mathbb{E}_{S^*\sim p_{t'}^{\pi^{(\ell)}_{\text{old}}}(\bullet|a,s)} f(S^*) - \mathbb{E}f(\widetilde{S}_{t'})\right| \leq \sum_{j=1}^{t'}\left|\mathbb{E}f(S_{t'}^{(j-1)}) - \mathbb{E}f(S_{t'}^{(j)})\right|.$$

令 $\widehat{p}_t^{\pi^{(\ell)}}{}_{\text{old}}(\bullet|a,s)$ 表示在时间 t 的状态向量 S_t 的分布函数，该状态向量遵循 π_{old}，并以 $(A_0 = a, S_0 = s)$ 为条件的估计转移函数为 $\mathcal{N}(\widehat{\mu}^{(\ell)}, \widehat{\Sigma}^{(\ell)})$。当 $t = 1$ 时，我们省略下标 t 和上标 $\pi^{(\ell)}$ old。假设 f 被某个常数 $c > 0$ 一致有界。那么，对于任意 s，$|\mathbb{E}_{S^*\sim\widehat{p}^{\pi^{(\ell)}}\text{ old }t'-j(\bullet|a^*,s)\, a^*\sim\pi^{(\ell)}\text{ old}(\bullet|s)} f(S^*)|$ 也被 c 一致有界。因此，

$$\left|\mathbb{E}f(S_{t'}^{(j-1)}) - \mathbb{E}f(S_{t'}^{(j)})\right| \leq \mathbb{E}\left|\mathbb{E}_{S^*\sim p(\bullet|A_{j-1}^{(j-1)},,S_{j-1}^{(j-1)})}\left\{\mathbb{E}_{\substack{S^{**}\sim \widehat{p}_{t'-j}^{\pi_{\text{old}}^{(\ell)}}(\bullet|a^{**},S^*)\\ a^{**}\sim\pi_{\text{old}}^{(\ell)}(\bullet|S^*)}} f(S^{**})\right\}\right.$$
$$\left. - \mathbb{E}_{S^*\sim \widehat{p}(\bullet|A_{j-1}^{(j-1)},S_{j-1}^{(j-1)})}\left\{\mathbb{E}_{\substack{S^{**}\sim p_{t'-j}^{\pi_{\text{old}}^{(\ell)}}(\bullet|a^{**},S^*)\\ a^{**}\sim\pi_{\text{old}}^{(\ell)}(\bullet|S^*)}} f(S^{**})\right\}\right|$$
$$\leq c\mathbb{E}\mathcal{D}_{\text{TV}}\{p(\bullet|A_{j-1}^{(j-1)}, S_{j-1}^{(j-1)}), \widehat{p}(\bullet|A_{j-1}^{(j-1)}, S_{j-1}^{(j-1)})\}.$$

由此可知，式（29）中的第一项可由以下式子给出上界

$$\begin{aligned}
&\sum_{t'=0}^{+\infty}(1-\gamma)\gamma^{t'}\left|\mathbb{E}_{S^*\sim p_{t'}^{\pi_{\text{old}}^{(\ell)}}(\bullet|a,s)} f(S^*) - \mathbb{E}f(\widetilde{S}_{t'})\right| \\
\leq\ & c\sum_{t'=1}^{+\infty}(1-\gamma)\gamma^{t'}\sum_{j=1}^{t'}\mathbb{E}\mathcal{D}_{\text{TV}}\{p(\bullet|A_{j-1}^{(j-1)}, S_{j-1}^{(j-1)}), \widehat{p}(\bullet|A_{j-1}^{(j-1)}, S_{j-1}^{(j-1)})\} \\
=\ & c\sum_{j=1}^{+\infty}\gamma^{\ell}\mathcal{D}_{\text{TV}}\{p(\bullet|A_{j-1}^{(j-1)}, S_{j-1}^{(j-1)}), \widehat{p}(\bullet|A_{j-1}^{(j-1)}, S_{j-1}^{(j-1)})\} \\
=\ & \frac{c\gamma}{1-\gamma}\mathbb{E}_{(A^*,S^*)\sim q^{\pi_{\text{old}}^{(\ell)}}(\bullet,\bullet;a,s)}\mathcal{D}_{\text{TV}}\{p(\bullet|A^*, S^*), \widehat{p}(\bullet|A^*, S^*)\},
\end{aligned} \tag{30}$$

其中，$q^{\pi^{(\ell)}_{\text{旧}}}(\bullet,\bullet;\, a, s)$ 表示状态-动作对的条件折扣访问概率，即

$$q^{\pi_{\text{old}}^{(\ell)}}(a', s'; a, s) = (1-\gamma)\mathbb{I}(a'=a, s'=s) + (1-\gamma)\sum_{t=1}^{+\infty}\gamma^t\pi_{\text{old}}^{(\ell)}(a'|s')p_t^{\pi_{\text{old}}^{(\ell)}}(s'|a,s).$$

令 $\widetilde{p}$ 表示均值为 $\widehat{\mu}^{(\ell)}$ 、协方差矩阵为 Σ 的正态密度函数。根据三角不等式可得

$$\mathcal{D}_{\text{TV}}(p, \widehat{p}) \leq \mathcal{D}_{\text{TV}}(p, \widetilde{p}) + \mathcal{D}_{\text{TV}}(\widetilde{p}, \widehat{p}).$$

根据 Devroye 等（2018）的命题 2.1，在最小特征值远离零的条件下，p 与 $\widetilde{p}$ 之间的全变差距离以上界 $0.5\sqrt{(\mu-\widehat{\mu}^{(\ell)})^\top\Sigma^{-1}(\mu-\widehat{\mu}^{(\ell)})} \leq c\|\mu-\widehat{\mu}^{(\ell)}\|_2$ 为界，其中$c>0$ 是某个常数。与此同时，根据 Devroye 等（2018）的定理 1.1，在 $\widehat{\Sigma}^{(\ell)}$ 和 Σ 均为正定矩阵的条件下，$\widehat{p}$ 与 $\widetilde{p}$ 之间的全变差距离以上界 $1.5\sqrt{\sum_i\lambda_i^2}$ 为界，其中 $\{\lambda_i\}_i$ 表示 $\Sigma^{-1}(\widehat{\Sigma}^{(\ell)}-\Sigma)$ 的特征值。特征值平方之和等于 $\Sigma^{-1}(\widehat{\Sigma}^{(\ell)}-\Sigma)$的弗罗贝尼乌斯范数的平方，而该平方

还可以进一步以上界 $\|\Sigma^{-1}\|_2^2\|\widehat{\Sigma}^{(\ell)}-\Sigma\|_F^2 \le c^2\|\widehat{\Sigma}^{(\ell)}-\Sigma\|_F^2$为界。因此，

$$\mathcal{D}_{\mathrm{TV}}(p,\widehat{p}) \le \frac{c}{2}\|\mu-\widehat{\mu}^{(\ell)}\|_2 + \frac{3c}{2}\|\Sigma-\widehat{\Sigma}^{(\ell)}\|_F.$$

综上所述，我们证明了

$$\begin{aligned}&\mathcal{D}_{\mathrm{TV}}(\widehat{d}^{(\ell)}(\bullet|a,s), d^{\pi_{\mathrm{old}}^{(\ell)}}(a,s)) \le \gamma^{T'} + M^{-1/2}\\ +\quad & O(1)\mathbb{E}_{(A^*,S^*)\sim q^{\pi_{\mathrm{old}}^{(\ell)}}(\bullet,\bullet;a,s)}[\|\mu(S^*,A^*)-\widehat{\mu}^{(\ell)}(S^*,A^*)\|_2 + \|\Sigma(S^*,A^*)-\widehat{\Sigma}^{(\ell)}(S^*,A^*)\|_F],\end{aligned}$$

对于某个正常数 $O(1)$。

根据柯西-施瓦茨不等式，聚合平方总变差距离的上界为

$$O(1)\mathbb{E}_{(a,s)\sim p_\infty}\{\mathbb{E}_{(A^*,S^*)\sim q^{\pi_{\mathrm{old}}^{(\ell)}}(\bullet,\bullet;a,s)}[\|\mu(S^*,A^*)-\widehat{\mu}^{(\ell)}(S^*,A^*)\|_2 + \|\Sigma(S^*,A^*)-\widehat{\Sigma}^{(\ell)}(S^*,A^*)\|_F]\}^2 + 3\gamma^{2T'} + \frac{3}{M},$$

对于某个正常数 $O(1)$。再次使用柯西-施瓦茨不等式，并注意到 p_∞ 有一个正的下界，第一项的上界还可进一步写为

$$\begin{aligned}&O(1)\mathbb{E}_{(a,s)\sim p_\infty}\mathbb{E}_{(A^*,S^*)\sim q^{\pi_{\mathrm{old}}^{(\ell)}}(\bullet,\bullet;a,s)}[\|\mu(S^*,A^*)-\widehat{\mu}^{(\ell)}(S^*,A^*)\|_2 + \|\Sigma(S^*,A^*)-\widehat{\Sigma}^{(\ell)}(S^*,A^*)\|_F]^2\\ &\le O(1)\mathbb{E}_{(A^*,S^*)\sim p_\infty}[\|\mu(S^*,A^*)-\widehat{\mu}^{(\ell)}(S^*,A^*)\|_2 + \|\Sigma(S^*,A^*)-\widehat{\Sigma}^{(\ell)}(S^*,A^*)\|_F]^2.\end{aligned}$$

至此，证明完成。

C 证明

在本节中，我们用 C 和 c 表示某些通用常数，其取值可以随上下文而变化。

C.1 定理 1 的证明

定理 1 的证明很直接。首先注意到，由于式 (20) 中的信赖域条件，我们有

$$(1-\gamma)\frac{1}{\mathbb{L}}\sum_{\ell=1}^{\mathbb{L}}\mathbb{E}_{S^*\sim\nu}\mathcal{D}_{\mathrm{KL}}(\pi_{\mathrm{old}}^{(\ell)}(\bullet|S^*), \pi_{\mathrm{new}}(\bullet|S^*)) \le \delta, \tag{31}$$

对于任意 s，如 $\widehat{d}^{(\ell)}, \nu(s) \ge (1-\gamma)\nu(s)$。

接下来，使用引理 1 证明中的类似论证，我们可以证明

$$|\mathcal{V}(\pi_{\text{new}}) - \mathcal{V}(\pi_{\text{old}})| = \left|\mathbb{E}_{S^* \sim d^{\pi_{\text{new}}}} \sum_a \{\pi_{\text{new}}(a|S^*) - \pi_{\text{old}}(a|S^*)\} A^{\pi_{\text{old}}}(a, S^*)\right|$$

$$\leq O(1)\mathbb{E}_{S^* \sim d^{\pi_{\text{new}}}} \|\pi_{\text{new}}(\bullet|S^*) - \pi_{\text{old}}(\bullet|S^*)\|_{\text{TV}} \leq O(1)\sqrt{\mathbb{E}_{S^* \sim d^{\pi_{\text{new}}}} \mathcal{D}_{\text{KL}}(\pi_{\text{old}}(\bullet|S^*) \ \pi_{\text{new}}(\bullet|S^*))},$$

其中， $O(1)$表示某个正常数。第一个等式由式 (5) 得出，第一个不等式源于即时奖励一致有界（因此优势函数也有界）这一条件，第二个不等式则由 Pinsker不等式 得出。在 $\nu(\cdot)$ 一致地远离零的条件下，我们得到

$$\mathbb{E}_{S^* \sim d^{\pi_{\text{new}}}} \mathcal{D}_{\text{KL}}(\pi_{\text{old}}(\bullet|S^*), \pi_{\text{new}}(\bullet|S^*)) \leq C\mathbb{E}_{S^* \sim \nu} \mathcal{D}_{\text{KL}}(\pi_{\text{old}}(\bullet|S^*), \pi_{\text{new}}(\bullet|S^*)),$$

对于某个常数 $C > 0$。由此可得

$$\left|\mathcal{V}(\pi_{\text{new}}) - \frac{1}{\mathbb{L}} \sum_{\ell=1}^{\mathbb{L}} \mathcal{V}(\pi_{\text{old}}^{(\ell)})\right| \leq \frac{1}{\mathbb{L}} \sum_{\ell=1}^{\mathbb{L}} |\mathcal{V}(\pi_{\text{new}}) - \mathcal{V}(\pi_{\text{old}}^{(\ell)})|$$

$$\leq c \sum_{\ell=1}^{\mathbb{L}} \sqrt{\mathbb{E}_{S^* \sim \nu} \mathcal{D}_{\text{KL}}(\pi_{\text{old}}^{(\ell)}(\bullet|S^*), \pi_{\text{new}}(\bullet|S^*))} \leq c\mathbb{L} \sqrt{\sum_{\ell=1}^{\mathbb{L}} \mathbb{E}_{S^* \sim \nu} \mathcal{D}_{\text{KL}}(\pi_{\text{old}}^{(\ell)}(\bullet|S^*), \pi_{\text{new}}(\bullet|S^*))},$$

