arXiv:2611.32947v1 [stat.ML] 24 Nov 2026

基于自适应纵向数据的在线 V 学习

Aniruddhan Ganesaraman 统计与运筹学系 北卡罗来纳大学教堂山分校 aniruddhan_ganesaraman@unc.edu

Michael R. Kosorok 生物统计学系 北卡罗来纳大学教堂山分校 kosorok@bios.unc.edu

摘要

用于估计最优动态治疗方案的强化学习方法正越来越多地应用于移动健康（mHealth）领域，在该领域中，治疗决策会持续做出。Luckett 等人 （2020）提出了 V 学习，该方法根据离线轨迹估计最优策略。在实践中，移动健康策略会在线更新。批次更新是指在每个新数据块之后重新估计策略，它是离线学习与每次观测后更新之间一种可行的折中方案。然而，由于每个批次的治疗策略都是根据早期批次中汇集的患者数据估计得到的，因此产生的数据是自适应收集的，且患者轨迹彼此相关；离线理论无法处理这种情形。我们针对批次大小变化且确定性的这一情形，建立了 V 学习的统计理论。对于每个固定批次，随着患者数量增加，我们建立了估计价值函数的一致一致性、估计策略的一致性、策略类上的函数型中心极限定理、所学习策略的估计价值的渐近正态性，以及一致的方差估计量。就一阶性质而言，该在线程序的表现与在第一个批次之后部署最优策略的预言机一样好。当患者数量固定而批次数量增加时，在持续可辨识性条件下，估计价值函数和策略仍保持一致。在模拟中，当批次数量固定且患者数量增加时，在自适应分配下，置信区间达到接近名义水平的覆盖率，且在线程序与预言机高度接近。对于一个经过多个批次随访的小队列，所学习的策略持续稳步改进。

1 引言

移动健康（mHealth）干预措施通过可穿戴设备和智能手机提供治疗，通常在数周或数月内每天进行多次。由于决策频繁且随访期没有自然终点，这类干预措施自然可表述为无限时域序贯决策问题，而用于估计最优策略的强化学习方法

动态治疗方案正越来越多地用于实现个性化。 Luckett 等人（2020）针对这一场景提出了 V 学习。V 学习通过重要性加权贝尔曼估计方程，估计每个候选策略的状态价值函数，并在预先指定的随机策略类别中最大化估计价值，其中价值是相对于临床上关注的参考分布定义的。

Luckett 等人的理论（2020）涉及离线估计，即从在固定行为策略下收集的轨迹中学习策略。在实践中，移动健康策略会随着数据累积而更新，实际部署的算法通常以批次方式进行更新（Trella 等人，2025）。批次更新是指在每一批新的观测值收集完成后重新估计策略，在离线学习与每次决策后都进行更新之间取得了可行的折中。这也从两个方面改变了统计问题。首先，每个批次都由根据更早批次估计出的策略生成，因此数据是以自适应方式收集的。其次，由于该策略是通过汇集多个患者的数据估计得到的，患者轨迹不再相互独立。离线理论无法处理这两个特征，并且已有充分文献表明，自适应数据收集可能导致标准推断失效（Zhang 等人，2020，2022）。

我们为累积在线 V 学习建立统计理论，研究批次大小确定但可能变化的情形，并考虑基于汇总患者数据估计出的策略反复部署到同一队列时所引入的依赖性。我们保留了Luckett 等人 (2020的估计方程和策略搜索过程。我们的分析基于一个结构性观察：当候选策略的价值函数具有精确的线性表示时，无论行为策略为何，重要性加权 Bellman 估计函数的条件均值都为零。因此，部署新估计出的策略会改变后续状态的分布，却不会改变总体目标；而且，无须对状态过程作出任何平稳性假设，估计函数仍构成鞅差序列。

当批次数固定、患者人数增长时，我们证明了一致性、策略类上的函数型中心极限定理、一阶预言机等价性，以及使用一致的患者层面方差估计量对学习策略的价值进行推断。当患者人数固定而批次数增长时，我们证明，在持续可辨识性条件下，价值函数估计量和估计策略几乎必然保持一致，因此，只要随访时间足够长，即使是小队列也足以满足要求。

仅有鞅结构还不足以完成分析；其中有三个步骤需要用到超出离线数据分析范畴的论证。首先，患者历史彼此相关，因为每位患者后续的动作取决于根据所有患者先前数据估计出的策略，因此，Luckett 等人 (2020) 理论所依据的独立轨迹大数定律并不适用。我们通过对批次进行归纳证明，尽管如此，历史的经验分布仍会收敛到一个确定性极限（定理 4.1），从而确定极限协方差。其次，策略是根据数据选出的，因此，要推断其价值，需要一个对策略类一致成立的中心极限定理。我们将 Azuma–Hoeffding 不等式与链式论证相结合，为鞅得分建立了这样的定理（引理4.2）。第三，在队列固定的情况下，信息仅通过患者内部的决策时点逐步累积，而这些时点既不独立，也非平稳。我们通过针对有界鞅差的一致强大数定律，建立了几乎处处成立且对策略类一致的相合性（引理 5.1 和定理 5.2）。

本文结构安排如下。第 2 节回顾相关研究。第 3 节介绍在线 V-学习、其总体目标以及我们的假设。第 4 节和 5节分别给出批次数固定和患者人数固定时的结果。第 6 节报告模拟研究，第 7 节总结全文。证明见附录。

2 相关研究

无限时域上的策略学习。通过贝尔曼方程和线性基估计价值函数，是强化学习领域由来已久的思想，其历史可追溯至最小二乘时序差分（LSTD）方法。Lagoudakis 和 Parr (2003) 将这一思想扩展到了控制领域，提出了

最小二乘策略迭代，该方法交替进行两步：估计当前策略的状态-动作价值函数，并改进该策略。这种方法在实践中的一项优势是，可以重复利用同一批状态转移来评估每个候选策略。对于离策略 LSTD，Yu (2012) 在底层马尔可夫过程满足一般条件时，证明了几乎必然收敛；随着批次数量增加，我们还将再次讨论这一结果。

在统计学文献中，该问题主要是在移动健康的背景下进行研究的，其中治疗决策需要频繁作出，且随访期没有固定的终点。Ertefaie 和 Strawderman （2018）利用贪心梯度 Q 学习，在无限期范围内构建动态治疗方案；Liao 等人 （2022）则研究了平均奖励马尔可夫决策过程中的批量策略学习。我们分析的程序是 Luckett 等人（2020）提出的 V 学习方法。V 学习在健康应用中具有两项优势：策略类可以限定为随机策略，从而确保每个状态下每种允许的治疗都以正概率分配；此外，策略价值是相对于参考分布定义的，而该参考分布可以选择为反映所关注的临床患者群体。尽管 Luckett 等人 （2020）也提出了 V 学习的在线更新方法，但其理论结果针对的是离线估计。

自适应数据收集后的推断。越来越多的研究表明，自适应数据收集可能会使原本对独立数据有效的推断程序失效。Zhang 等人（2020）表明，在批量老虎机实验中，当治疗分配概率不集中时，通常的正态近似可能失效。Hadad 等人 （2021）构造了自适应加权估计量，以稳定自适应实验中的逆概率权重；Zhang 等人 （2021）则为基于自适应收集数据计算的适当加权 M 估计量建立了渐近正态性。

与我们工作最接近的是 Zhang 等人 （2022）的研究。他们研究了由跨个体汇总信息的算法收集的纵向数据，这正是在线 V 学习中产生依赖性的来源。他们表明，由于治疗策略本身是根据较早的数据估计得到的，标准方差估计量可能低估真实方差；为此，他们提出了一种自适应估计量，其中加入了一个额外项，用于考虑策略参数的估计。该项在他们称为策略不变性的条件下消失：即，期望估计函数在一阶意义下不依赖于策略参数。在精确线性设定下，贝尔曼估计函数在任意行为策略下的条件均值为零，因此，尽管存在自适应治疗分配，标准患者层面估计量仍保持一致性。在相关工作中，Zhang 等人 （2025）将自适应学习策略的可复现性与基于所得数据计算的估计量的一致性和渐近正态性联系起来，而 Trella 等人 （2025）则描述了在临床试验中部署此类算法的过程。Dwivedi 等人 （2025）研究了在治疗以自适应方式分配的序贯实验中，对反事实结果进行推断的问题。

其他研究关注如何利用自适应收集的数据进行学习，而不只是推断。 Zhan 等人(2024) 从自适应收集的上下文老虎机数据中学习策略，并证明了加权增广逆概率估计量的有限样本遗憾界。 Syrgkanis 和 Zhan (2026) 针对多阶段强化学习算法收集的数据，开发了基于矩的估计与推断方法，并使用自适应权重处理时变条件方差。我们的设定与两者都不同，因为估计策略会在连续马尔可夫决策过程中被反复部署。我们不引入方差稳定化权重，而是利用贝尔曼估计函数的条件均值为零性质，并证明行为策略在每个固定批次索引处都会收敛。