对于某个常数 $c > 0$。由式 (31) 即可得出定理 1。

C.2 定理2的证明

我们先介绍一些辅助引理。

引理 2 在 (C5) 条件下，存在某个常数 $c_0 > 0$ ，使得对于任意 π,

$$\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi) \geq c_0 \{\mathbb{E}_{S^* \sim d^{\pi^{opt}, \nu}} \|\pi^{opt}(\bullet|S^*) - \pi(\bullet|S^*)\|_{\text{TV}}\}^{1+1/\alpha},$$

其中 α 是 (C5) 中的指数。

引理 3 令 $\widehat{\eta}_1^*(\pi, \pi^{(\ell)\text{旧}}) = (|\mathcal{I}_\ell|T)^{-1} \sum_{i \in \mathcal{I}_\ell} \sum_{t<T} \psi(O_{i,\ t};\ \pi, \pi^{(\ell)\text{旧}} Q^{\pi^{(\ell)\text{旧}}}\ \omega^{\pi^{(\ell)\text{旧}}}\ d^{\pi^{(\ell)\text{旧}}})$ 和 $\widehat{\eta}_1^{(\ell)}(\pi) = (|\mathcal{I}_\ell|T)^{-1} \sum_{i \in \mathcal{I}_\ell} \sum_{t<T} \psi(O_{i,\ t};\ \pi, \pi^{(\ell)\text{旧}}\ \widehat{Q}^{(\ell)}, \widehat{\omega}^{(\ell)}, \widehat{d}^{(\ell)})$。在给定条件下，

$$\sup_{\pi_1, \pi_2 \in \Pi} \frac{|\widehat{\eta}_1^{(\ell)}(\pi_1) - \widehat{\eta}_1^*(\pi_1, \pi_{\text{old}}^{(\ell)}) - \widehat{\eta}_1^{(\ell)}(\pi_2) + \widehat{\eta}_1^*(\pi_2, \pi_{\text{old}}^{(\ell)})|}{\mathbb{E}_{S^* \sim d^{\pi^{opt}, \nu}} \|\pi_1(\bullet|S^*) - \pi_2(\bullet|S^*)\|_{\text{TV}}} = o_p(\{NT\}^{-1/(2+2\alpha)}) + O_p(\{(NT)\}^{\kappa_4/2-1/2}).$$

引理 4 设 $\{Z_t : t \geq 0\}$ 是平稳的 β-混合过程，其 β-混合系数 $\{\beta(q) : \ q \geq 0\}$。设 $\mathcal{F}$ 是一个逐点可测函数类，其输入为 Z_t 。对于任意 $f \in \mathcal{F}$，假设 $\mathbb{E}\{f(Z_0)\} = 0$。设 $\sigma^2 > 0$ 为正数,

使得 $\sup_{f\in\mathcal{F}}\mathbb{E}\{f^2(Z_0)\}\le\sigma^2$。假设包络函数被某个常数 $C>0$ 一致有界。此外，假设 $\mathcal{F}$ 属于 VC 型函数类，使得 $\sup_Q N(\mathcal{F},e_Q,\varepsilon\|F\|_{Q,\,2})\le(A/\varepsilon)^v$，其中 $A\ge e,v\ge 1$。那么

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1}f(Z_t)\right|=O_p\left\{\sqrt{vq\sigma^2T\log\left(\frac{AC}{\sigma}\right)}+vC\log\left(\frac{AC}{\sigma}\right)+q\right\},$$

对于任意满足 $T\beta(q)/q=o(1)$ 的 $1\le q<T/2$。

接下来，我们概述证明思路。我们的目标是为价值差 $\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{new}})$ 提供一个上界。它可以表示为

$$\frac{1}{(1-\gamma)\mathbb{L}}\sum_{\ell=1}^{\mathbb{L}}\left\{\eta_1(\pi^{opt},\pi_{\text{old}}^{(\ell)})+\eta_2(\pi^{opt},\pi_{\text{old}}^{(\ell)})-\eta_1(\pi_{\text{new}},\pi_{\text{old}}^{(\ell)})-\eta_2(\pi_{\text{new}},\pi_{\text{old}}^{(\ell)})\right\}.$$

我们将证明的其余部分分为几个步骤。在第一步中，我们为高阶余项 $\eta_2(\pi^{opt},\pi^{(\ell)}{}_{\text{old}})$ 和 $\eta_2(\pi_{\text{new}},\pi^{(\ell)}{}_{\text{old}})$ 提供上界。

我们首先证明，当 δ 被设定为一个足够小的常数时，高阶余项 $|\eta_2(\pi^{opt},\pi^{(\ell)}{}_{\text{old}})|+|\eta_2(\pi_{\text{new}},\pi^{(\ell)}{}_{\text{old}})|$ 可以被上界 $\epsilon\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{new}})\}+c\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi^{(\ell)}{}_{\text{old}})\}^{(\alpha+2)/(\alpha+1)}$ 所上界，其中 $c>0$、$0<\epsilon\le(1-\gamma)/2$ 为某些常数，并且该结论以趋近于 1 的概率成立（w.p.a.1）。因此

$$\begin{aligned}\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{new}})-\frac{\epsilon}{1-\gamma}\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{new}})\}\le{}&\frac{c}{\mathbb{L}}\sum_{\ell=1}^{\mathbb{L}}\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{old}}^{(\ell)})\}^{(\alpha+2)/(\alpha+1)}\\&+\frac{1}{(1-\gamma)\mathbb{L}}\sum_{\ell=1}^{\mathbb{L}}\{\eta_1(\pi^{opt},\pi_{\text{old}}^{(\ell)})-\eta_1(\pi_{\text{new}},\pi_{\text{old}}^{(\ell)})\}.\end{aligned}$$

在给定的初始策略条件下，我们有

$$\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{new}})\le O_p\{(NT)^{-\frac{\alpha+2}{\alpha+1}\kappa_0}\}+\frac{2}{(1-\gamma)\mathbb{L}}\sum_{\ell=1}^{\mathbb{L}}\{\eta_1(\pi^{opt},\pi_{\text{old}}^{(\ell)})-\eta_1(\pi_{\text{new}},\pi_{\text{old}}^{(\ell)})\},$$

以概率趋近于 1。

只需对 $\mathbb{L}^{-1}\sum_{\ell=1}^{\mathbb{L}}\{\eta_1(\pi^{opt},\pi^{(\ell)}{}_{\text{旧}})-\eta_1(\pi_{\text{new}},\pi^{(\ell)}{}_{\text{旧}})\}$ 给出上界。注意到 π_{new} 是通过最大化 $\widehat{\eta}_1(\pi)$ 得到的，因此有 $\widehat{\eta}_1(\pi_{\text{new}})\ge\widehat{\eta}_1(\pi^{opt})$。由此，$\mathbb{L}^{-1}\sum_{\ell=1}^{\mathbb{L}}\{\eta_1(\pi^{opt},\pi_{\text{old}})-\eta_1(\pi_{\text{new}},\pi_{\text{old}})\}\le\mathbb{L}^{-1}\sum_{\ell=1}^{\mathbb{L}}\{\eta_1(\pi^{opt},\pi_{\text{old}})-\eta_1(\pi_{\text{new}},\pi_{\text{old}})-\widehat{\eta}_1(\pi_{\text{new}})+\widehat{\eta}_1(\pi^{opt})\}$。因此，只需为其给出上界

$$\left|\frac{1}{\mathbb{L}}\sum_{\ell=1}^{\mathbb{L}}\{\eta_1(\pi^{opt},\pi_{\text{old}}^{(\ell)})-\eta_1(\pi_{\text{new}},\pi_{\text{old}}^{(\ell)})-\widehat{\eta}_1(\pi_{\text{new}})+\widehat{\eta}_1(\pi^{opt})\}\right|.$$

根据引理 3，上述量的上界为

$$\left|\frac{1}{\mathbb{L}}\sum_{\ell=1}^{\mathbb{L}}\{\eta_1(\pi^{opt},\pi_{\text{old}}^{(\ell)})-\eta_1(\pi_{\text{new}},\pi_{\text{old}}^{(\ell)})-\widehat{\eta}_1^*(\pi_{\text{new}})+\widehat{\eta}_1^*(\pi^{opt})\}\right|$$
$$+\quad [o_p\{(NT)^{-1/(2+2\alpha)}\}+O_p\{(NT)\}^{\kappa_4/2-1/2}]\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\|\pi^{opt}(\bullet|S^*)-\pi_{\text{new}}(\bullet|S^*)\|_{\text{TV}}.$$

在第二步中，我们证明第一项可以被上界为

$$\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\mathcal{D}_{\text{TV}}\{\pi_{\text{new}}(\bullet|S^*),\pi^{opt}(\bullet|S^*)\}O_p\{(NT)^{\kappa_4/2-1/2}\log(NT)\}.$$

根据引理 2，我们得到

$$\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{new}})\le O_p\{(NT)^{-\frac{\alpha+2}{\alpha+1}\kappa_0}\}+\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{new}})\}^{\alpha/(1+\alpha)}(I_1+I_2),$$

其中 $I_1=O_p\{(NT)^{\kappa_4/2-1/2}\log(NT)\}$和 $I_2=o_p\{(NT)^{-1/(2+2\alpha)}\}$。利用 H¨older 不等式，右侧的第二项可以被上界为 $\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{new}})\}/2+I_1^{(1+\alpha)}+I_2^{(1+\alpha)}$。由此，在给定的关于 κ_4 的条件下，可得 $\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{new}})=O\{(NT)^{-\kappa_0\frac{\alpha+2}{\alpha+1}}\}+o_p\{(NT)^{-1/2}\}$。至此，证明完成。

在最后三个步骤中，我们给出引理 2、3 和 4 的证明。接下来，我们分别详细介绍每个步骤。

步骤 1. 我们旨在界定高阶余项 $|\eta_2(\pi^{opt},\pi^{(\ell)}{}_{\text{old}})|$ 和 $|\eta_2(\pi_{\text{new}},\pi^{(\ell)}{}_{\text{old}})|$。我们首先考虑 $|\eta_2(\pi^{opt},\pi^{(\ell)}{}_{\text{old}})|$。根据定义，对于任意 π，有 $\eta_2(\pi^{opt},\pi)=\eta_2^{(1)}(\pi^{opt},\pi)+\eta_2^{(2)}(\pi^{opt},\pi)$，其中

$$\begin{aligned}\eta_2^{(1)}(\pi^{opt},\pi)&=\sum_{a\in\mathcal{A},s\in\mathcal{S}}\{\pi^{opt}(a|s)-\pi(a|s)\}A^{\pi^{opt}}(a,s)\{d^{\pi^{opt},\nu}(s)-d^{\pi,\nu}(s)\},\\ \eta_2^{(2)}(\pi^{opt},\pi)&=\sum_{a\in\mathcal{A},s\in\mathcal{S}}\{\pi^{opt}(a|s)-\pi(a|s)\}\{A^{\pi}(a,s)-A^{\pi^{opt}}(a,s)\}\{d^{\pi^{opt},\nu}(s)-d^{\pi,\nu}(s)\}.\end{aligned}$$

因此，只需对 $|\eta_2^{(j)}(\pi^{opt},\pi^{(\ell)}{}_{\text{old}})|$ 进行界定，其中 $j=1$ 为 2。

我们首先考虑 $|\eta_2^{(1)}(\pi^{opt},\pi^{(\ell)}{}_{\text{old}})|$。利用引理 1 的证明中类似的论证（见公式 24），我们可以证明其上界为

$$O(1)\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\|\pi^{opt}(\bullet|S^*)-\pi_{\text{old}}^{(\ell)}(\bullet|S^*)\|_{\text{TV}}\mathbb{E}_{S^*\sim d^{\pi_{\text{old}}^{(\ell)},\nu}}\sum_{a\in\mathcal{A}}|\pi_{\text{old}}^{(\ell)}(a|S^*)-\pi^{opt}(a|S^*)||A^{\pi^{opt}}(a,s)|,$$

其中，$O(1)$ 表示某个正常数。在引理 2 的证明中，我们证明了对于任意 a、s 和 π，$\{\pi(a|s)-\pi^{opt}(a|s)\}A^{\pi^{opt}}(a,s)$ 非正。因此，对于任意 π，$\mathbb{E}_{S^*\sim d^{\pi,\nu}}\sum_{a\in\mathcal{A}}|\pi(a|S^*)-\pi^{opt}(a|S^*)||A^{\pi^{opt}}(a,S^*)|=\mathbb{E}_{S^*\sim d^{\pi,\nu_{\text{old}}}}\sum_{a\in\mathcal{A}}\{\pi^{opt}(a|S^*)-\pi(a|S^*)\}A^{\pi^{opt}}(a,S^*)=\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi)$。由引理 2 可知，对于任意 π，

$$|\eta_2^{(1)}(\pi^{opt},\pi_{\text{old}}^{(\ell)})| \le c\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{old}}^{(\ell)})\}\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\|\pi^{opt}(\bullet|S^*)-\pi_{\text{old}}^{(\ell)}(\bullet|S^*)\|_{\text{TV}} \le c_1\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{old}}^{(\ell)})\}^{\frac{2\alpha+1}{\alpha+1}}, \tag{32}$$

其中 $c_1>0$ 为某个常数。

接下来考虑 $|\eta_2^{(2)}(\pi^{opt},\pi^{(\ell)\text{旧}})|$。注意，对于任意 π，$A^{\pi}(a,s)-A^{\pi^{opt}}(a,s)=\gamma\mathbb{E}_{S^*\sim p(\bullet|a,s)}\{V^{\pi}(S^*)-V^{\pi^{opt}}(S^*)\}+V^{\pi^{opt}}(s)-V^{\pi}(s)$。在给定条件下，存在某个常数 $C>0$，使得对于任意 s，都有 $\nu(s)\ge C$。利用测度变换可得

$$\mathbb{E}_{S^*\sim p(\bullet|a,s)}\{V^{\pi^{opt}}(S^*)-V^{\pi}(S^*)\} \le \mathbb{E}_{S^*\sim\nu}\{V^{\pi^{opt}}(S^*)-V^{\pi}(S^*)\}\frac{p(S^*|a,s)}{\nu(S^*)} \le C^{-1}\mathbb{E}_{S^*\sim\nu}\{V^{\pi^{opt}}(S^*)-V^{\pi}(S^*)\}=C^{-1}\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi)\}, \tag{33}$$

和 $V^{\pi^{opt}}(s)-V^{\pi}(s)\le C^{-1}\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi)\}$。

使用与式 (24) 类似的论证，可得

$$|\eta_2^{(2)}(\pi^{opt},\pi)| \le O(1)\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\sum_{a\in\mathcal{A}}|\pi(a|S^*)-\pi^{opt}(a|S^*)|\max_{a,s}|A^{\pi^{opt}}(a,s)-A^{\pi}(a,s)| \le O(1)\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi)\}\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\|\pi(\bullet|S^*)-\pi^{opt}(\bullet|S^*)\|_{\text{TV}},$$

其中 $O(1)$ 表示某个正的常数。类似于式 (32)，可得 $|\eta_2^{(2)}(\pi^{opt},\pi^{(\ell)\text{旧}})|\le c_2\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi^{(\ell)\text{旧}})\}^{(2\alpha+1)/(\alpha+1)}$，其中 $c_2>0$ 为某个常数。结合式 (32)，得到

$$|\eta_2(\pi^{opt},\pi_{\text{old}}^{(\ell)})| \le (c_1+c_2)\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi_{\text{old}}^{(\ell)})\}^{\frac{2\alpha+1}{\alpha+1}}.$$

接下来我们对界 $|\eta_2(\pi_{\text{new}},\pi^{(\ell)}{}_{\text{old}})|$。我们注意到，$|\eta_2(\pi_{\text{new}},\pi^{(\ell)}{}_{\text{old}})|$ 可以被上界限制为 y

$$|\eta_2(\pi_{\text{new}},\pi_{\text{old}}^{(\ell)})| \le |\eta_2^{(3)}(\pi_{\text{new}},\pi_{\text{old}}^{(\ell)})|+|\eta_2^{(4)}(\pi_{\text{new}},\pi_{\text{old}}^{(\ell)})| \equiv \sum_{a,s}|\pi_{\text{new}}(a|s)-\pi_{\text{old}}^{(\ell)}(a|s)||A^{\pi^{opt}}(a,s)-A^{\pi_{\text{old}}^{(\ell)}}(a,s)||d^{\pi_{\text{new}},\nu}(s)-d^{\pi_{\text{old}}^{(\ell)},\nu}(s)| +\sum_{a,s}|\pi_{\text{new}}(a|s)-\pi_{\text{old}}^{(\ell)}(a|s)||A^{\pi^{opt}}(a,s)||d^{\pi_{\text{new}},\nu}(s)-d^{\pi_{\text{old}}^{(\ell)},\nu}(s)|.$$

采用对 $|\eta_2^{(2)}(\pi^{opt},\pi^{(\ell)}{}_{\text{old}})|$ 进行界定时的类似论证，$|\eta_2^{(3)}(\pi_{\text{new}},\pi^{(\ell)}{}_{\text{old}})|$ 可以被 $O(1)\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi^{(\ell)}{}_{\text{old}})\}\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\|\pi_{\text{new}}(a|S^*)-\pi^{(\ell)}{}_{\text{old}}(a|S^*)\|_{\text{TV}}$ 上界限制，其中 $O(1)$ 表示某个正数常数。根据三角不等式，我们有

$$\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\|\pi_{\text{new}}(a|S^*)-\pi_{\text{old}}^{(\ell)}(a|S^*)\|_{\text{TV}} \le \mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\|\pi^{opt}(a|S^*)-\pi_{\text{old}}^{(\ell)}(a|S^*)\|_{\text{TV}} +\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\|\pi^{opt}(a|S^*)-\pi_{\text{new}}(a|S^*)\|_{\text{TV}}.$$

根据引理 2，右端的两项分别可以由 $c_0^{-1}\{\mathcal{V}(\pi^{opt})-\mathcal{V}(\pi^{(\ell)\text{旧}})\}^{\alpha/(1+\alpha)}$

和 $c_0^{-1}\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{new}})\}^{\alpha/(1+\alpha)}$ 上界限定。因此，

$$|\eta_2^{(3)}(\pi_{\text{new}}, \pi_{\text{old}}^{(\ell)})| \leq O(1)\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{old}}^{(\ell)})\}^{\frac{2\alpha+1}{\alpha+1}}$$
$$+O(1)\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{new}})\}^{\frac{\alpha}{\alpha+1}}\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{old}}^{(\ell)})\},$$

对于某个正数 $O(1)$。

类似地， $\eta_2^{(4)}(\pi_{\text{new}}, \pi^{(\ell)\text{旧}})$ 可以用以下上界表示：

$$|\eta_2^{(4)}(\pi_{\text{new}}, \pi_{\text{old}}^{(\ell)})| \leq \sum_{a,s} |\pi_{\text{new}}(a|s) - \pi^{opt}(a|s)||A^{\pi^{opt}}(a,s)||d^{\pi_{\text{new}},\nu}(s) - d^{\pi_{\text{old}}^{(\ell)},\nu}(s)|$$
$$+\sum_{a,s} |\pi_{\text{old}}^{(\ell)}(a|s) - \pi^{opt}(a|s)||A^{\pi^{opt}}(a,s)||d^{opt,\nu}(s) - d^{\pi_{\text{old}}^{(\ell)},\nu}(s)|$$
$$+\sum_{a,s} |\pi_{\text{old}}^{(\ell)}(a|s) - \pi^{opt}(a|s)||A^{\pi^{opt}}(a,s)||d^{\pi_{\text{new}},\nu}(s) - d^{opt,\nu}(s)|.$$

利用引理 1 和定理 1 的证明中的类似论证，右端的第一项可以用 $O(\sqrt{\delta}\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{new}})\})$ 作为上界。第二项和第三项可以用以下上界表示：

$$O(1)\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{old}}^{(\ell)})\}^{\frac{2\alpha+1}{\alpha+1}} + O(1)\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{new}})\}^{\frac{\alpha}{\alpha+1}}\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{old}}^{(\ell)})\},$$

采用对 $|\eta_2^{(3)}(\pi_{\text{new}}, \pi^{(\ell)\text{旧}})|$ 进行上界估计时的类似论证。

综上所述，我们已经证明

$$|\eta_2(\pi_{\text{new}}, \pi_{\text{old}}^{(\ell)})| \leq O(1)\sqrt{\delta}\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{new}})\} + O(1)\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{old}}^{(\ell)})\}^{\frac{2\alpha+1}{\alpha+1}}$$
$$+O(1)\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{new}})\}^{\frac{\alpha}{\alpha+1}}\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{old}}^{(\ell)})\},$$

对于某个正常数 $O(1)$。根据 Hölder 不等式，第二行的最后一项可由 $\sqrt{\delta}\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{new}})\} + O(1)\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi^{(\ell)\text{旧}})\}^{\alpha+1}/\sqrt{\delta}$ 上界。当 $\mathcal{V}(\pi^{(\ell)\text{旧}})$ 与 $\mathcal{V}(\pi^{opt})$ 一致时， $\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi^{(\ell)\text{旧}})\}^{\alpha+1}/\sqrt{\delta} = \{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi^{(\ell)\text{旧}})\}^{\frac{2\alpha+1}{\alpha+1}}/\sqrt{\delta}$。由此可得

$$|\eta_2(\pi_{\text{new}}, \pi_{\text{old}}^{(\ell)})| \leq O(1)\sqrt{\delta}\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{new}})\} + O(1)\{\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi_{\text{old}}^{(\ell)})\}^{\frac{2\alpha+1}{\alpha+1}}.$$

证明至此完成。

步骤 2。首先介绍一些记号。对于 $\ell = 1, \cdots, \mathbb{L}$，我们用 $\psi(o;\ \pi, \pi^{(\ell)}\text{ old }\widehat{Q}^{(\ell)}, \widehat{\omega}^{(\ell)}, \widehat{d}^{(\ell)})$ 表示 $\psi(o;\ \pi, \pi^{(\ell)}\text{ old }\widehat{V}^{(\ell)}\ \widehat{A}^{(\ell)}, \widehat{\omega}^{(\ell)}, \widehat{d}^{(\ell)})$，因为 $\widehat{V}^{(\ell)}$ 和 $\widehat{A}^{(\ell)}$ 都是由 $\widehat{Q}^{(\ell)}$ 得出的。类似地，我们用记号 $\psi_1(\pi, \pi^{(\ell)}\text{ old }\widehat{Q}^{(\ell)}, \widehat{d}^{(\ell)})$、 $\psi_2(o;\ \pi, \pi^{(\ell)}\text{ old }\widehat{Q}^{(\ell)}, \widehat{\omega}^{(\ell)}, \widehat{d}^{(\ell)})$ 和 $\psi_3(o;\ \pi, \pi^{(\ell)}\text{ old }\widehat{Q}^{(\ell)}, \widehat{\omega}^{(\ell)}, \widehat{d}^{(\ell)})$ 表示 $\psi_1(\pi, \pi^{(\ell)}\text{ old }\widehat{A}^{(\ell)}, \widehat{d}^{(\ell)})$、 $\psi_2(o;\ \pi, \pi^{(\ell)}\text{ old }\widehat{V}^{(\ell)}\ \widehat{A}^{(\ell)}, \widehat{\omega}^{(\ell)}, \widehat{d}^{(\ell)})$

以及 $\psi_3(o;\ \pi, \pi^{(\ell)}\text{old}\ \widehat{A}^{(\ell)}, \widehat{\omega}^{(\ell)}, \widehat{d}^{(\ell)})$。记号 $\psi_1(\pi, \pi^{(\ell)}\text{old}\ Q^{\pi^{(\ell)}}\text{old}\ d^{\pi^{(\ell)}}\text{old})$、$\psi_2(o;\ \pi, \pi^{(\ell)}\text{old}\ Q^{\pi^{(\ell)}}\text{old}\ \omega^{\pi^{(\ell)}}\text{old}\ d^{\pi^{(\ell)}}\text{old})$、$\psi_3(o;\ \pi, \pi^{(\ell)}\text{old}\ Q^{\pi^{(\ell)}}\text{old}\ \omega^{\pi^{(\ell)}}\text{old}\ d^{\pi^{(\ell)}}\text{old})$ 和 $\psi(o;\ \pi, \pi^{(\ell)}\text{old}\ Q^{\pi^{(\ell)}}\text{old}\ \omega^{\pi^{(\ell)}}\text{old}\ d^{\pi^{(\ell)}}\text{old})$ 也可类似地定义。