长时域下的离策略评估与推断。 离策略评估，即根据在不同策略下收集的数据估计某一策略的价值，在强化学习和统计学领域都受到了广泛关注。 Kallus 和 Uehara (2020, 2022) 研究马尔可夫决策过程中的高效离策略评估，包括利用边际密度比和时间齐次性来处理长时域估计问题。 Madhow 等人 (2025) 针对策略依赖于过去数据的情形，给出了表格型马尔可夫决策过程中离线策略评估和策略学习的样本复杂度结果。

在推断方面，Shi 等人(2022) 利用状态-动作价值函数的级数近似，构造策略价值的置信区间；针对数据依赖的目标策略，他们提出了一种序贯价值评估程序，当轨迹数量或决策点数量增长时，该程序仍然有效。 Wu 等人 (2026) 研究采用线性函数逼近的时序差分学习，并针对独立样本推导了Berry–Esseen界和渐近协方差的在线估计量。 Liu 等人 (2025) 针对重尾奖励和污染下的策略评估，开发了在线估计与推断方法；Lu 等人 (2026) 则研究最优价值的离线推断。这些论文将数据生成过程视为固定，并研究策略价值的估计与推断。我们还研究学习过程本身，探讨当每个估计策略都会生成下一批数据时，该过程是否收敛。

3 在线V学习

3.1 数据与估计量

我们跟踪 n 名患者，索引为 $i = 1, .., n$，且无脱落。令 $S_t^i \in \mathcal{S} \subset \mathbb{R}^p$ 为时间 t治疗前立即观察到的状态， $A_t^i \in \mathcal{A} = \{a_1, \ldots, a_K\}$ 为分配的治疗，$U_t^i = u(S_{t+1}^i, A_t^i, S_t^i)$为效用。状态根据动作依赖的转移核$P_a(s, ds')$演变，该转移核在所有患者和时间点上均相同。记

$$(P_a f)(s) = \int_{\mathcal{S}} f(s') P_a(s, ds'), \qquad r_a(s) = \int_{\mathcal{S}} u(s', a, s) P_a(s, ds'),$$

其中， f 有界且可测。我们固定折扣因子 $\gamma \in (0, 1)$，并将平稳策略 π 的状态价值函数定义为

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_{m=0}^{\infty} \gamma^m U_{t+m} \,\middle|\, S_t = s \right].$$

此处， $\mathbb{E}_\pi$ 指的是由具有转移核 P_a的 π 所生成的后续治疗决策，而不是生成观测数据的治疗概率。

正如 Luckett 等人 （2020）一样，我们考虑多项逻辑斯蒂类 $\Pi = \{\pi_\beta : \beta \in \mathcal{B}\}$，其中 $\beta = (\beta_1^\top$、……、$\beta_{K-1}^\top)^\top \in \mathcal{B} \subset \mathbb{R}^{p(K-1)}$ 且

$$\pi_\beta(a_j; s) = \frac{\exp(s^\top \beta_j)}{1 + \sum_{\ell=1}^{K-1} \exp(s^\top \beta_\ell)}, \quad j < K, \qquad \pi_\beta(a_K; s) = \frac{1}{1 + \sum_{\ell=1}^{K-1} \exp(s^\top \beta_\ell)}.$$

系数空间 $\mathcal{B}$ 假设为紧集且具有非空内部。我们将 π 的价值近似为 $V(\pi, s; \theta^\pi) = \Phi(s)^\top \theta^\pi$，其中 $\Phi : \mathcal{S} \to \mathbb{R}^q$ 是给定的特征映射，且 $\theta^\pi \in \Theta \subseteq \mathbb{R}^q$。

第一个批次是在已知策略 π^0下生成的。令 $B_k \in \mathbb{N}$ 表示批次 k的确定性长度k，并设 $T_0 = 0$、$T_k = \sum B_j$、 $N_k = nT_k$。 在时刻 t，令 $j(t)$表示满足$_{j=1}T_{j(t)-1} < t \le T_{j(t)}$的批次索引。一旦批次 j 结束，截至时刻 T_j 的累积数据便会产生一个估计策略 $\hat{}\pi_n^j = \pi_{\hat{\beta}_n^j}$。因此，在时刻 t 用于分配治疗的概率为

$$\mu_{n,t}(a;s) = \begin{cases} \pi^0(a;s), & t \le B_1, \\ \pi_{\hat{\beta}_n^{j(t)-1}}(a;s), & t > B_1. \end{cases} \tag{1}$$

策略 $\mu_{n,\ t}$ 是随机的，但由时间 t治疗分配之前收集的观测值所决定。

对于候选策略 $\pi \in \Pi$ 和 $\theta \in \Theta$，令

$$w_t^i(\pi) = \frac{\pi(A_t^i; S_t^i)}{\mu_{n,t}(A_t^i; S_t^i)}, \qquad e_t^i(\theta) = U_t^i + \gamma \Phi(S_{t+1}^i)^\top \theta - \Phi(S_t^i)^\top \theta.$$

根据重要性加权贝尔曼估计方程，定义

$$\Lambda_{n,k}(\pi, \theta) = \frac{1}{N_k} \sum_{t=1}^{T_k} \sum_{i=1}^{n} w_t^i(\pi) e_t^i(\theta) \Phi(S_t^i) = \hat{b}_{\mathrm{cum},k}(\pi) - \hat{W}_{\mathrm{cum},k}(\pi)\theta, \tag{2}$$

其中

$$\hat{W}_{\mathrm{cum},k}(\pi) = \frac{1}{N_k} \sum_{t=1}^{T_k} \sum_{i=1}^{n} w_t^i(\pi) \Phi(S_t^i) \{\Phi(S_t^i) - \gamma \Phi(S_{t+1}^i)\}^\top,$$
$$\hat{b}_{\mathrm{cum},k}(\pi) = \frac{1}{N_k} \sum_{t=1}^{T_k} \sum_{i=1}^{n} w_t^i(\pi) U_t^i \Phi(S_t^i).$$

令 Ω 为固定的正定 $q \times q$ 矩阵， $J:\ \Theta \to [0, \infty)$为惩罚项，$\lambda_{n,\ k} \ge 0$ 为确定性调优参数。我们的估计量为

$$\hat{\theta}_{\pi,n}^k = \arg\min_{\theta \in \Theta} \left\{ \Lambda_{n,k}(\pi, \theta)^\top \Omega \Lambda_{n,k}(\pi, \theta) + \lambda_{n,k} J(\theta) \right\}. \tag{3}$$

对于 $J(\theta) = \|\theta\|^2$ 和 $\Theta = \mathbb{R}^q$，这就简化为通常的岭正则化最小距离估计量。对于固定的参考分布 R，令 $\nu = \int_{\mathcal{S}} \Phi(s)\, dR(s)$并定义

$$\hat{V}_{n,R,k}(\pi) = \nu^\top \hat{\theta}_{\pi,n}^k, \qquad \hat{\beta}_n^k \in \arg\max_{\beta \in \mathcal{B}} \hat{V}_{n,R,k}(\pi_\beta), \qquad \hat{\pi}_n^k = \pi_{\hat{\beta}_n^k}. \tag{4}$$

$\hat{}$在有限样本中，矩阵 $W_{\mathrm{cum},\ k}\ (\pi_\beta)$ 对于某些 β可能是奇异的。由于每个估计策略都会生成下一个批次，因此该过程仍必须为每个事件指定一个策略。因此，我们固定默认值 $\bar{\theta} \in \Theta$ 和 $\bar{\beta} \in \mathcal{B}$。如果下确界 $_{\beta \in \mathcal{B}} \sigma_{\min}\{\ \hat{}W_{\mathrm{cum},\ k}\ (\pi_\beta)\ \} = 0$，则我们令$\hat{}\theta_{\pi_\beta,\ n}^k = \bar{\theta}$ 对所有$\beta \in \mathcal{B}$ 以及$\hat{}\beta_n^k = \bar{\beta}$均如此。否则，估计量由（3）–（4）定义。

算法1： 具有确定性批次的累积在线 V 学习

输入： n 患者；初始策略 π^0；类别 $\{\pi_\beta : \beta \in \mathcal{B}\}$；特征映射 Φ；参考向量$\nu = \int \Phi \, dR$；正定矩阵 Ω；惩罚项 J；调节序列 λ_n、k；折扣因子 γ；确定性批次大小 $\{B_k\}$。设置 $T_0 = 0$、$\hat{\pi}_n^0 = \pi^0$，对于 $k = 1$、2、.....执行：设置 $T_k = T_{k-1} + B_k$ 和 $N_k = nT_k$，对于 $t = T_{k-1} + 1$、.....、T_k 执行：对于 $i = 1$、.....、n 执行：从 $\hat{\pi}_n^{k-1}(\cdot; S_t^i)$ 中采样 A_t^i，并观测 U_t^i、S_{t+1}^i；记录 μ_n、t =；将 $\hat{\pi}_n^{k-1}$作为时间 t的行为策略；对于每个查询到的 $\beta \in \mathcal{B}$，使用（3）计算 $\hat{\theta}_{\pi_\beta}^k$、$n$；根据（4）设置 $\hat{\beta}_n^k$ 和 $\hat{\pi}_n^k$

3.2 总体目标与假设

对于每个候选策略 π 和状态 s，令 $V_R(\pi) = \int V^\pi(s) \, dR(s)$表示其在$\mathcal{S}$参考分布 R下的价值。下文假设价值函数具有精确的线性表示$V^\pi(s) = \Phi(s)^\top \theta_0^\pi$。由此得到的总体目标函数及其最大化点为

$$M(\beta) = V_R(\pi_\beta) = \nu^\top \theta_0^{\pi_\beta}, \qquad \beta^* = \arg\max_{\beta \in \mathcal{B}} M(\beta), \qquad \pi^* = \pi_{\beta^*}. \tag{5}$$

T总体目标 β^* 的定义与批次数量及各批次的长度无关 .