我们旨在应用引理 4 来证明

$$\sup_{\pi\in\Pi} \frac{|\eta_1(\pi^{opt}, \pi_{\text{old}}^{(\ell)}) - \eta_1(\pi, \pi_{\text{old}}^{(\ell)}) + \widehat{\eta}_1^*(\pi, \pi_{\text{old}}^{(\ell)}) - \widehat{\eta}_1^*(\pi^{opt}, \pi_{\text{old}}^{(\ell)})|}{\mathbb{E}_{S^*\sim d^{\pi^{opt}}} \|\pi(\bullet|S^*) - \pi_{\text{old}}^{(\ell)}(\bullet|S^*)\|_{\text{TV}}} = O_p\{(NT)^{\frac{\kappa_4-1}{2}} \log(NT)\},$$

其中 $\widehat{\eta}_1^*(\pi, \pi^{(\ell)}\text{旧}) = (|\mathcal{I}_\ell|T)^{-1} \sum_{i\in\mathcal{I}_\ell} \sum_{t=0}^{T} \psi(O_{i,\ t};\ \pi, \pi^{(\ell)}\text{旧}\ Q^{\pi^{(\ell)}}\text{旧}\ \omega^{\pi^{(\ell)}}\text{旧}\ d^{\pi^{(\ell)}}\text{旧})$。为此，我们首先注意到，由 (C6) 可知，$\{(S_t, A_t, R_t, S_{t+1})\}_{t\geq 0}$ 是指数级 β-混合的。由于观测数据集由多条相互独立的轨迹组成，因此它也具有指数级 β-混合性。因此，在引理 4 中将整数 q 设为某个充分大的常数 c 乘以 $\log(NT)$，可得 $NT\beta(q)/q = o(1)$。

在给定条件下，即时奖励和条件概率比均一致有界。根据 (C4)，Π 属于一个具有有界包络函数且 VC 指数有界于 $O\{(NT)^{\kappa_4}\}$ 的 VC 型函数类。函数类$\{\psi(o;\ \pi, \pi^{(\ell)}\text{旧}\ Q^{\pi^{(\ell)}}\text{旧}\ \omega^{\pi^{(\ell)}}\text{旧}\ d^{\pi^{(\ell)}}\text{旧}) - \psi(o;\ \pi, \pi^{opt}, Q^{\pi^{(\ell)}}\text{旧}\ \omega^{\pi^{(\ell)}}\text{旧}\ d^{\pi^{(\ell)}}\text{旧}) :\ \pi\in\Pi\}$同样如此。根据引理 4，只需证明

$$\begin{aligned}\text{Var}\{\psi(O_0;\pi, \pi_{\text{old}}^{(\ell)}, Q^{\pi_{\text{old}}^{(\ell)}}, \omega^{\pi_{\text{old}}^{(\ell)}}, d^{\pi_{\text{old}}^{(\ell)}}) - \psi(O_0;\pi, \pi^{opt}, Q^{\pi_{\text{old}}^{(\ell)}}, \omega^{\pi_{\text{old}}^{(\ell)}}, d^{\pi_{\text{old}}^{(\ell)}})\}\\ \leq c\{\mathbb{E}_{S^*\sim d^{\pi^{opt}}} \|\pi(\bullet|S^*) - \pi_{\text{old}}^{(\ell)}(\bullet|S^*)\|_{\text{TV}}\}^2,\end{aligned}$$

and

$$\begin{aligned}|\psi(O_0;\pi, \pi_{\text{old}}^{(\ell)}, Q^{\pi_{\text{old}}^{(\ell)}}, \omega^{\pi_{\text{old}}^{(\ell)}}, d^{\pi_{\text{old}}^{(\ell)}}) - \psi(O_0;\pi, \pi^{opt^{(\ell)}}, Q^{\pi_{\text{old}}^{(\ell)}}, \omega^{\pi_{\text{old}}^{(\ell)}}, d^{\pi_{\text{old}}^{(\ell)}})|\\ \leq c\mathbb{E}_{S^*\sim d^{\pi^{opt}}} \|\pi(\bullet|S^*) - \pi_{\text{old}}^{(\ell)}(\bullet|S^*)\|_{\text{TV}},\end{aligned}$$

以概率 1 对某个常数 $c > 0$ 成立。根据 ψ 的定义，这一点显而易见。为简洁起见，我们省略细节。

步骤 3。我们在此步骤中证明引理 2。根据式 (5)，可得

$$\begin{aligned}\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi) &= -\frac{1}{1-\gamma}\mathbb{E}_{S^*\sim d^{\pi,\nu}} \sum_{a\in\mathcal{A}} \pi(a|S^*) A^{\pi^{opt}}(a, S^*)\\ &= -\frac{1}{1-\gamma}\mathbb{E}_{S^*\sim d^{\pi,\nu}} \sum_{a\in\mathcal{A}} \{\pi(a|S^*) - \pi^{opt}(a|S^*)\} A^{\pi^{opt}}(a, S^*).\end{aligned} \tag{34}$$

如第 4.1 节所述，$A^{\pi^{opt}}(a, s) \leq 0$ 对任意 a 和 s 均成立。接下来我们提出

$$\{\pi(a|s) - \pi^{opt}(a|s)\} A^{\pi^{opt}}(a, s) \leq 0, \quad \forall a, s. \tag{35}$$

若 $a = \text{argmax}_{a'} Q^{\pi^{opt}}(a', s)$，则 $A^{\pi^{opt}}(a, s) = 0$，且（35）自动成立。否则，我们有 $A^{\pi^{opt}}(a, s) < 0$，且 $\pi^{opt}(a|s) = 0$。由此可得（35）成立。

将 (34) 与 (35) 结合可得

$$\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi) = \frac{1}{1-\gamma}\mathbb{E}_{S^* \sim d^{\pi,\nu}} \sum_{a \in \mathcal{A}} |\pi(a|S^*) - \pi^{opt}(a|S^*)||A^{\pi^{opt}}(a, S^*)|.$$

由于 $d^{\pi,\nu}(s) \geq (1-\gamma)\nu(s)$ 且 ν 一致地远离零，因此存在某个常数 $C > 0$，使得 $d^{\pi,\nu}(s) \geq (1-\gamma)C$ 对任意 s 均成立。类似于 (33)，我们可以证明

$$\begin{aligned}\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi) &= \frac{1}{1-\gamma}\mathbb{E}_{S^* \sim d^{\pi,\nu}} \sum_{a \in \mathcal{A}} |\pi(a|S^*) - \pi^{opt}(a|S^*)||A^{\pi^{opt}}(a, S^*)| \frac{d^{\pi,\nu}(S^*)}{d^{\pi^{opt},\nu}(S^*)} \\ &\geq C\mathbb{E}_{S^* \sim d^{\pi^{opt},\nu}} \sum_{a \in \mathcal{A}} |\pi(a|S^*) - \pi^{opt}(a|S^*)||A^{\pi^{opt}}(a, S^*)|.\end{aligned}$$

记 $a^* = \text{argmax}_{a'} Q^{\pi^{opt}}(a', S^*)$。则 $A^{\pi^{opt}}(a^*, S^*) = 0$。对于任意 $\epsilon > 0$，可得

$$\begin{aligned}\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi) &\geq C\mathbb{E}_{S^* \sim d^{\pi^{opt},\nu}} \sum_{a \in \mathcal{A}, a \neq a^*} |\pi(a|S^*) - \pi^{opt}(a|S^*)||A^{\pi^{opt}}(a, S^*)| \mathbb{I}(A^{\pi^{opt}}(a, S^*) \leq -\epsilon) \\ &\geq \epsilon C\mathbb{E}_{S^* \sim d^{\pi^{opt},\nu}} \sum_{a \in \mathcal{A}, a \neq a^*} |\pi(a|S^*) - \pi^{opt}(a|S^*)| \mathbb{I}(A^{\pi^{opt}}(a, S^*) \leq -\epsilon) \\ &\geq \epsilon C\mathbb{E}_{S^* \sim d^{\pi^{opt},\nu}} \sum_{a \in \mathcal{A}, a \neq a^*} |\pi(a|S^*) - \pi^{opt}(a|S^*)| \\ &\quad - \epsilon C\mathbb{E}_{S^* \sim d^{\pi^{opt},\nu}} \sum_{a \in \mathcal{A}, a \neq a^*} |\pi(a|S^*) - \pi^{opt}(a|S^*)| \mathbb{I}(-\epsilon < A^{\pi^{opt}}(a, S^*) < 0).\end{aligned}$$

在 (C5) 中的间隔型条件下，最后一行是 $O(\epsilon^{\alpha+1})$。我们取

$$\epsilon = c\{\mathbb{E}_{S^* \sim d^{\pi^{opt},\nu}} \sum_{a \in \mathcal{A}, a \neq a^*} |\pi(a|S^*) - \pi^{opt}(a|S^*)|\}^{1/\alpha},$$

其中 $c > 0$ 是某个常数。对 c 进行适当的性质选择后，最后一行的下界为 $\epsilon C\mathbb{E}_{S^* \sim d^{\pi^{opt},\nu}} \sum_{a \in \mathcal{A}, a \neq a^*} |\pi(a|S^*) - \pi^{opt}(a|S^*)|/2$。注意

$$\sum_{a \in \mathcal{A}, a \neq a^*} |\pi(a|S^*) - \pi^{opt}(a|S^*)| = |\pi(a^*|S^*) - \pi^{opt}(a^*|S^*)|$$

因此，

$$\mathcal{V}(\pi^{opt}) - \mathcal{V}(\pi) \geq \frac{cC}{4}\{\mathbb{E}_{S^* \sim d^{\pi^{opt},\nu}} \sum_{a \in \mathcal{A}} |\pi(a|S^*) - \pi^{opt}(a|S^*)|\}^{1+1/\alpha}.$$

注意到 $\sum_{a \in \mathcal{A}} |\pi(a|S^*) - \pi^{opt}(a|S^*)| = 2\|\pi(\bullet|S^*) - \pi^{opt}(\bullet|S^*)\|_{\text{TV}}$，证明即告完成。

步骤 4. 我们在这一步证明引理 3。首先，我们证明偏差项

$$\Delta\psi(\pi_1,\pi_2,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\widehat{\omega}^{(\ell)},\widehat{d}^{(\ell)})=\mathbb{E}[\psi(O_0;\pi_1,\pi_{\text{old}},\widehat{Q}^{(\ell)},\widehat{\omega}^{(\ell)},\widehat{d}^{(\ell)})|\{O_{i,t}\}_{i\in\mathcal{I}_\ell,0\le t<T}]$$
$$-\mathbb{E}[\psi(O_0;\pi_2,\pi_{\text{old}},\widehat{Q}^{(\ell)},\widehat{\omega}^{(\ell)},\widehat{d}^{(\ell)})|\{O_{i,t}\}_{i\in\mathcal{I}_\ell,0\le t<T}]-\eta_1(\pi_1,\pi_{\text{old}}^{(\ell)})+\eta_1(\pi_2,\pi_{\text{old}}^{(\ell)})$$

其中，小 o 项关于 π_1、π_2 一致。接下来，我们证明 $\sup_{\pi_1,\ \pi_2\in\Pi}|\widehat{\eta}^{(\ell)}(\pi_1)-\widehat{\eta}^{*}(\pi_1,\pi^{(\ell)}{}_{\text{旧}})-\widehat{\eta}^{(\ell)}(\pi_2)+\widehat{\eta}^{*}(\pi_2,\pi^{(\ell)}{}_{\text{旧}})-\Delta\psi(\pi_1,\pi_2,\pi^{(\ell)}{}_{\text{旧}})|=o_p\{(NT)^{\kappa_4/2-1/2}\}$。证明至此完成。

为了界定偏差项，我们首先注意到，当 $\widehat{\omega}^{(\ell)}=\omega^{\pi^{(\ell)}{}_{\text{旧}}}$ 时，

$$\psi_1(\pi,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\widehat{d}^{(\ell)})+\mathbb{E}\{\psi_2(O_0;\pi,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\omega^{\pi_{\text{old}}^{(\ell)}},\widehat{d}^{(\ell)})|\widehat{Q}^{(\ell)},\widehat{d}^{(\ell)}\}=\psi_1(\pi,\pi_{\text{old}}^{(\ell)},Q^{\pi_{\text{old}}^{(\ell)}},\widehat{d}^{(\ell)}),$$
$$\psi_1(\pi,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\widehat{d}^{(\ell)})+\mathbb{E}\{\psi_3(O_0;\pi,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\omega^{\pi_{\text{old}}^{(\ell)}},\widehat{d}^{(\ell)})|\widehat{Q}^{(\ell)},\widehat{d}^{(\ell)}\}=\psi_1(\pi,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},d^{\pi_{\text{old}}^{(\ell)}}).$$

这两条断言可通过与 A.2 节中类似的论证证明。因此，我们有

$$\begin{aligned}\Delta\psi(\pi_1,\pi_2,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\omega^{\pi_{\text{old}}^{(\ell)}},\widehat{d}^{(\ell)})&=\psi_1(\pi_1,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},d^{\pi_{\text{old}}^{(\ell)}})+\psi_1(\pi_1,\pi_{\text{old}}^{(\ell)},Q^{\pi_{\text{old}}^{(\ell)}},\widehat{d}^{(\ell)})\\&-\psi_1(\pi_1,\pi_{\text{old}}^{(\ell)},Q^{\pi_{\text{old}}^{(\ell)}},d^{\pi_{\text{old}}^{(\ell)}})-\psi_1(\pi_1,\pi_{\text{old}}^{(\ell)},Q^{\pi_{\text{old}}^{(\ell)}},d^{\pi_{\text{old}}^{(\ell)}})-\psi_1(\pi_2,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},d^{\pi_{\text{old}}^{(\ell)}})\\&-\psi_1(\pi_2,\pi_{\text{old}}^{(\ell)},Q^{\pi_{\text{old}}^{(\ell)}},\widehat{d}^{(\ell)})+\psi_1(\pi_2,\pi_{\text{old}}^{(\ell)},Q^{\pi_{\text{old}}^{(\ell)}},d^{\pi_{\text{old}}^{(\ell)}})+\psi_1(\pi_2,\pi_{\text{old}}^{(\ell)},Q^{\pi_{\text{old}}^{(\ell)}},d^{\pi_{\text{old}}^{(\ell)}})\\&=\sum_{s,a}\{\pi_1(a|s)-\pi_2(a|s)\}\{\widehat{Q}^{(\ell)}(a,s)-Q^{\pi_{\text{old}}^{(\ell)}}(a,s)\}\{d^{\pi_{\text{old}},\nu}(s)-\widehat{d}^{(\ell),\nu}(s)\}.\end{aligned}$$

根据柯西-施瓦茨不等式，右端可由以下上界界定

$$\sqrt{\sum_{s,a}|\pi_1(a|s)-\pi_2(a|s)||\widehat{Q}^{(\ell)}(a,s)-Q^{\pi_{\text{old}}^{(\ell)}}(a,s)|^2p_\infty(a,s)}\tag{36}$$
$$\times\ \sqrt{\sum_{s,a}|\pi_1(a|s)-\pi_2(a|s)||\widehat{d}^{(\ell),\nu}(s)-d^{\pi_{\text{old},\nu}^{(\ell)}}(s)|^2/p_\infty(a,s)}.\tag{37}$$

注意，在 ν 一致地远离零的假设下，$\max_{a,\ s}|\pi_1(a|s)-\pi_2(a|s)|\le\sum_{a,\ s}|\pi_1(a|s)-\pi_2(a|s)|\le C\sum_a\mathbb{E}_{S^*\sim d^{\pi^{opt,\ \nu}}}|\pi_1(a|S^*)-\pi_2(a|S^*)|$。在 (C1) 条件下，(36) 的上界为 $O_p\{(NT)^{-\kappa_1}\}\mathbb{E}_{S^*\sim d^{\pi^{opt,\ \nu}}}\|\pi_1(\bullet|S^*)-\pi_2(\bullet|S^*)\|_{\text{TV}}$。由于 p_∞ 一致地远离零，(37) 可由以下上界界定

$$c\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}|\pi_1(a|S^*)-\pi_2(a|S^*)|\sqrt{\max_s|\widehat{d}^{(\ell),\nu}(s)-d^{\pi_{\text{old}}^{(\ell)},\nu}(s)|^2},$$

对于某个常数 $c>0$。注意，$\widehat{d}^{(\ell),\nu}(s)-d^{\pi^{(\ell)}}\text{old}^{\nu}(s)$ 可以表示为 $\mathbb{E}_{S^*\sim\widehat{d}^{(\ell),\nu}}\mathbb{I}(S^*=s)-\mathbb{E}_{S^*\sim d^{\pi^{(\ell)}}}\text{old}_{\nu}\mathbb{I}(S^*=s)=\sum_{a^*,s^*}\pi^{(\ell)}\text{old}(a^*|s^*)\nu(s^*)(\mathbb{E}_{S^*\sim\widehat{d}^{(\ell)}(\bullet|a^*,s^*)}\mathbb{I}(S^*=s)-\mathbb{E}_{S^*\sim d^{\pi^{(\ell)}}}\text{old}(\bullet|a^*,s^*)\mathbb{I}(S^*=$

s）因此，它可以由以下内容给出上界

$$O(1)\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}|\pi_1(a|S^*)-\pi_2(a|S^*)|\sum_a \mathbb{E}_{S_0\sim\nu}\pi_{\text{old}}^{(\ell)}(a|S_0)\mathcal{D}_{\text{TV}}\{d^{\pi_{\text{old}}^{(\ell)}}(\bullet|a,S_0),\widehat{d}^{(\ell)}(\bullet|a,S_0)\},$$

其中 $O(1)$ 表示某个正常数。由于 b 一致地远离零，因此可以进一步将其上界为

$$O\left[\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}|\pi_1(a|S^*)-\pi_2(a|S^*)|\mathbb{E}_{(A,S)\sim p_\infty}\mathcal{D}_{\text{TV}}\{d^{\pi_{\text{old}}^{(\ell)}}(\bullet|A,S),\widehat{d}^{(\ell)}(\bullet|A,S)\}\right].$$

因此，在（C3）条件下，（37）为 $O_p\{(NT)^{-\kappa_3}\}\mathbb{E}_{S^*\sim d^{\pi^{opt},\ \nu}}|\pi_1(a|S^*)-\pi_2(a|S^*)|$。在满足 $\kappa_1+\kappa_3>1/(2+2\alpha)$ 的条件下，我们得到 $\Delta\psi(\pi_1,\pi_2,\pi^{(\ell)}\text{旧}\,\widehat{Q}^{(\ell)},\omega^{\pi^{(\ell)}}\text{旧}\,\widehat{d}^{(\ell)})=o_p\{(NT)^{-1/(2+2\alpha)}\}\mathbb{E}_{S^*\sim d^{\pi^{opt},\ \nu}}|\pi_1(a|S^*)-\pi_2(a|S^*)|$。

只需对 $\Delta\psi(\pi_1,\pi_2,\pi^{(\ell)}\text{旧}\,\widehat{Q}^{(\ell)},\omega^{\pi^{(\ell)}}\text{旧}\,\widehat{d}^{(\ell)})-\Delta\psi(\pi_1,\pi_2,\pi^{(\ell)}\text{旧}\,\widehat{Q}^{(\ell)},\widehat{\omega}^{(\ell)},\widehat{d}^{(\ell)})$ 给出上界，或者等价地，

$$\begin{aligned}&\mathbb{E}\{\psi_2(O_0;\pi_1,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\omega^{\pi_{\text{old}}^{(\ell)}},\widehat{d}^{(\ell)})-\psi_2(O_0;\pi_1,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\widehat{\omega}^{(\ell)},\widehat{d}^{(\ell)})\}\\-&\mathbb{E}\{\psi_2(O_0;\pi_2,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\omega^{\pi_{\text{old}}^{(\ell)}},\widehat{d}^{(\ell)})-\psi_2(O_0;\pi_2,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\widehat{\omega}^{(\ell)},\widehat{d}^{(\ell)})\}\end{aligned}\tag{38}$$

$$\begin{aligned}&\mathbb{E}\{\psi_3(O_0;\pi_1,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\omega^{\pi_{\text{old}}^{(\ell)}},\widehat{d}^{(\ell)})-\psi_3(O_0;\pi_1,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\widehat{\omega}^{(\ell)},\widehat{d}^{(\ell)})\}\\-&\mathbb{E}\{\psi_3(O_0;\pi_2,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\omega^{\pi_{\text{old}}^{(\ell)}},\widehat{d}^{(\ell)})-\psi_3(O_0;\pi_2,\pi_{\text{old}}^{(\ell)},\widehat{Q}^{(\ell)},\widehat{\omega}^{(\ell)},\widehat{d}^{(\ell)})\}.\end{aligned}\tag{39}$$

同样，我们可以证明 (38) 和 (39) 可以由以下内容给出上界

$$C\sqrt{\mathbb{E}_{(A,S)\sim p_\infty}|\widehat{Q}^{(\ell)}(A,S)-Q^{\pi_{\text{old}}^{(\ell)}}(A,S)|^2}\sqrt{\mathbb{E}_{\substack{(A,S)\sim p_\infty\\(\widetilde{A},\widetilde{S}\sim p_\infty)}}|\widehat{\omega}^{(\ell)}(\widetilde{A},\widetilde{S};A,S)-\omega^{\pi_{\text{old}}^{(\ell)}}(\widetilde{A},\widetilde{S};A,S)|^2}$$
$$\times\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\|\pi_1(\bullet|S^*)-\pi_2(\bullet|S^*)\|_{\text{TV}},$$

and

$$C\sqrt{\mathbb{E}_{(A,S)\sim p_\infty}\mathcal{D}_{\text{TV}}^2\{d^{\pi_{\text{old}}^{(\ell)}}(\bullet|A,S),\widehat{d}^{(\ell)}(\bullet|A,S)\}}\sqrt{\mathbb{E}_{\substack{(A,S)\sim p_\infty\\(\widetilde{A},\widetilde{S})\sim p_\infty}}|\widehat{\omega}^{(\ell)}(\widetilde{A},\widetilde{S};A,S)-\omega^{\pi_{\text{old}}^{(\ell)}}(\widetilde{A},\widetilde{S};A,S)|^2}$$
$$\times\mathbb{E}_{S^*\sim d^{\pi^{opt},\nu}}\|\pi_1(\bullet|S^*)-\pi_2(\bullet|S^*)\|_{\text{TV}},$$

对于某个常数 $C>0$。在 (C2) 和 (C3) 的条件下，我们得到 $\Delta\psi(\pi_1,\pi_2,\pi^{(\ell)}\text{旧}\,\widehat{Q}^{(\ell)},\omega^{\pi^{(\ell)}}\text{旧}\,\widehat{d}^{(\ell)})-\Delta\psi(\pi_1,\pi_2,\pi^{(\ell)}\text{旧}\,\widehat{Q}^{(\ell)},\widehat{\omega}^{(\ell)},\widehat{d}^{(\ell)})=o_p\{(NT)^{-1/(2+2\alpha)}\}$。

剩下需要证明的是 $\sup_{\pi_1,\ \pi_2\in\Pi}|\widehat{\eta}^{(\ell)}(\pi_1)-\widehat{\eta}^*(\pi_1,\pi^{(\ell)}\text{旧})-\widehat{\eta}^{(\ell)}(\pi_2)+\widehat{\eta}^*(\pi_2,\pi^{(\ell)}\text{旧})-\Delta\psi(\pi_1,\pi_2,\pi^{(\ell)}\text{旧})|=o_p\{(NT)^{\kappa_4/2-1/2}\}$。这可以通过步骤 2 中的类似论证来证明。为节省篇幅，我们省略细节。

步骤 5。最后一步证明引理 4。我们首先使用 Berbee耦合引理

（参见 Dedecker 和 Louhichi，2002 年的引理 4.1），通过独立同分布随机变量之和来近似 $\sup_{f\in\mathcal{F}}|\sum_{t=0}^{T-1} f(Z_t)|$。然后，我们应用 Chernozhukov等人（2014 年）推论 5.1 中的最大不等式，来界定经验过程的期望。

根据 Dedecker 和 Louhichi (2002) 引理 4.1 下方的讨论，我们可以构造一列随机变量 $\{Z_t^0:\ t\geq 0\}$，使得

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t)\right| = \sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t^0)\right|, \tag{40}$$

其成立的概率至少为 $1-T\beta(q)/q$，且序列 $\{U_{2i}^0:\ i\geq 0\}$ 和 $\{U_{2i+1}^0:\ i\geq 0\}$ 独立同分布，其中 $U_i^0=(Z_{iq}^0, Z_{iq+1}^0,\cdots,Z_{iq+q-1}^0)$。

回顾 $\mathcal{I}_r=\{q\lfloor T/q\rfloor, q\lfloor T/q\rfloor+1,\cdots,T-1\}$，我们有

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t^0)\right| \leq \sum_{j=0}^{q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/q\rfloor} f(Z_{tq+j}^0)\right| + \sup_{f\in\mathcal{F}}\left|\sum_{t\in\mathcal{I}_r} f(Z_t^0)\right|.$$