对于每个固定批次 k，估计矩阵的总体极限取决于时刻 $t = 1$、.....、T_k患者状态的分布。我们根据连续批次中实施的策略的大样本行为来描述这些分布。

回顾一下，初始策略 π^0 是固定且已知的，而批次 $j \geq 2$中使用的策略是根据截至批次 $j-1$收集的数据估计得到的。在假设 7下，总体目标函数具有一个共同且唯一的最大化点 β^*，其对应的策略为 $\pi^* = \pi_{\beta^*}$。正如我们将在定理 4.1中证明的，对于每个固定的 j，批次 j 结束时的估计策略收敛于 π^*，当 $n \to \infty$时。因此，在大样本极限下，第一个批次的行为策略为 π^0，而后续每个批次的行为策略均为 π^*。我们将这一行为策略的极限序列记为

$$\mu_t^\circ = \begin{cases} \pi^0, & 1 \leq t \leq B_1, \\ \pi^*, & t > B_1. \end{cases}$$

令 η_1° 表示共同的初始状态分布。对于任意策略 π，定义转移核

$$P_\pi(s, D) = \sum_{a \in \mathcal{A}} \pi(a; s) P_a(s, D).$$

极限行为策略序列下的状态分布满足

$$\eta_{t+1}^\circ = \eta_t^\circ P_{\mu_t^\circ}, \qquad t \geq 1, \tag{6}$$

其中（ηP_π）（D）$=\int P_\pi$（s，D）$d\eta$（s）。策略切换s后，这些分布仍会继续演化，并且不一定处于稳态。

利用这些分布，将批次-k 总体矩阵定义为

$$W_{\mathrm{cum},k}(\pi)=\frac{1}{T_k}\sum_{t=1}^{T_k}\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\pi(a;s)\Phi(s)\{\Phi(s)-\gamma(P_a\Phi)(s)\}^\top d\eta_t^\circ(s), \tag{7}$$

$$b_{\mathrm{cum},k}(\pi)=\frac{1}{T_k}\sum_{t=1}^{T_k}\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\pi(a;s)r_a(s)\Phi(s)\,d\eta_t^\circ(s). \tag{8}$$

在以下假设4下，这些矩阵仍然识别出相同的系数 θ_0^π 对于每个 k，如引理3.1所示。

假设 1（批次大小）。序列 $\{B_k\}_{k\geq1}$ 是确定性的，并且满足 $0<B_{\min}\leq B_k\leq B_{\max}<\infty$ 对所有 $k\geq1$ 均成立。

假设 2（处理分配与条件独立性）。初始状态 S_1^1、……、S_1^n 独立同分布，且共同分布为 η_1°。令

$$\mathcal{G}_{t-1}=\sigma\big(S_1^1,\dots,S_1^n,\,A_s^i,U_s^i,S_{s+1}^i:1\leq i\leq n,\,1\leq s<t\big)$$

表示截至时间 $t-1$ 收集的信息，包括当前状态 S_t^1、……、S_t^n。令 $\mu_{n,\,t}$ 表示在前一个批次边界选择的行为策略，其中在第一个批次中为 $\mu_{n,\,t}=\pi^0$。定义

$$\mathcal{F}_{t,0}=\mathcal{G}_{t-1}\vee\sigma(\mu_{n,1},\dots,\mu_{n,t}),$$

并且，对于 $i=1$、……、n，

$$\mathcal{F}_{t,i}=\mathcal{F}_{t,0}\vee\sigma\big(A_t^\ell,U_t^\ell,S_{t+1}^\ell:1\leq\ell\leq i\big).$$

因此，$\mathcal{F}_{t,\,0}$ 包含时间 t 分配治疗前可用的所有信息，包括先前实施的策略，并且

$$\mathcal{F}_{t,n}\subseteq\mathcal{F}_{t+1,0}.$$

C在给定 $\mathcal{F}_{t,\,0}$的条件下，数对 $(A_t^i,\;S_{t+1}^i)$、$i=1,\;\dots,\;n$ 相互独立，且满足 y

$$\mathbb{P}(A_t^i=a,S_{t+1}^i\in D\mid\mathcal{F}_{t,0})=\mu_{n,t}(a;S_t^i)P_a(S_t^i,D).$$

转移 核 P_a 不依赖于 t、i或 n。效用为 $U_t^i=u(S_{t+1}^i,\;A_t^i,\;S_t^i$).

假设3（状态空间与转移正则性）。状态空间 $\mathcal{S}\subset\mathbb{R}^p$ 是紧的；Φ 和 u 在各自的定义域上连续。对于每个 $a\in\mathcal{A}$ 以及每个有界连续函数 $f\colon\mathcal{S}\to\mathbb{R}$，函数

$$s\longmapsto(P_af)(s)=\int_{\mathcal{S}}f(s')P_a(s,ds')$$

是连续的。已知的初始策略 $\pi^0(a;s)$ 连续且严格为正，对每个动作 a和每个状态 $s\in\mathcal{S}$均成立。

假设 4（线性价值函数）。对于每个 $\pi\in\Pi$，存在 $\theta_0^\pi\in\mathrm{int}(\Theta)$，使得 $V^\pi(s)=\Phi(s)^\top\theta_0^\pi$对每个 $s\in\mathcal{S}$都成立。集合 $\Theta\subset\mathbb{R}^q$ 是闭且凸的。

假设 4 要求候选策略类上的价值函数满足精确的线性设定。这确保了对于任何可容许的行为策略，重要性加权贝尔曼估计方程的条件均值为零；这一性质将在后文证明，并贯穿整个分析过程。

假设5 （可识别性）。 矩阵 Ω 是对称正定矩阵。对于每个固定的 $k \geq 1$，存在$c_{1,\ k} > 0$使得

$$\inf_{\beta \in \mathcal{B}} \sigma_{\min}\left\{W_{\mathrm{cum},k}(\pi_\beta)\right\} \geq c_{1,k},$$

其中 $\sigma_{\min}(A)$表示 A 的最小奇异值。

假设6 （惩罚项）。 惩罚项 $J\colon \Theta \to [0,\infty)$ 是连续且凸的。在 θ_0 的一个开邻域上，它是连续可微的，且其导数在该邻域的紧子集上有界。对于每个固定的 $k \geq 1$，假设

$$\lambda_{n,k} = o(N_k^{-1/2}) \qquad \text{as } n \to \infty,$$

并且，对于每个固定的 $n \geq 1$，

$$\lambda_{n,k} \to 0 \qquad \text{as } k \to \infty.$$

假设7 （唯一总体最优解）。 函数 M 在（5）中定义，并且具有唯一的最大化点$\beta^* \in \mathcal{B}$。

注1。 假设 3、 $\mathcal{S}$ 的紧致性以及 $\mathcal{B}$的逻辑斯蒂形式意味着存在$c_0 > 0$ 使得

$$\inf_{n,t,a,s} \mu_{n,t}(a;s) \geq c_0, \qquad \inf_{t,a,s} \mu_t^{\circ}(a;s) \geq c_0$$

几乎必然。因此，重要性权重是一致有界的；由紧致性和连续性可知，奖励和特征也有界。我们不假设策略切换后状态过程是平稳的。

引理3.1 (总体方程与连续性)。在假设3–5下，对所有 $\pi \in \Pi$ 以及所有 $k \geq$ 1，

$$b_{\mathrm{cum},k}(\pi) = W_{\mathrm{cum},k}(\pi)\theta_0^{\pi}. \tag{9}$$

此外， $\beta \mapsto \theta_0^{\pi_\beta}$ 和 $M(\beta) = \nu^\top \theta_0^{\pi_\beta}$ 在 $\mathcal{B}$ 上有界且利普希茨连续 .