在对 F 的有界性假设下，右端（RHS）的第二项有上界 Cq。不失一般性，假设 $\lfloor T/q\rfloor$ 是偶数。右端的第一项有上界 $\sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)|$。综上，我们已证明

$$\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{T-1} f(Z_t^0)\right| \leq \sum_{j=0}^{2q-1}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq+j}^0)\right| + Cq. \tag{41}$$

根据构造，$\{Z_{2tq}^0:\ t\geq 0\}$ 是独立同分布的。接下来需要界定 $\mathbb{E}\sup_{f\in\mathcal{F}}|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)|$。由 Chernozhukov等人 (2014) 的推论 5.1 可得

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z_{2tq}^0)\right| \preceq \sqrt{\frac{v\sigma^2 T}{q}\log\left(\frac{AC}{\sigma}\right)} + vC\log\left(\frac{AC}{\sigma}\right).$$

因此，根据 (40)、(41) 和马尔可夫不等式即可得出该结论。

C.3 定理3的证明

为简化记号，我们省略 $\pi^{(\ell)}{}_{\text{old}}$ 中的下标 ℓ。与引理 3 类似，我们可以证明，对于任意 π、$\pi_{\text{old}}\in\Pi$，当干扰函数以快于 $o_p((NT)^{-1/4})$ 的速率收敛时，$\widehat{\eta}_1(\pi,\pi_{\text{old}}) = \widehat{\eta}_1^*(\pi,\pi_{\text{old}}) + o_p\{(NT)^{-1/2}\}$。注意，$\widehat{\eta}_1^*(\pi,\pi_{\text{old}})$ 对 $\eta_1(\pi,\pi_{\text{old}})$ 是无偏的。只需证明

$$\frac{\widehat{\eta}_1^*(\pi,\pi_{\text{old}})-\eta_1(\pi,\pi_{\text{old}})}{\sqrt{\text{EB}(\pi,\pi_{\text{old}})}}\xrightarrow{d}N(0,1). \tag{42}$$

根据定义，$\widehat{\eta}_1^*(\pi,\pi_{\text{old}})-\eta_1(\pi,\pi_{\text{old}})=(NT)^{-1}\sum_{i=1}^{N}\sum_{t=0}^{T-1}\{\psi_2(O_{i,\ t};\ \pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})+\psi_3(O_{i,\ t};\ \pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\}$。右端的和可以表示为

$$\frac{1}{NT}\sum_{g=1}^{NT}\{\psi_2(O_{i(g),t(g)};\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})+\psi_3(O_{i(g),t(g)};\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\}, \tag{43}$$

其中，$i(g)$ 和 $t(g)$ 是满足 $\{i(g)-1\}T+t(g)=g-1$ 的唯一整数。在 (A1) 和 (A2) 下，(43) 对应于一个鞅差序列的和。假设我们可以证明

$$\begin{aligned}\text{Var}\{\psi_2(O_{i(g),t(g)};\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})+\psi_3(O_{i(g),t(g)};\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\}\\ =NT\text{EB}(\pi,\pi_{\text{old}}).\end{aligned} \tag{44}$$

在给定的假设下，我们可以证明，Brown 等人 (1971年) 的定理 1 中的条件自动得到满足。根据 Brown 等人 (1971年) 提出的鞅中心极限定理，可得 (42) 成立。因此，只需证明 (44) 成立即可。

注意，η_1 仅取决于转移函数 $\bar{p}$。对于给定的元组 $O=(S,A,R,S')$，满足 $(A,S)\sim p_\infty$，$(S',R|A,S)\sim\bar{p}(\bullet,\bullet|A,S)$，假设我们可以证明 (45) 成立，

$$\nabla_{\theta_1}\eta_1(\theta_1^*)=\mathbb{E}\sum_{j=2}^{3}\psi_j(O;\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\nabla_{\theta_1}\log\bar{p}_{\theta_1^*}(\bullet,\bullet|A,S). \tag{45}$$

由柯西-施瓦茨不等式可得

$$\begin{aligned}NT\text{EB}(N,T)&=\sup\nabla_{\theta_1}\eta_1(\theta_1^*)\left\{\mathbb{E}\nabla_{\theta_1}\log\bar{p}_{\theta_1^*}(S',R|A,S)\nabla_{\theta_1}\log\bar{p}_{\theta_1^*}^{\top}(S',R|A,S)\right\}^{-1}\{\nabla_{\theta_1}\eta_1(\theta_1^*)\}^{\top}\\&\le\mathbb{E}\left[\sum_{j=2}^{3}\psi_j(O;\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\right]\left[\sum_{j=2}^{3}\psi_j(O;\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\right]^{\top}.\end{aligned}$$

这意味着，所提出的估计量的方差渐近地大于或等于效率界。此外，注意到对于 $j=2$ 或 3，我们几乎处处有$\mathbb{E}\{\psi_j(O;\ \pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})|A,\ S\}=0$。利用 Kallus 和 Uehara（2019）定理 2 证明中的类似论证，我们可以证明，存在一列正则参数子模型，其 Cr´amer-Rao 下界趋近于

$$\mathbb{E}\left[\sum_{j=2}^{3}\psi_j(O;\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\right]\left[\sum_{j=2}^{3}\psi_j(O;\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\right]^{\top}.$$

这意味着，所提出的估计量的方差渐近地等于效率界。因此，证明完成。

接下来只需证明（45）。首先注意，优势函数和折扣访问概率完全由转移函数 $\bar{p}$ 决定。根据链式法则，

$$\nabla_{\theta_1}\eta_1(\theta_1^*)=\sum_{a,s}\pi(a|s)\{\nabla_{\theta_1}A^{\pi_{\text{old}}}(a,s;\theta_1^*)\}d^{\pi_{\text{old}},\nu}(s)+\sum_{a,s}\pi(a|s)A^{\pi_{\text{old}}}(a,s)\nabla_{\theta_1}d^{\pi_{\text{old}},\nu}(s;\theta_1^*),$$

其中，$A^{\pi_{\text{old}}}(\bullet,\bullet;\theta_1)$ 和 $d^{\pi_{\text{old}},\,\nu}(\bullet;\theta_1)$ 分别表示由 θ_1 索引的某个关于 $\bar{p}$ 的参数子模型下的优势函数和折扣访问概率。

运用 Kallus 和 Uehara（2019）中定理 2 证明里的类似论证，我们可以证明

$$\begin{aligned}\{\nabla_{\theta_1}A^{\pi_{\text{old}}}(a,s;\theta_1^*)\}=\frac{1}{1-\gamma}\mathbb{E}\left\{\omega^{\pi_{\text{old}}}(A,S;a,s)-\sum_{a'}\pi_{\text{old}}(a'|s)\omega^{\pi_{\text{old}}}(A,S;a',s)\right\}\\ \times\{R+\gamma V^{\pi_{\text{old}}}(S')-Q^{\pi_{\text{old}}}(A,S)\}\nabla_{\theta_1}\log\bar{p}_{\theta_1^*}^{\top}(S',R|A,S).\end{aligned}$$

由此可得

$$\sum_{a,s}\pi(a|s)\{\nabla_{\theta_1}A^{\pi_{\text{old}}}(a,s;\theta_1^*)\}d^{\pi_{\text{old}},\nu}(s)=\mathbb{E}\psi_2(O;\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\nabla_{\theta_1}\log\bar{p}_{\theta_1^*}^{\top}(S',R|A,S).$$

接下来只需证明

$$\sum_{a,s}\pi(a|s)A^{\pi_{\text{old}}}(a,s)\nabla_{\theta_1}d^{\pi_{\text{old}},\nu}(s;\theta_1^*)=\mathbb{E}\psi_3(O;\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\nabla_{\theta_1}\log\bar{p}_{\theta_1^*}^{\top}(S',R|A,S).$$

对于给定的参数子模型 $\{\bar{p}_{\theta_1}:\ \theta_1\in\Theta_1\}$，令 $p(s'|a,s;\,\theta_1)=\sum_r\bar{p}(s',r|a,s;\,\theta_1)$。经过一些计算，我们得到

$$\frac{1}{\gamma(1-\gamma)}\nabla_{\theta_1} d^{\pi_{\text{old}},\nu}(s;\theta_1^*) = \sum_{t\geq 0}\gamma^t \sum_{\{(a_j,s_j)\}_{j=0}^t} \nabla_{\theta_1}\left\{\prod_{j=0}^t \pi_{\text{old}}(a_j|s_j)p(s|a_j,s_j;\theta_1^*)\right\}\nu(s_0)$$

$$= \sum_{t\geq 0}\gamma^t \sum_{\{(a_j,s_j)\}_{j=0}^t, s_{t+1}} \mathbb{I}(s_{t+1}=s)\nabla_{\theta_1}\left\{\prod_{j=0}^t \pi_{\text{old}}(a_j|s_j)p(s_{j+1}|a_j,s_j;\theta_1^*)\right\}\nu(s_0)$$

$$= \sum_{t\geq 0}\gamma^t \sum_{\substack{\{(a_j,s_j)\}_{j=0}^t \\ s_{t+1}}} \mathbb{I}(s_{t+1}=s)\prod_{j=0}^t \pi_{\text{old}}(a_j|s_j)p(s_{j+1}|a_j,s_j)\sum_{k=0}^t \nabla_{\theta_1}\log p(s_{k+1}|a_k,s_k;\theta_1^*)\nu(s_0)$$

$$= \frac{1}{1-\gamma}\sum_{k=0}^{+\infty}\gamma^k \sum_{\substack{\{(a_j,s_j)\}_{j=0}^k \\ s_{k+1},a}} \pi_{\text{old}}(a|s_{k+1})d^{\pi_{\text{old}}}(s;a,s_{k+1})\nabla_{\theta_1}\log p(s_{k+1}|a_k,s_k;\theta_1^*)$$

$$\times\left\{\prod_{j=0}^k \pi_{\text{old}}(a_j|s_j)p(s_{j+1}|a_j,s_j)\right\}\nu(s_0).$$

根据 $\omega^{\pi_{\text{old}}}$ 的定义，最后一个方程可改写为

$$\frac{1}{(1-\gamma)^2}\sum_a \pi_{\text{old}}(a|S')d^{\pi_{\text{old}}}(s;a,S')\omega^{\pi_{\text{old}},\nu}(A,S)\nabla_{\theta_1}\log p(S'|A,S;\theta_1^*).$$

N注意，对于任意函数 h，由 $\mathbb{E}h(A,S)\nabla_{\theta_1}\log p(S'|A,S;\theta_1^*)=0$，我们得到

$$\nabla_{\theta_1} d^{\pi_{\text{old}},\nu}(s;\theta_1^*) = \frac{1}{1-\gamma}\mathbb{E}\left[\gamma\sum_a \pi_{\text{old}}(a|S')d^{\pi_{\text{old}}}(s;a,S') - \mathbb{E}\left\{\gamma\sum_a \pi_{\text{old}}(a|S')d^{\pi_{\text{old}}}(s;a,S')|A,S\right\}\right]$$
$$\times\omega^{\pi_{\text{old}},\nu}(A,S)\nabla_{\theta_1}\log p(S'|A,S;\theta_1^*)$$
$$= \frac{1}{1-\gamma}\mathbb{E}\left[\gamma\sum_a \pi_{\text{old}}(a|S')d^{\pi_{\text{old}}}(s;a,S') - d^{\pi_{\text{old}}}(s;A,S) + (1-\gamma)\mathbb{I}(s=S)\right]$$
$$\times\omega^{\pi_{\text{old}},\nu}(A,S)\nabla_{\theta_1}\log p(S'|A,S;\theta_1^*).$$

因此，我们得到

$$\sum_{a,s}\pi(a|s)A^{\pi_{\text{old}}}(a,s)\nabla_{\theta_1} d^{\pi_{\text{old}},\nu}(s;\theta_1^*) = \mathbb{E}\psi_3(O;\pi,\pi_{\text{old}},Q^{\pi_{\text{old}}},\omega^{\pi_{\text{old}}},d^{\pi_{\text{old}}})\nabla_{\theta_1}\log\bar{p}_{\theta_1^*}^{\top}(S',R|A,S).$$

至此，证明完成。

D 一些补充数值细节

D.1 玩具示例的补充结果

在本小节中，我们首先列出玩具示例的一些细节。具体而言，我们考虑以下转移矩阵：

$$p = \begin{array}{c} \\ (S=0,A=0) \\ (S=0,A=1) \\ (S=1,A=0) \\ (S=1,A=1) \end{array} \begin{array}{c} \begin{array}{cc} S'=0 & S'=1 \end{array} \\ \begin{pmatrix} 0.75 & 0.25 \\ 0.4 & 0.6 \\ 0.1 & 0.9 \\ 0.85 & 0.15 \end{pmatrix} \end{array}$$