证明见附录 A。因此，批次大小和行为策略会改变总体矩阵，但不会改变其解：自适应性影响的是对 θ_0^{π} 和 β^* 的估计精度，而非它们的取值。

4 固定批次数下的渐近性质

我们首先研究这样一种情形：纳入了许多患者，而策略仅更新少数几次。形式化地说，我们令 $n \to \infty$，并固定批次索引 k，因此时间范围 T_k 保持不变，而观测到的患者历史数量不断增加。由于每个批次的策略都是根据汇总数据估计的，这些历史彼此相关，并且其分布会随批次而变化。

对于每位患者 i 和时间 t，令

$$H_{i,t} = (S_1^i, A_1^i, S_2^i, \ldots, A_{t-1}^i, S_t^i), \qquad \mathbb{Q}_{n,t} = \frac{1}{n}\sum_{i=1}^{n} \delta_{H_{i,t}}.$$

历史 $H_{i,\ t}$ 取值于紧空间 $\mathcal{H}_t = \mathcal{S}^t \times \mathcal{A}^{t-1}$。

设 $Q_1 = \eta_1^\circ$，并将 Q_t 递归定义为极限行为策略序列下 H_t 的分布。给定 $H_t = h$，按照 $\mu_t^\circ(\cdot;\ S_t)$ 抽取 A_t，并按照 $P_{A_t}(S_t, \cdot)$ 抽取 S_{t+1}，再附加 (A_t, S_{t+1}) 以得到 H_{t+1}。效用 $U_t = u(S_{t+1}, A_t, S_t)$ 由这些变量决定。由构造可知，S_t 在 Q_t 下的边缘分布为 η_t°。

我们记

$$d_{\mathrm{BL}}(Q, Q') = \sup_{\|f\|_\infty \le 1,\, \mathrm{Lip}(f) \le 1} \left| \int f\, dQ - \int f\, dQ' \right|$$

为紧历史空间上的有界利普希茨距离。该度量生成该空间上概率测度的弱收敛。

4.1 一致性

我们的第一个结果证明，当患者数量增加时，在批次数量固定的情况下，累积 V 学习估计量具有一致性。我们首先证明其经验分布及相应估计矩阵的收敛性，然后利用共同的总体方程得到价值函数和策略估计量的一致性。

定理 4.1 (一致性)。在假设 1–7下，对于任意固定的 $k \ge 1$，当 $n \to \infty$时，以下结论成立：

(i) 经验历史分布的收敛。对于每个 $t = 1, \ldots, T_k + 1$,

$$d_{\mathrm{BL}}(\mathbb{Q}_{n,t}, Q_t) \xrightarrow{\mathbb{P}} 0.$$

特别地，时刻 t 的经验状态分布依概率弱收敛至 η_t°。

(ii) 估计矩阵的一致收敛。

$$\begin{aligned} \sup_{\pi \in \Pi} \|\hat{W}_{\mathrm{cum},k}(\pi) - W_{\mathrm{cum},k}(\pi)\| &\xrightarrow{\mathbb{P}} 0, \\ \sup_{\pi \in \Pi} \|\hat{b}_{\mathrm{cum},k}(\pi) - b_{\mathrm{cum},k}(\pi)\| &\xrightarrow{\mathbb{P}} 0. \end{aligned} \tag{10}$$

(iii) 价值函数估计量和策略估计量的一致性。

$$\begin{aligned} \sup_{\pi \in \Pi} \|\hat{\theta}^k_{\pi,n} - \theta_0^\pi\| &\xrightarrow{\mathbb{P}} 0, \\ \hat{\beta}^k_n &\xrightarrow{\mathbb{P}} \beta^*. \end{aligned} \tag{11}$$

证明见附录 B，通过对批次进行归纳得出。结果表明，尽管行为策略和状态分布随批次而变化，累积估计量仍能一致地识别出相同的总体价值函数参数和最优策略。

4.2 渐近正态性与预言机等价性

对于 $\beta \in \mathcal{B}$ 缩写为 $\theta_\beta = \theta_0^{\pi_\beta}$ 并定义

$$\xi_t^i(\beta) = \frac{\pi_\beta(A_t^i; S_t^i)}{\mu_{n,t}(A_t^i; S_t^i)} e_t^i(\theta_\beta) \Phi(S_t^i), \qquad \mathbb{G}_{n,k}(\beta) = \frac{1}{\sqrt{N_k}} \sum_{t=1}^{T_k} \sum_{i=1}^{n} \xi_t^i(\beta). \tag{12}$$

利用滤波 $\{\mathcal{F}_{t,i}\}$ 在假设 2中定义，条件处理分配机制和贝尔曼方程给出

$$\mathbb{E}[\xi_t^i(\beta) \mid \mathcal{F}_{t,i-1}] = \Phi(S_t^i) \sum_{a\in\mathcal{A}} \pi_\beta(a; S_t^i)\{r_a(S_t^i) + \gamma(P_a\Phi)(S_t^i)^\top\theta_\beta - \Phi(S_t^i)^\top\theta_\beta\} = 0. \tag{13}$$

因此，12式中的各项均为鞅差分。
对于 β， $\beta' \in \mathcal{B}$ 定义

$$\begin{aligned} g_t(\beta, \beta'; s) = \sum_{a\in\mathcal{A}} \frac{\pi_\beta(a; s)\pi_{\beta'}(a; s)}{\mu_t^\circ(a; s)} \Phi(s)\Phi(s)^\top \\ \times \int_{\mathcal{S}} \Big\{u(s', a, s) + \gamma\Phi(s')^\top\theta_\beta - \Phi(s)^\top\theta_\beta\Big\} \\ \times \Big\{u(s', a, s) + \gamma\Phi(s')^\top\theta_{\beta'} - \Phi(s)^\top\theta_{\beta'}\Big\} P_a(s, ds'), \end{aligned}$$

并设

$$\Gamma_k(\beta, \beta') = \frac{1}{T_k}\sum_{t=1}^{T_k} \int_{\mathcal{S}} g_t(\beta, \beta'; s)\, d\eta_t^\circ(s), \qquad \Sigma_k(\beta, \beta') = W_{\mathrm{cum},k}(\pi_\beta)^{-1}\Gamma_k(\beta, \beta')W_{\mathrm{cum},k}(\pi_{\beta'})^{-\top}. \tag{14}$$

引理 4.2 （贝尔曼得分的随机等度连续性）。在假设 *1–5*下，对每个固定的$k \geq 1$ 和 $\epsilon > 0$,

$$\lim_{\delta\downarrow 0} \sup_{n\geq 1} \mathbb{P}\left\{ \sup_{\substack{\beta,\beta'\in\mathcal{B} \\ \|\beta-\beta'\|\leq\delta}} \|\mathbb{G}_{n,k}(\beta) - \mathbb{G}_{n,k}(\beta')\| > \epsilon \right\} = 0.$$

证明见附录 B。接下来，我们将一致性加强为在策略类上均匀成立的函数型中心极限定理，以便对依赖于数据的 $\hat{}\beta_n^k$进行推断。

定理 4.3 （函数型渐近正态性）。在假设 *1–7*下，对每个固定的 $k \geq 1$,

$$\left\{\sqrt{N_k}(\hat{\theta}^k_{\pi_\beta,n} - \theta_0^{\pi_\beta}) : \beta \in \mathcal{B}\right\} \rightsquigarrow \{\mathbb{Z}_k(\beta) : \beta \in \mathcal{B}\} \quad in\ \ell^\infty(\mathcal{B})^q, \tag{15}$$

其中， $\mathbb{Z}_k$ 是一个紧的均值为零的高斯过程，其协方差为

$$\mathrm{Cov}\{\mathbb{Z}_k(\beta), \mathbb{Z}_k(\beta')\} = \Sigma_k(\beta, \beta').$$

因此，过程 $\sqrt{N_k}\{\hat{}V_{n,\ R,\ \overline{k}}(\pi_\beta) - M(\beta)\}$ 在 $\ell^\infty(\mathcal{B})$ 中弱收敛于 $\nu^\top\mathbb{Z}_k(\beta)$。

证明见附录B。尽管行为策略是根据较早的批次估计得到的，Σ_k 仅通过它们的极限值 μ_t°依赖于这些策略，而它们的抽样变异不会给协方差带来额外项。这是因为对于任意行为策略，贝尔曼得分的条件均值都为零，因此估计策略中的误差不会以一阶效应传递。按照Zhang 等人（2022）的术语，该估计方程具有策略不变性，因此无需对方差进行自适应校正。

推论 4.4 （一阶预言机等价性）。固定 $k \geq 1$，并令$\hat{}\theta^{k}$、π°、n 表示基于 n 条在 μ_t°下生成的独立患者轨迹计算得到的累积 V 学习估计量；在重要性权重分母中使用该策略，并采用相同的λ_{n}、k、 J、 Φ、Ω。在假设1–7下，

$$\left\{\sqrt{N_k}(\hat{\theta}^{k,\circ}_{\pi_\beta,n} - \theta_0^{\pi_\beta}) : \beta \in \mathcal{B}\right\} \rightsquigarrow \mathbb{Z}_k$$

在 $\ell^\infty(\mathcal{B})^q$中，具有相同的高斯过程 $\mathbb{Z}_k$ 和相同的协方差 Σ_k，如定理 4.3所述。相应的价值过程也具有相同的极限。

证明见附录B。

4.3 学习策略的推断

定理 4.5 (学习策略的推断**)**。 在假设 1–7下，对于每个固定的 $k \geq 1$，令

$$\sigma^2_{V,k} = \nu^\top \Sigma_k(\beta^*, \beta^*)\nu.$$

Then

$$\sqrt{N_k}\{\hat{V}_{n,R,k}(\pi_{\hat{\beta}^k_n}) - M(\hat{\beta}^k_n)\} \xrightarrow{d} N(0, \sigma^2_{V,k}), \tag{16}$$

并且，将 $M(\hat{}\beta^k_n)$ 替换为 $M(\beta^*)$后，相同的极限仍然成立。

证明见附录 B。该定理允许采用两种中心化方式。第一种， $M(\hat{}\beta^k_n)$，是随机的：它是批次 k之后实际部署策略的价值，这是实践中关注的量，也是数据驱动的治疗方案推断中的标准目标（Chakraborty 等人，2014）。第二种， $M(\beta^*)$，是策略类中的最优价值。由于两种中心化方式产生相同的极限，因此所得区间对任一目标在渐近意义下均有效。