用于生成模拟数据的行为策略为

$$b = \begin{array}{c} \\ S=0 \\ S=1 \end{array} \begin{array}{c} \begin{array}{cc} A=0 & A=1 \end{array} \\ \begin{pmatrix} 0.7 & 0.3 \\ 0.2 & 0.8 \end{pmatrix} \end{array}.$$

以下是用于检验价值提升性质和三重稳健性的各个场景的详细设计。

(i) 原点：所有干扰函数均设为预言机值。(ii) 模块1：向边际密度比中注入均匀分布 uniform(0, 2) 的噪声，而其他干扰函数均设为预言机值。(iii) 模块 2：向 Q 函数中注入均匀分布 uniform(0, 2) 的噪声，而其他干扰函数均设为预言机值。(iv) 模块3：将转移矩阵 p 乘以一个服从指数分布 exponential(1 的随机变量

)

并将所有值截断到 $[0,1],$，而其他干扰函数均设为预言机值。(vi) 模块4：按照 (ii)–(iv) 中所述的步骤，向全部三个干扰函数注入随机误差。

接下来，我们报告玩具示例中的估计策略取值（参见第5.1节），其中δ 设为 0.05 和 0.2。这些取值分别如图4和图5所示。

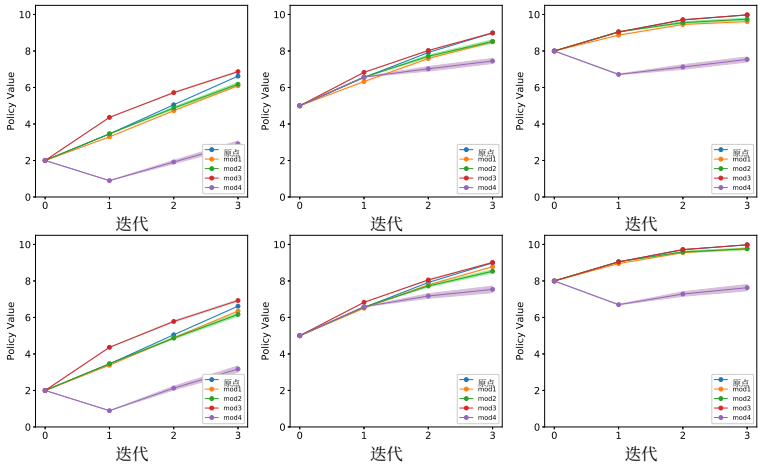

图4：玩具示例中估计策略的价值。第一行展示使用 (T, N) 组合为 (30, 30) 时的结果，第二行展示使用 (50, 50) 时的结果。三列分别表示初始策略因子 κ 取值为 0.8、0.5、0.2 的情况。横轴表示价值增强过程中使用的迭代次数。当迭代次数为零时，绘制初始策略的评估价值。最优价值为 10，且 δ 固定为 0.05。置信带根据 100 次重复试验计算得出。

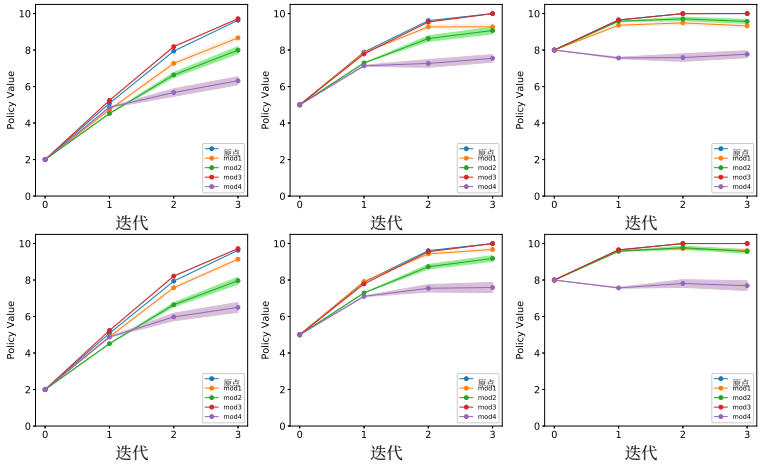

图5：玩具示例中估计策略的价值。第一行展示使用 (T, N) 对为 (30, 30) 时的结果，第二行展示使用 (50, 50) 时的结果。三列分别表示初始策略因子 κ 取值为 0.8、0.5 和 0.2 的情况。横轴表示价值增强过程中使用的迭代次数。当迭代次数为零时，我们绘制初始策略的评估价值。最优价值为 10，且 δ 固定为 0.2。置信带根据 100 次重复试验计算得出。

此外，为了展示所提方法的优势，我们使用查找表（例如，带有查表特征的线性模型）而非深度学习模型来参数化所有干扰函数（包括 Q函数、概率比和

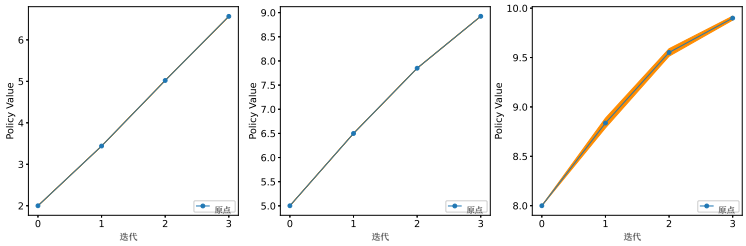

图6：玩具示例中估计策略的价值。所有干扰参数和策略均由线性函数建模。结果基于（T, N）取值为（100, 100）时计算得出。这三幅图分别对应初始策略因子 κ 取值为 0.8、0.5、0.2 的情况。横轴表示价值增强过程中使用的迭代次数。当迭代次数为零时，我们绘制初始策略的评估价值。最优价值为 10，且 δ 固定为 0.05。置信带基于 100 次重复试验计算得出。前两幅图中价值的方差非常小，因此上下置信带大部分重叠。

转移核），并将所提出的方法应用于正文第5.1节中的玩具示例设定。结果见图6。可以看出，所提出的方法仍能够提升初始策略的表现。

D.2 另一项模拟研究

在本小节中，我们开展另一项模拟研究，以展示所提出方法在有限样本下的性能。考虑一个 15 维状态向量 $S_t = (S_t^{(1)}, \cdots, S_t^{(15)})$。我们将初始状态设为一个标准正态向量。令前两个状态变量按照以下方式演化：对于 $t \geq 1$

$$S_t^{(1)} = (3/4)(2A_{t-1} - 1)S_{t-1}^{(1)} + (1/4)S_{t-1}^{(2)} + \epsilon_t^{(1)},$$
$$S_t^{(2)} = (3/4)(1 - 2A_{t-1})S_{t-1}^{(2)} + (1/4)S_{t-1}^{(1)} + \epsilon_t^{(2)},$$

其中，A_t 以相等概率取 $\{0, 1\}$ 中的值，$\{\epsilon_t^{(1)}\}_t$和 $\{\epsilon_t^{(2)}\}_t$ 是 i.i.d. $N(0, 1/4)$ 随机误差。这两个变量的数据生成机制与 Luckett 等人（2020）和 Liao 等人（2020）的模拟设定中所考虑的机制类似。对于所有$0 \leq t \leq T - 1$，其他状态变量均独立地从标准正态分布中抽取。将奖励函数定义为 $R_t = 2S_{t+1}^{(1)} + S_{t+1}^{(2)} - (1/4)(2A_t - 1)$。在我们的实验中，样本量对 (T, N) 设为 (50, 100)、(25, 200) 或 (100, 50)。我们考虑 γ的两种选择，分别对应于 0.9 和 0.95。

为实现我们的方法，我们将折数 $\mathbb{L}$ 设为 2，并将常数 δ 设为

0.05。为获得 π_{old}，如第3.2.1节所述，我们在实验中采用价值学习、拟合Q迭代和保守Q学习，计算出三种不同的初始策略。拟合Q迭代和保守Q学习都需要对最优Q函数建模。在我们的实现中，我们使用具有两个隐藏层的修正线性单元（ReLU）神经网络。为实现V学习，我们使用Luckett 等人（2020年）开发的R软件包。具体而言，我们使用径向基函数对状态价值函数建模，并使用线性基函数对策略类建模。我们还使用具有两个隐藏层的修正线性单元（ReLU）神经网络对 π_{new}建模。

结果汇总于图7。可以看出，所有这些初始策略都有可能通过我们的方法得到改进。这证明了我们的价值增强方法具有卓越的性能。此外，当 π_{old} 与最优策略相距较远时，改进效果十分显著。这与我们在玩具示例中的观察一致。还可以看出，当 π_{old} 与最优策略非常接近时，我们的方法可能会遭受轻微的价值损失。这可能是因为我们在模拟中使用了固定的 δ 。如玩具示例所示，当π_{old} 与最优策略相距较远时，我们应选择较大的 δ；否则则应选择较小的 δ 。研究如何自适应地选择 δ将会很有意义。不过，这超出了本文的范围。我们将其留作未来工作。最后，我们还开展了一些额外研究，以考察所提出的方法在样本量较小、噪声水平较高的设定下的性能。具体而言，图8报告了 $N \times T = 3000$ 时的结果。图9报告了 $\{\epsilon_t^{(1)}\}_t$、 $\{\epsilon_t^{(2)}\}_t$ 为独立同分布的 $N(0, 1)$随机误差时的结果。所得结论与图7中的结果相似。

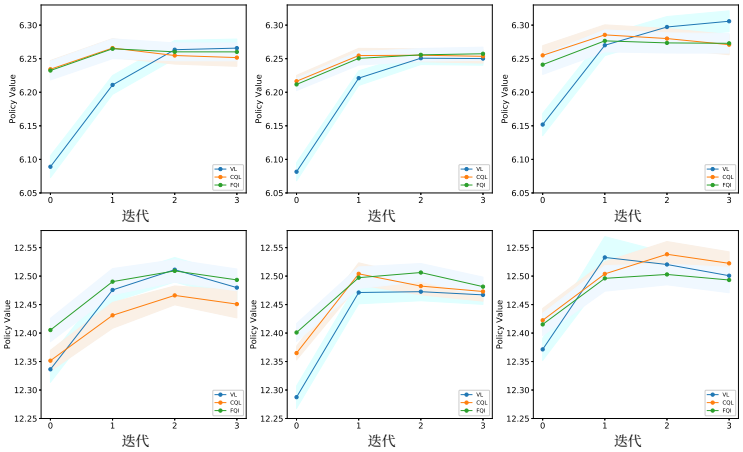

图7：初始策略由VL、保守Q学习和拟合Q迭代计算得到的估计策略的价值。第一行表示使用 $\gamma = 0.9$ 的结果，第二行表示使用 $\gamma = 0.95$的结果。三列分别表示使用(T, N)组合为(50, 100)、(25, 200)、(100, 50)的结果。在$\gamma = 0.9$ 下，最优价值约为6.89；在 $\gamma = 0.95$下约为13.47。置信带根据100次重复试验计算得出。

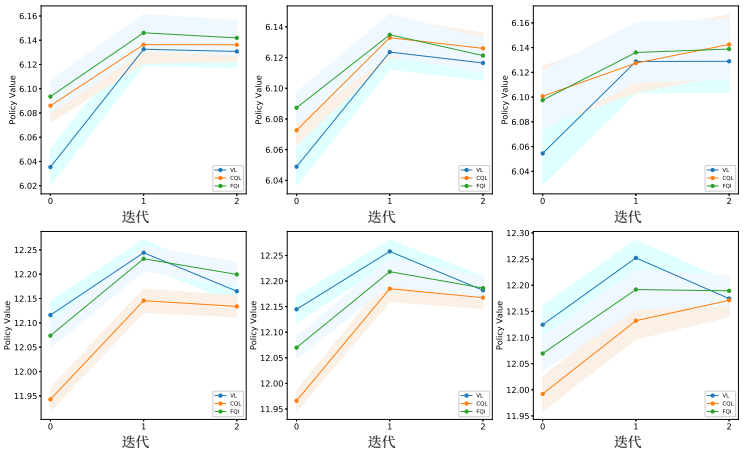

图8：初始策略分别由 VL、保守Q学习和拟合Q迭代计算得出的估计策略价值。第一行展示使用 $\gamma = 0.9$ 时的结果，第二行展示使用 $\gamma = 0.95$时的结果。三列分别展示使用（T, N）对 (30, 100)、(15, 200)、(100, 30) 时的结果。在$\gamma = 0.9$ 下，最优价值约为 6.89；在 $\gamma = 0.95$下，约为 13.47。置信带基于 100 次重复试验计算得出。