对于 $\beta \in \mathcal{B}$，定义估计的患者级评分

$$\hat{\psi}^k_i(\beta) = \frac{1}{\sqrt{T_k}} \sum_{t=1}^{T_k} \frac{\pi_\beta(A^i_t; S^i_t)}{\mu_{n,t}(A^i_t; S^i_t)} e^i_t(\hat{\theta}^k_{\pi_\beta,n})\Phi(S^i_t),$$

并令

$$\hat{\Gamma}_k(\beta, \beta') = \frac{1}{n}\sum_{i=1}^n \hat{\psi}^k_i(\beta)\hat{\psi}^k_i(\beta')^\top, \qquad \hat{\Sigma}_k(\beta, \beta') = \hat{W}_{\mathrm{cum},k}(\pi_\beta)^{-1}\hat{\Gamma}_k(\beta, \beta')\hat{W}_{\mathrm{cum},k}(\pi_{\beta'})^{-\top}. \tag{17}$$

逆矩阵定义在经验矩阵为非奇异的事件上，而该事件的概率趋于 1。

命题 4.6（估计量的一致性）。 在假设 1–7下，对于每个固定的 $k \geq 1$,

$$\sup_{\beta,\beta' \in \mathcal{B}} \|\hat{\Sigma}_k(\beta, \beta') - \Sigma_k(\beta, \beta')\| \xrightarrow{\mathbb{P}} 0.$$

因此，$\hat{}\sigma^2_{V,\,k} = \nu^\top \hat{}\Sigma_k(\hat{}\beta^k_n, \hat{}\beta^k_n)\,\nu \to \sigma^2_{V,\,k}$依概率收敛。若 $\sigma^2_{V,\,k} > 0$，则将式（16）中的统计量除以$\hat{}\sigma_{V,\,k}$所得结果收敛于标准正态分布。

证明见附录B。

5 固定患者数量下的渐近性

现在考虑互补情形：小队列被随访多个批次，这类似于参与者较少但长期随访的研究。形式化地，固定 n 并令 $k \to \infty$。在假设1下，$T_k \to \infty$，因此观测值的累积数量 $N_k = nT_k$ 也趋于无穷。信息现在仅通过患者内部的重复决策逐渐积累，因此一致性必须来自随访时长，而不是队列规模。

对于 $\beta \in \mathcal{B}$，定义

$$\overline{W}_{n,k}(\pi_\beta) = \frac{1}{N_k} \sum_{t=1}^{T_k} \sum_{i=1}^{n} \sum_{a \in \mathcal{A}} \pi_\beta(a; S_t^i) \Phi(S_t^i) \{\Phi(S_t^i) - \gamma(P_a \Phi)(S_t^i)\}^\top. \tag{18}$$

假设 8 （持续可辨识性）。 对于每个固定的 $n \geq 1$,

$$\liminf_{k \to \infty} \inf_{\beta \in \mathcal{B}} \sigma_{\min}\{\overline{W}_{n,k}(\pi_\beta)\} > 0 \text{ almost surely},$$

其中， $\sigma_{\min}(A)$ 表示 A的最小奇异值。

假设 8 要求实现的轨迹持续探索状态空间，这类似于自适应回归中的持续激励条件（LaiandWei, 1982）。下面的引理为经验估计矩阵和贝尔曼得分提供所需的一致强大数定律。

引理 5.1 （鞅差分的一致强大数定律）。 设 $\{X_t(\beta) : \beta \in \mathcal{B}\}$、 $t \geq 1$ 是取值于 $\mathbb{R}^d$ 的随机函数，其中 d 固定且 $\mathcal{B}$ 为紧集。假设 $X_t(\beta)$ 是$\mathcal{F}_{t+1,\,0}$-可测的，并且

$$\mathbb{E}[X_t(\beta) \mid \mathcal{F}_{t,0}] = 0$$

对每个 t 和 β 均成立。另假设存在确定性常数 C、 $L < \infty$ ，使得几乎必然地，对所有 t 和 β， $\beta' \in \mathcal{B}$,

$$\|X_t(\beta)\| \leq C, \qquad \|X_t(\beta) - X_t(\beta')\| \leq L\|\beta - \beta'\|.$$

Then

$$\sup_{\beta \in \mathcal{B}} \left\| \frac{1}{T} \sum_{t=1}^{T} X_t(\beta) \right\| \xrightarrow{\text{a.s.}} 0 \qquad \text{as } T \to \infty.$$

证明见附录C。

许多移动健康研究会长期跟踪一个小队列。下一个定理表明，只要满足持续可识别性条件，无论固定队列的规模如何，都可以实现一致性。

定理 5.2 (固定患者数下的一致性)。 假设条件 *1–8* 均成立，且在线程序在每个批次上均定义良好。那么，对每个固定的 $n \geq 1$，以下结论在 $k \to \infty$ 时成立：

(i) 经验估计矩阵满足

$$\sup_{\beta \in \mathcal{B}} \|\hat{W}_{\text{cum},k}(\pi_\beta) - \overline{W}_{n,k}(\pi_\beta)\| \xrightarrow{\text{a.s.}} 0. \tag{19}$$

（*ii*） 累积价值函数估计量是一致收敛的：

$$\sup_{\beta \in \mathcal{B}} \|\hat{\theta}^k_{\pi_\beta,n} - \theta_0^{\pi_\beta}\| \xrightarrow{\text{a.s.}} 0. \tag{20}$$

(iii) 估计的策略参数满足

$$\hat{\beta}_n^k \xrightarrow{\text{a.s.}} \beta^*. \tag{21}$$

证明见附录C。

6 模拟研究

我们通过两项实验对在线 V 学习进行模拟，每项实验包含 100 次重复实验：第一项固定批次数量并增加患者数量，以评估价值估计和推断；第二项固定患者数量，并随着批次累积跟踪估计策略。

6.1 模拟设置

我们采用与Luckett 等人(2020)类似的生成模型。令 $S_t = (S_{t,1}, S_{t,2})^\top$ 表示状态，$A_t \in \{0,1\}$ 表示时间 t的治疗。省略患者索引后，我们生成

$$\begin{aligned} X_{t+1,1} &= 0.75(2A_t - 1)S_{t,1} + 0.25S_{t,1}S_{t,2} + \epsilon_{t,1}, \\ X_{t+1,2} &= 0.75(1 - 2A_t)S_{t,2} + 0.25S_{t,1}S_{t,2} + \epsilon_{t,2}, \end{aligned} \tag{22}$$

其中，$\epsilon_{t,j}$是患者、时间和坐标之间相互独立的 $N(0, 1/4)$ 随机变量，并设定

$$S_{t+1,j} = \max\{-6, \min(6, X_{t+1,j})\}$$

对于 $j = 1, 2$，以防止数值不稳定性。效用

$$U_t = 2S_{t+1,1} + S_{t+1,2} - 0.25(2A_t - 1)$$

会奖励更大的状态值，并对治疗施加成本。初始状态相互独立且服从标准正态分布，并在均匀治疗分配下经过 50 个时间点的预热期后记录观测值。

我们取 $\gamma = 0.9$和随机策略

$$\pi_\beta(1; s) = \exp(s^\top\beta)/\{1 + \exp(s^\top\beta)\}, \qquad \beta \in [-5, 5]^2.$$

按照 Luckett 等人（2020）的做法，我们使用高斯径向基函数$\phi(z; c, \tau) = \exp\{-(z - c)^2/(2\tau^2)\}$对状态价值函数进行建模，其中 $\tau = 0.25$和 $c \in \{0, 0.25, 0.5, 0.75, 1\}$分别应用于每个状态坐标。令 $\{s_j^R\}_{j=1}^{m_R}$ 为包含 $m_R = 5{,}000$ 个状态的参考样本。首先，使用该样本的第一个和第九十九个百分位数，将每个坐标映射到 $[0,1]$；超出此~区间的值将被截断。再加上一个截距，即得到一个包含 11 个函数的基函数 Φ。

~由于每个坐标对应的五个高斯函数在 $[0,1]$ 上的和几乎为常数，Φ 与截距几乎共线：其格拉姆矩阵 $G_R = m_R^{-1}\sum_{j=1}^{m_R}\widetilde{\Phi}(s_j^R)\widetilde{\Phi}(s_j^R)^\top$ 在参考样本上的条件数约为 10^5。因此，我们使用白化基函数 $\Phi(s) = G_R^{-1/2}\widetilde{\Phi}(s)$ (Kessy 等人，2018年），并在各批次和重复实验中保持不变。白化保留了基函数的线性张成空间，并改善了数值条件性。然而，岭惩罚是在白化后的

坐标中施加的，并且对这种重新参数化并不具有不变性（Golub 和 Van Loan，2013年）。我们取 $\Omega = I_{11}$, $J(\theta) = \|\theta\|^2$ 和 $\lambda_{n,\,k} = 0.01/N_k$。
第一个批次中，每位患者有 12 个决策时点，采用 $\pi^0(1; s) = 1/2$，之后每个批次有 6 个决策时点，因此 $T_k = 6k + 6$ for $k \geq 1$。每个批次结束后，利用所有可用数据重新估计策略，并在下一个批次中部署。整个过程中持续随访同一批患者，策略改变时不重置其状态。

与 Luckett 等人 (2020一样，我们加入一个 L_2 策略系数惩罚项，因此在批次 k 估计策略最大化

$$\hat{M}_{n,k}(\beta) = \hat{V}_{n,R,k}(\pi_\beta) - \rho\|\beta\|^2$$

其中 $\rho = 0.5$。该惩罚项可使最优解远离参数空间的边界；在边界处，策略几乎是确定性的，重要性权重也会变得极端。由于该惩罚项仅用于模拟，目标是 $M_\rho(\beta) = V_R(\pi_\beta) - \rho\|\beta\|^2$ 的最大化点，而非 β^* 中的（5）。

我们通过最大化 $\hat{M}_{n,\,k}$ 来计算 $\hat{\beta}_n^k$，使用 L-BFGS-B 算法在 $[-5, 5]^2$ 上进行优化，该算法由 R 中的 optim 函数实现。初始值包括零向量、从 $[-5, 5]^2$ 中均匀抽取的八个点，以及第一个批次之后的前一次估计值 $\hat{\beta}_n^{k-1}$。若候选策略的经验估计矩阵为奇异矩阵或条件数大于 200，则予以排除；如果找不到符合条件的解，则保留前一次的策略。

参考分布 R 是上述初始化程序生成的 $m_R = 5{,}000$ 个参考状态的经验分布。该样本在各次重复实验中保持固定，也用于计算 ν。我们通过对从参考状态出发、长度为 120 的 5,000 条轨迹上的折扣效用取平均，来近似 $V_R(\pi_\beta)$，并将所得近似值记为 $\tilde{V}_R(\pi_\beta)$ 和 $\tilde{M}_\rho(\beta) = \tilde{V}_R(\pi_\beta) - \rho\|\beta\|^2$。通过依次采用间距为 0.5、0.1 和 0.025 的网格搜索，确定 M_ρ 的最大化点。随后在独立样本上重新评估排名最高的 25 个候选策略，选定策略的价值根据单独的蒙特卡洛样本报告，并在比较策略时使用共同随机数。

所得基准为 $\tilde{\beta}_\rho^* = (1.550, -0.925)^\top$，其中 $\tilde{V}_R(\pi_{\tilde{\beta}_\rho^*}) = 5.941$（蒙特卡洛标准误 0.061），且 $M_\rho(\beta_\rho^*) = 4.312$。

固定批次数量。我们取 $n \in \{25, 100, 250\}$，并在批次$k \in \{1, 3, 5\}$后评估估计量。作为比较，我们考虑一项预言机实验：其中第一批次是在 π^0 下生成的，之后所有批次均在基准策略 $\pi_{\tilde{\beta}_\rho^*}$ 下生成。在两项实验中，每批次结束后都会重新估计策略，但在预言机方案中，处理分配仍遵循基准策略。

令 $\hat{\pi}_n^k = \pi_{\hat{\beta}_n^k}$。我们考察估计误差 $\hat{V}_{n,R,k}(\hat{\pi}_n^k) - \tilde{V}_R(\hat{\pi}_n^k)$ 以及名义95%置信区间 $\hat{V}_{n,R,k}(\hat{\pi}_n^k) \pm 1.96\,\hat{\sigma}_{V,k}/\sqrt{N_k}$，其中 $\hat{\sigma}^2_{V,k}$ 是方差估计量。表1报告了100次重复实验中估计误差的经验偏差和标准差（SD）、平均估计标准误（SE）以及覆盖概率（CP）。覆盖率是相对于 $\tilde{V}_R(\hat{\pi}_n^k)$评估的，其蒙特卡洛误差未反映在估计标准误中。

在两种方案下，随着 n 和 k 增大，标准差均有所下降；估计的标准误与标准差接近，但在最小的设置下例外，这些设置中也出现了最大的偏差。自适应抽样的覆盖率为 92% 至 96%，预言机抽样的覆盖率为 91% 至 98%；考虑到蒙特卡洛标准误约为 2.2 个百分点，这与名义覆盖率一致。

在批次 5 中，随着 n 增大，两种方案下的标准差变得越来越相似，这与一阶预言机等价性一致。对于 $n = 250$，自适应偏差保持在 0.030 和

		自适应				预言机			
n	k	Bias	SD	SE	CP（%）	Bias	SD	SE	CP (%)
25	1	0.481	2.253	1.646	92	0.937	2.936	2.070	96
	3	0.059	0.785	0.758	94	0.225	0.830	0.787	95
	5	0.038	0.575	0.589	96	0.139	0.663	0.615	91
100	1	-0.020	0.640	0.608	94	0.050	0.689	0.619	91
	3	0.011	0.400	0.384	92	0.052	0.386	0.396	98
	5	-0.003	0.298	0.304	96	0.049	0.306	0.310	97
250	1	0.030	0.388	0.382	95	0.031	0.409	0.386	95
	3	0.062	0.284	0.245	93	0.049	0.255	0.248	97
	5	0.058	0.201	0.194	94	0.046	0.203	0.196	98

表1：在自适应和预言机治疗分配下对估计策略价值的估计：基于每种设定下100次重复实验，报告估计误差的经验偏差和标准差（SD）、平均估计标准误（SE），以及名义95%置信区间的覆盖概率（CP）。

0.062 之间；相较于标准差而言，它们较小，可能反映了基函数逼近误差，以及有限样本和蒙特卡洛效应。

固定患者数量。接下来，我们取 $n \in \{2, 5, 10\}$，并对每位患者进行跟踪，历经40个批次，即246个决策时点。每个批次结束后，我们记录惩罚价值损失（$L_{n,\,k}$）、策略参数偏离超过10%的情况（$I_{n,\,k}$），以及估计价值的误差（$D_{n,\,k}$），其定义如下：

$$
\begin{aligned}
L_{n,k} &= \tilde{M}_\rho(\tilde{\beta}^*_\rho) - \tilde{M}_\rho(\hat{\beta}^k_n), \\
I_{n,k} &= \mathbb{I}\left\{ \|\hat{\beta}^k_n - \tilde{\beta}^*_\rho\| > 0.1 \|\tilde{\beta}^*_\rho\| \right\}, \\
D_{n,k} &= \left| \hat{V}_{n,R,k}(\hat{\pi}^k_n) - \tilde{V}_R(\hat{\pi}^k_n) \right|.
\end{aligned} \tag{23}
$$

图1汇总了各重复实验中的这些量。

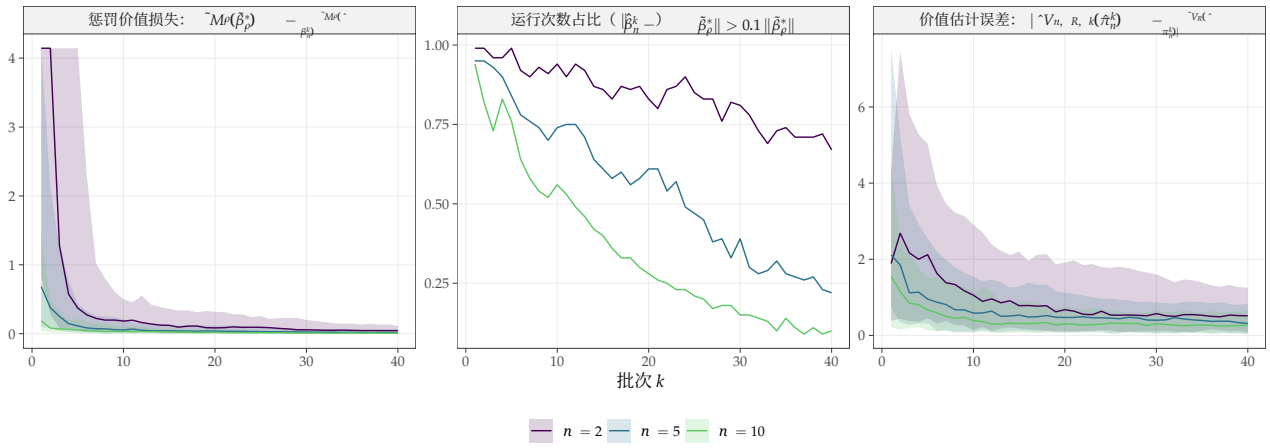

图1：在线 V 学习，患者数为 $n = 2$、5、10，历经40个批次。(a) 惩罚价值损失$L_{n,\,k}$ 和 (c) 估计价值误差 $D_{n,\,k}$：各重复实验中的中位数（线）和第10至第90百分位数（带状区域）。(b) 出现$I_{n,\,k} = 1$的重复实验比例。

图 1 显示，延长队列的随访时间可以改进估计策略。面板（a）所示的惩罚价值损失中位数随 k 增加而下降，对三种队列规模均如此，降至0。045、0。015

在 $n = 2$ 为 5 和 10 时，第 40 批次分别为 0。008。面板（b）所示的参数误差超过 10% 的重复实验比例也有所下降，分别降至 67%、22% 和 10%；不过，对于 $n = 2$，下降较为缓慢。在这种情况下，即使参数误差仍然较大，损失也很小，因此，接近最优的价值并不要求对 β 作出准确估计。面板（c）所示的估计价值误差中位数则降至 0。516、0。315 和 0。266。由于基准是通过网格搜索和蒙特卡洛模拟计算得出的，损失可能为负；解读较小的损失时应考虑其精度。