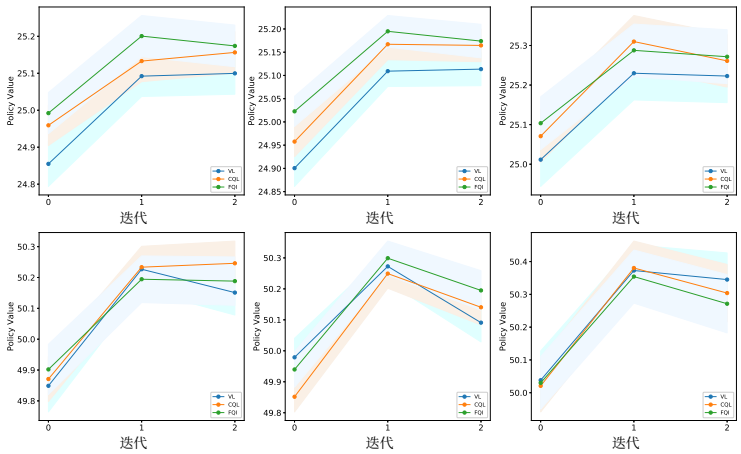

图9：采用价值学习、保守Q学习和拟合Q迭代计算初始策略后，各估计策略的价值。第一行表示使用 $\gamma = 0.9$ 的结果，第二行表示使用 $\gamma = 0.95$的结果。三列分别表示使用(T, N)参数对(50, 100)、(25, 200)、(100, 50)的结果。置信带基于100次重复试验计算。

参考文献

Abbeel, P. 和 Ng, A. Y.（2004），通过逆强化学习进行学徒学习，载于《第二十一届国际机器学习会议论文集》，第1页。

Audibert, J.-Y., Tsybakov, A. B. 等（2007），《插件分类器的快速学习率》，《统计年鉴》35(2)，608–633。

Bertsekas, D.P. 和 Tsitsiklis, J. N.（1996），《神经动态规划》，第5卷，Athena Scientific，马萨诸塞州贝尔蒙特。

Bhandari, J.、Russo, D. 和 Singal, R.（2018年），《采用线性函数逼近的时序差分学习的有限时间分析》，arXiv 预印本 arXiv:1806.02450。

Bishop, C.（1994年），《混合密度网络》，技术报告，第 1–26 页　　.

Bradley, R. C.（2005），《强混合条件的基本性质：综述与若干开放问题》，《概率调查》2，107–144。

Brown, B. M. 等（1971年），《鞅中心极限定理》，《数学统计年鉴》42(1)，59–66。

Casella, G. 和 Berger, R. L.（2002），《统计推断》，第2卷，Duxbury Pacific Grove，CA。Chakraborty, B. 和 Moodie, E.（2013），《动态治疗方案的统计方法》，施普林格。Chen, X. 和 Qi, Z.（2022），《离策略评估中非参数 q 函数估计的适定性与极小极大最优速率》，arXiv 预印本 arXiv:2201.06169。Chernozhukov, V.、Chetverikov, D.、Demirer, M.、Duflo, E.、Hansen, C.、Newey, W. 和 Robins, J.（2018），《用于治疗和结构参数的双重／去偏机器学习》。Chernozhukov, V.、Chetverikov, D.、Kato, K. 等（2014），《经验过程上确界的高斯逼近》，《统计年鉴》42(4)，1564–1597。

Dedecker, J. 和 Louhichi, S.（2002），极大不等式与经验中心极限定理，载于《相关数据的经验过程技术》，Birkh¨auser Boston，波士顿，马萨诸塞州，第 137–159 页。

Degris, T.、White, M. 和 Sutton, R. S.（2012年），离策略演员-评论家，载于《第 29 届国际机器学习会议论文集》，第 179–186 页。

Devroye, L.、Mehrabian, A. 和 Reddad, T.（2018年），《高维高斯分布之间的全变差距离》，arXiv 预印本 arXiv:1810.08693。Ernst, D.、Geurts, P.、Wehenkel, L. 和 Littman, L.（2005），《基于树的批量模式强化学习》，《机器学习研究杂志》6，503–556。Ertefaie, A. 和 Strawderman, R. L.（2018年），《在无限时间范围内构建动态治疗方案》，《生物统计学》105(4)，963–977。Fan, J.、Wang, Z.、Xie, Y. 和 Yang, Z.（2020a），《深度 q 学习的理论分析》，载于《动力学与控制学习》，PMLR，第486–489页。Fan, J.、Wang, Z.、Xie, Y. 和 Yang, Z.（2020b），《深度 q 学习的理论分析》，载于《动力学与控制学习》，PMLR，第486–489页。

Farahmand, A.-m.、Ghavamzadeh, M.、Szepesv´ari, C. 和 Mannor, S.（2016年），《基于非参数函数空间的正则化策略迭代》，《机器学习研究期刊》17(1)，4809–4874。

Feng, Y., Ren, T., Tang, Z. 和 Liu, Q.（2020年），使用核贝尔曼统计量进行可问责的离策略评估，载于《国际机器学习会议》，PMLR，第 3102–3111 页。

Harvey, N., Liaw, C. 和 Mehrabian, A.（2017），分段线性神经网络近乎紧致的 VC 维数界，载于《学习理论会议》，PMLR，第 1064–1068 页。

Hu, X., Qian, M., Cheng, B. 和 Cheung, Y. K.（2021年），《利用纵向移动健康数据进行个性化策略学习》，《美国统计学会会刊》116(533)，410–420。

Hubbs, C. D., Perez, H. D., Sarwar, O., Sahinidis, N. V., Grossmann, I. E. 和 Wassick, J. M.（2020年），《Or-gym：面向运筹学问题的强化学习库》，arXiv 预印本 arXiv:2008.06319。

Hunter, D. R. and Lange, K.（2004年），《MM 算法教程》，The American Statistician 58(1), 30–37。

Jiang, N. and Li, L.（2016年），《用于强化学习的双重稳健离策略价值评估》，载于《International Conference on Machine Learning》，第 652–661 页。

Jiang, Z., Yang, S. 和 Ding, P.（2020），《主层可忽略性下因果效应的多重稳健估计》，arXiv 预印本 arXiv:2012.01615。Kakade, S. 和 Langford, J.（2002），《近似最优近似强化学习》，载于《ICML》，第 2 卷，第 267–274 页。Kallus, N. 和 Uehara, M.（2019），《高效打破时域灾难：无限时域过程中的双重强化学习》，arXiv 预印本 arXiv:1909.05850。Kallus, N. 和 Uehara, M.（2020），《统计高效的离策略策略梯度》，载于《国际机器学习会议》，PMLR，第 5089–5100 页。

Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. 和 Faisal, A. A.（2018），《人工智能临床医生学会了在重症监护中针对脓毒症的最优治疗策略》，《自然医学》24(11)，1716–1720。Komunjer, I. 和 Vuong, Q.（2010），《条件分位数时间序列模型的半参数效率界》，《计量经济理论》第 383–405 页。Kosorok, M. R. 和 Laber, E. B.（2019），《精准医学》，《统计学及其应用年度评论》6，263–286。Kumar, A., Zhou, A., Tucker, G. 和 Levine, S.（2020），《用于离线强化学习的保守Q学习》，arXiv 预印本 arXiv:2006.04779。Laber, E. B., Lizotte, D. J., Qian, M., Pelham, W. E. 和 Murphy, S. A.（2014），《动态治疗方案：技术挑战与应用》，《电子统计学期刊》8(1)，1225。Le, H., Voloshin, C. 和 Yue, Y.（2019），《约束条件下的批量策略学习》，载于《国际机器学习会议》，第 3703–3712 页。LeCun, Y., Bengio, Y. 和 Hinton, G.（2015），《深度学习》，《自然》521(7553)，436–444。Levine, S., Kumar, A., Tucker, G. 和 Fu, J.（2020），《离线强化学习：教程、综述及开放问题展望》，arXiv 预印本 arXiv:2005.01643。Li, Y.（2017），《深度强化学习：概述》，arXiv 预印本 arXiv:1701.07274。Liao, P., Klasnja, P. 和 Murphy, S.（2019），《离策略估计长期平均结果及其在移动健康中的应用》，arXiv 预印本 arXiv:1912.13088。Liao, P., Qi, Z. 和 Murphy, S.（2020），《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。Liu, Q., Li, L., Tang, Z. 和 Zhou, D.（2018），《打破时域灾难：无限时域离策略估计》，载于《神经信息处理系统进展》，第 5356–5366 页。

Luckett, D. J., Laber, E. B., Kahkoska, A. R., Maahs, D. M., Mayer-Davis, E. 和 Kosorok, M. R.（2020年），《利用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。

Luedtke, A. R. 和Van Der Laan, M. J.（2016年），《针对可能不唯一的最优治疗策略下平均结果的统计推断》，《统计年鉴》44(2)，713。

Marcolino, M. S., Oliveira, J. A. Q., D'Agostino, M., Ribeiro, A. L., Alkmim, M. B. M. 和 Novillo-Ortiz, D.（2018年），《移动健康干预的影响：系统综述的系统综述》，《JMIR移动健康与电子健康》6(1)，e23。

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G. 等（2015年），'通过深度强化学习实现人类水平的控制'，《自然》518(7540)，529–533。

Murphy, S. A.（2003），'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。

Puterman, M. L.（1994年），《马尔可夫决策过程：离散随机动态规划》，John Wiley & Sons, Inc.

Qian, M. 和 Murphy, S. A.（2011），'个体化治疗规则的性能保证'，《统计年鉴》39(2)，1180。

Rust, J.（1987年），《GMC公交车发动机的最优更换：Harold Zurcher的经验模型》，《计量经济学：计量经济学会期刊》，第999–1033页。

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

Shi, C.、Fan, A.、Song, R. 和 Lu, W.（2018年），《用于最优动态治疗方案的高维A学习》，《统计年鉴》46(3)，第925页。

Shi, C., Lu, W. 和 Song, R.（2020年），《通过子抽样聚合打破非正则性诅咒：最优治疗策略下平均结果的推断》，已被《机器学习研究杂志》接收。

Shi, C., Wan, R., Chernozhukov, V. 和 Song, R.（2021年），深度去偏的离策略区间估计，载于《第38届国际机器学习会议论文集》，机器学习研究论文集第139卷，PMLR，第9580–9591页。

Shi, C., Wan, R., Song, R., Lu, W. 和 Leng, L.（2020年），马尔可夫决策过程是否适用于数据：检验序贯决策中的马尔可夫性质，载于《国际机器学习会议》，PMLR，第8807–8817 页。

Shi, C., Zhang, S., Lu, W. 和 Song, R.（2020年），《无限时域情形下强化学习价值函数的统计推断》，arXiv 预印本 arXiv:2001.04515。

Shi, X., Miao, W., Nelson, J. C. 和 Tchetgen Tchetgen, E. J.（2020年），《针对类别型未测量混杂的双重负对照调整与乘重稳健因果推断》，《英国皇家统计学会会刊：B辑（统计方法学）》82(2)，521–540。

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., VanDen Driessche, G., Schrit-twieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M. 等（2016年），《借助深度神经网络和树搜索掌握围棋》，《自然》529(7587)，484–489。

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

Tchetgen Tchetgen, E. J. 和 Shpitser, I.（2012年），《因果中介分析的半参数理论：效率界、多重稳健性与敏感性分析》，《统计年鉴》40(3)，1816。

Tsiatis, A.（2007年），《半参数理论与缺失数据》，施普林格科学与商业媒体。

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

Tsybakov，A. B. 等（2004年），“统计学习中的分类器最优聚合”，《统计年鉴》32(1)，135–166。

Wang，L. 和 Tchetgen Tchetgen，E.（2018年），“使用工具变量对平均处理效应进行有界、高效且多重稳健的估计”，英国皇家统计学会会刊：B辑（统计方法学）80(3)，531–550。

Wang，L.、Zhou，Y.、Song，R. 和 Sherwood，B.（2018年），《分位数最优处理方案》，《美国统计学会会刊》113(523)，1243–1254。

Watkins，C. J. 和 Dayan，P.（1992），《Q学习》，《机器学习》8(3-4)，279–292。

Wu，Y.、Tucker，G. 和 Nachum，O.（2019年），《行为正则化离线强化学习》，arXiv 预印本 arXiv:1911.11361。

Yu，T.、Thomas，G.、Yu，L.、Ermon，S.、Zou，J. Y.、Levine，S.、Finn，C. 和 Ma，T.（2020年），《Mopo：基于模型的离线策略优化》，《神经信息处理系统进展》33，14129–14142。

Zhao，Y.-Q.、Zeng，D.、Laber，E. B. 和 Kosorok，M. R.（2015年），《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。

此图“toy_est.png”提供“PNG”格式版本，来源：

http://arxiv.org/ps/2301.02220v1