图中未展示另外两项结果。在批次 1 中，对于 $n = 2$、5 和 10，估计矩阵的条件数过高，无法得到有效的价值估计；这类重复实验分别占 80%、17% 和 4%。这些重复实验仅从面板 (c) 的汇总结果中排除；对于 $n = 2$，回退规则是面板 (a) 中初期急剧下降的部分原因。到批次 40 时，所有重复实验均得到了有效估计，中位条件数介于 10.89 和 12.15 之间。

7 讨论

我们针对固定队列纵向数据下、经过重复策略更新的累积在线 V 学习，建立了渐近理论。对于固定数量的批次，价值函数估计量一致收敛，并满足函数型中心极限定理，其一阶极限分布与预言机程序相同。对学习策略价值的推断可基于患者层面方差估计量进行。在患者数量固定的情况下，只要满足持续可辨识性，即使批次数量增加，一致性仍然成立。

一个核心局限是，候选策略价值函数采用了精确的线性设定，这确保总体目标在自适应行为策略下保持不变。理解近似误差对这一性质及后续推断的影响，是未来研究的重要方向。还有若干值得进一步探索的方向。在固定队列情形下，我们证明了一致性，但没有得到极限分布；要在这种情形下进行推断，需要满足一些条件，使状态过程和行为策略随时间趋于稳定，并且还需要收敛速率。若患者人数和批次数量同时增长，则可建立联合渐近理论，从而衔接这两种情形，并阐明随访时长与队列规模之间的权衡。将结果推广到数据依赖的批次长度也是一个研究方向，尤其是因为自适应更新计划在实践中很常见。最后，我们的结果假设最优策略唯一，且分配概率有正的下界。当最优策略近乎确定时，有限样本下的重要性权重可能变得很大；而当最优解不唯一时，估计的最优价值通常不再适用高斯极限。这两种情况在实践中都很自然地出现，值得进一步研究。

第3节的证明

引理 *3.1* 的证明。固定 $\pi \in \Pi$。由于 $\Phi(s)^\top \theta_0^\pi$ 是 π的价值，因此贝尔曼方程对每个$s \in \mathcal{S}$,

$$\Phi(s)^\top \theta_0^\pi = \sum_{a \in \mathcal{A}} \pi(a; s)\{r_a(s) + \gamma(P_a\Phi)(s)^\top \theta_0^\pi\}.$$

两边乘以 Φ(*s*)并重新整理，得到

$$\sum_{a \in \mathcal{A}} \pi(a; s) r_a(s)\Phi(s) = \left[\sum_{a \in \mathcal{A}} \pi(a; s)\Phi(s)\{\Phi(s) - \gamma(P_a\Phi)(s)\}^\top\right] \theta_0^\pi.$$

对 η_t°积分，并对 $t = 1, \ldots, T_k$取平均，得到

$$\begin{aligned} b_{\mathrm{cum},k}(\pi) &= \frac{1}{T_k}\sum_{t=1}^{T_k}\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\pi(a;s)r_a(s)\Phi(s)\,d\eta_t^\circ(s) \\ &= \left[\frac{1}{T_k}\sum_{t=1}^{T_k}\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\pi(a;s)\Phi(s)\{\Phi(s)-\gamma(P_a\Phi)(s)\}^\top\,d\eta_t^\circ(s)\right]\theta_0^\pi \\ &= W_{\mathrm{cum},k}(\pi)\theta_0^\pi, \end{aligned}$$

这证明了（9）。

对于 $\beta \in \mathcal{B}$，令 $W_\beta = W_{\mathrm{cum},\ 1}(\pi_\beta)$ 以及 $b_\beta = b_{\mathrm{cum},\ 1}(\pi_\beta)$。假设5 蕴含

$$\sup_{\beta\in\mathcal{B}}\|W_\beta^{-1}\| \le c_{1,1}^{-1}.$$

由紧致性、u 和 Φ的有界性，以及（8），有 $C_b := \text{上确界}_\beta \|b_\beta\| < \infty$。因此，由（9），

$$\sup_{\beta\in\mathcal{B}}\|\theta_0^{\pi_\beta}\| = \sup_{\beta}\|W_\beta^{-1}b_\beta\| \le c_{1,1}^{-1}C_b.$$

由于 $\mathcal{S}$ 和 $\mathcal{B}$ 是紧集，逻辑斯蒂概率关于 β 的导数一致有界。因此，存在某个 $L_\pi < \infty$,

$$\sup_{s,a}|\pi_\beta(a;s) - \pi_{\beta'}(a;s)| \le L_\pi\|\beta - \beta'\|.$$

W_β、b_β的定义由此给出有限常数 L_W、L_b，使得

$$\|W_\beta - W_{\beta'}\| \le L_W\|\beta-\beta'\|, \qquad \|b_\beta - b_{\beta'}\| \le L_b\|\beta-\beta'\|.$$

利用 $W_\beta^{-1} - W_{\beta'}^{-1} = W_\beta^{-1}(W_{\beta'} - W_\beta)W_{\beta'}^{-1}$ 可得

$$\begin{aligned} \|\theta_0^{\pi_\beta} - \theta_0^{\pi_{\beta'}}\| &\le \|W_\beta^{-1}(b_\beta - b_{\beta'})\| + \|(W_\beta^{-1} - W_{\beta'}^{-1})b_{\beta'}\| \\ &\le (c_{1,1}^{-1}L_b + c_{1,1}^{-2}C_bL_W)\|\beta-\beta'\|. \end{aligned}$$

由于固定了 ν，因此对于 $M(\beta) = \nu^\top\theta_0^{\pi_\beta}$，同样的结论成立。

■

第 B 节：第4节的证明

定理 4.1的证明。 我们通过对批次索引 j进行归纳，联合证明 (i)–(iii)。

由于$H_{i,\ 1} = S_1^i$、$i = 1$、.....、n服从分布 Q_1且独立同分布，根据大数定律以及对有界-利普希茨单位球的有限网逼近（如下方归纳步骤中所使用的），可得

$$d_{\mathrm{BL}}(\mathbb{Q}_{n,1}, Q_1) \xrightarrow{\text{a.s.}} 0.$$

假设在批次 j 开始时，

$$d_{\mathrm{BL}}(\mathbb{Q}_{n,T_{j-1}+1}, Q_{T_{j-1}+1}) \xrightarrow{\mathbb{P}} 0,$$

并且，如果 $j > 1$，则 $\hat{\beta}_n^{j-1} \xrightarrow{\mathbb{P}} \beta^*$。对于 $j = 1$，行为策略为 $\pi^0 = \mu_t^\circ$。对于 $j > 1$，逻辑斯蒂策略类的一致利普希茨连续性给出

$$\sup_{T_{j-1}<t\le T_j} \sup_{a,s} |\mu_{n,t}(a;s) - \mu_t^\circ(a;s)| \xrightarrow{\mathbb{P}} 0. \tag{24}$$

我们在批次 j中证明每一项断言。

(i) 固定 $T_{j-1} < t \le T_j$，并假设 $d_{\mathrm{BL}}(\mathbb{Q}_{n,\,t},\ Q_t) \xrightarrow{\mathbb{P}} 0$。对于 $f \in C(\mathcal{H}_{t+1})$，定义

$$(K_\mu f)(h) = \sum_{a\in\mathcal{A}} \mu(a;s_t) \int_{\mathcal{S}} f(h,a,s') P_a(s_t, ds'),$$

其中 $h = (s_1, a_1, \ldots, s_t) \in \mathcal{H}_t$。

我们证明 $K_{\mu_t^\circ} f \in C(\mathcal{H}_t)$。令 $h_m \to h$ 属于 $\mathcal{H}_t$，并分别将它们当前的状态记为s_m和 s，从而 $s_m \to s$。对于每个 $a \in \mathcal{A}$，记$g_{m,\,a}(s') = f(h_m,\ a,\ s')$，以及 $g_a(s') = f(h,\ a,\ s')$。由 f和 u在相应紧空间上的连续性，

$$\|g_{m,a} - g_a\|_\infty \to 0.$$

因此，

$$\begin{aligned}\left| \int g_{m,a}(s') P_a(s_m, ds') - \int g_a(s') P_a(s, ds') \right| \\ \le \|g_{m,a} - g_a\|_\infty + |(P_a g_a)(s_m) - (P_a g_a)(s)| \longrightarrow 0,\end{aligned}$$

其中第二项根据假设 3消失，因为 $g_a \in C(\mathcal{S})$。由 $\mu_t^\circ(a;\ \cdot)$ 的连续性以及 $\mathcal{A}$ 的有限性，现在可推出 $K_{\mu_t^\circ} f(h_m) \to K_{\mu_t^\circ} f(h)$，如所要求。

此外，

$$\|K_{\mu_{n,t}} f - K_{\mu_t^\circ} f\|_\infty \le |\mathcal{A}| \|f\|_\infty \sup_{a,s} |\mu_{n,t}(a;s) - \mu_t^\circ(a;s)| = o_{\mathbb{P}}(1).$$

条件于 $\mathcal{F}_{t,\,0}$，变量 $f(H_{i,\,t+1})$ 相互独立，其条件均值为 $(K_{\mu_{n,\,t}} f)(H_{i,\,t})$。因此，

$$\mathbb{E}[\mathbb{Q}_{n,t+1} f \mid \mathcal{F}_{t,0}] = \mathbb{Q}_{n,t} K_{\mu_{n,t}} f, \qquad \mathrm{Var}(\mathbb{Q}_{n,t+1} f \mid \mathcal{F}_{t,0}) \le \frac{\|f\|_\infty^2}{n}.$$

利用 $Q_{t+1} f = Q_t K_{\mu_t^\circ} f$，我们得到

$$\begin{aligned}\mathbb{Q}_{n,t+1} f - Q_{t+1} f = \{\mathbb{Q}_{n,t+1} f - \mathbb{Q}_{n,t} K_{\mu_{n,t}} f\} \\ + \mathbb{Q}_{n,t}(K_{\mu_{n,t}} - K_{\mu_t^\circ}) f \\ + (\mathbb{Q}_{n,t} - Q_t) K_{\mu_t^\circ} f = o_{\mathbb{P}}(1).\end{aligned}$$

第一项根据条件方差界收敛于零，第二项根据（24）收敛于零，第三项则根据归纳假设以及 $K_{\mu_t^\circ} f$ 的连续性收敛于零。

设 $\mathcal{N}_\epsilon$ 是 $\mathcal{H}_{t+1}$ 上有界-利普希茨单位球在上确界范数下的一个有限 ϵ-网。则

$$d_{\mathrm{BL}}(\mathbb{Q}_{n,t+1}, Q_{t+1}) \le \max_{f\in\mathcal{N}_\epsilon} |(\mathbb{Q}_{n,t+1} - Q_{t+1}) f| + 2\epsilon.$$

有限最大值为 $o_{\mathbb{P}}(1)$，而 ϵ 是任意的。遍历批次 j 中的有限个观测时间，即可证明截至时间 $T_j + 1$的 (i)。

(ii) 根据归纳假设和（24），

$$\max_{1\le t\le T_j}\sup_{a,s}|\mu_{n,t}(a;s)-\mu_t^\circ(a;s)|=o_{\mathbb{P}}(1).$$

令 $c_0>0$为　　行为概率的一致下界。对于任意 $a\in\mathcal{A}$ 和 $s\in\mathcal{S}$,

$$\left|\frac{\pi_\beta(a;s)}{\mu_{n,t}(a;s)}-\frac{\pi_\beta(a;s)}{\mu_t^\circ(a;s)}\right|\le c_0^{-2}|\mu_{n,t}(a;s)-\mu_t^\circ(a;s)|.$$

由于 Φ 和 u 有界，因此（24）意味着，在经验矩阵中用 μ_t° 替换$\mu_{n,t}$ 会引入一个关于 $\beta\in\mathcal{B}$ 一致的 $o_{\mathbb{P}}(1)$ 误差。
为证明所得矩阵的收敛性，考虑其中任意一个坐标。它可以写成 $\mathbb{Q}_{n,T_j+1}F_\beta$，其中，对于完整历史 $h=(s_1,a_1,s_2,\dots,{}^{a_{T_j}},{}^{s_{T_j+1}})$,

$$F_\beta(h)=\frac{1}{T_j}\sum_{t=1}^{T_j}\frac{\pi_\beta(a_t;s_t)}{\mu_t^\circ(a_t;s_t)}G_t(h).$$

这里，$G_t(h)$表示 $\Phi(s_t)\{\Phi(s_t)-\gamma\Phi(s_{t+1})\}^\top$ 或 $u(s_{t+1},a_t,s_t)\Phi(s_t)$中相应的坐标，并且独立于 β。由于特征、效用和行为概率有界且连续，F_β 在 $\mathcal{H}_{T_j+1}$上有界且连续。

此外，存在 $C_G<\infty$ 使得 $\sup_{t,\ h}|G_t(h)|\le C_G$。由于逻辑斯蒂策略类满足

$$\sup_{a,s}|\pi_\beta(a;s)-\pi_{\beta'}(a;s)|\le L_\pi\|\beta-\beta'\|,$$

一致正性给出

$$\sup_{h\in\mathcal{H}_{T_j+1}}|F_\beta(h)-F_{\beta'}(h)|\le\frac{C_GL_\pi}{c_0}\|\beta-\beta'\|.$$

因此，所需的利普希茨界在 $L_F=C_GL_\pi/c_0$下成立。
对于 $\mathcal{B}$的有限 δ-网 $\mathcal{B}_\delta$，

$$\sup_{\beta\in\mathcal{B}}|(\mathbb{Q}_{n,T_j+1}-Q_{T_j+1})F_\beta|\le\max_{\beta\in\mathcal{B}_\delta}|(\mathbb{Q}_{n,T_j+1}-Q_{T_j+1})F_\beta|+2L_F\delta.$$

根据 (i)，该最大值依概率收敛至零；令 $\delta\downarrow0$ 即可得到一致收敛。

还需确定极限矩阵。在 Q_{T_j+1} 条件下，S_t 的边缘分布为 η_t°；条件于 $S_t=s$，治疗 A_t 服从分布 $\mu_t^\circ(\cdot;s)$。因此，

$$\begin{aligned}&\mathbb{E}_{Q_{T_j+1}}\left[\frac{\pi_\beta(A_t;S_t)}{\mu_t^\circ(A_t;S_t)}\Phi(S_t)\{\Phi(S_t)-\gamma\Phi(S_{t+1})\}^\top\right]\\&=\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\pi_\beta(a;s)\Phi(s)\{\Phi(s)-\gamma(P_a\Phi)(s)\}^\top\,d\eta_t^\circ(s),\end{aligned}$$

在给定 S_t 的条件下，行为概率相互抵消。同理，

$$\mathbb{E}_{Q_{T_j+1}}\left[\frac{\pi_\beta(A_t;S_t)}{\mu_t^\circ(A_t;S_t)}U_t\Phi(S_t)\right]=\int_{\mathcal{S}}\sum_{a\in\mathcal{A}}\pi_\beta(a;s)r_a(s)\Phi(s)\,d\eta_t^\circ(s).$$

对 $t=1,$ ……, T_j取平均，分别得到（7）和（8）。结合上述一致收敛结果，即可证明批次 j 中的（10）。

（iii）根据假设5 和（ii），当概率趋于 1 时，

$$\sup_{\beta\in\mathcal{B}} \|\hat{W}_{\mathrm{cum},j}(\pi_\beta)^{-1}\| \le 2c_{1,j}^{-1}.$$

在此事件下，(3) 中的惩罚准则在ˆ 闭凸集 Θ 上连续、严格凸且强制的。事实上，$b_{\mathrm{cum},j}$ 的一致可逆性和有界性意味着，其二次部分在 β 中随着 $\|\theta\| \to \infty$ 而一致趋于无穷大。由于 J 非负且凸，对每个 β，该准则都有唯一极小值点。一致强制性、准则关于 (β, θ) 的连续性以及唯一性也意味着 $\beta \mapsto \hat{}\theta^j_{\pi_\beta,n}$ 连续。

记作 $\theta_\beta = \theta_0^{\pi_\beta}$，$\hat{}\theta_\beta = \hat{}\theta^j_{\pi_\beta,n}$，以及 $\Lambda_\beta(\theta) = \Lambda_{n,j}(\pi_\beta, \theta)$。引理 3.1 和 (ii) 给出

$$\sup_{\beta\in\mathcal{B}} \|\Lambda_\beta(\theta_\beta)\| = o_{\mathbb{P}}(1).$$

集合 $\{\theta_\beta : \beta \in \mathcal{B}\}$ 根据假设 4 是紧集且包含于 Θ；因此 上确界$_\beta$ $J(\theta_\beta) < \infty$。ˆ根据 θ_β的极小化性质，

$$\begin{aligned}\lambda_{\min}(\Omega)\|\Lambda_\beta(\hat{\theta}_\beta)\|^2 &\le \Lambda_\beta(\hat{\theta}_\beta)^\top \Omega \Lambda_\beta(\hat{\theta}_\beta) + \lambda_{n,j} J(\hat{\theta}_\beta) \\ &\le \lambda_{\max}(\Omega)\|\Lambda_\beta(\theta_\beta)\|^2 + \lambda_{n,j} J(\theta_\beta).\end{aligned}$$

由于$\lambda_{n,\ j} \to 0$,

$$\sup_{\beta\in\mathcal{B}} \|\Lambda_\beta(\hat{\theta}_\beta)\| = o_{\mathbb{P}}(1).$$

利用 Λ_β的线性,

$$\hat{W}_{\mathrm{cum},j}(\pi_\beta)(\hat{\theta}_\beta - \theta_\beta) = \Lambda_\beta(\theta_\beta) - \Lambda_\beta(\hat{\theta}_\beta).$$

因此，一致逆界给出

$$\sup_{\beta\in\mathcal{B}} \|\hat{\theta}_\beta - \theta_\beta\| = o_{\mathbb{P}}(1).$$

由此可得

$$\sup_{\beta\in\mathcal{B}} |\hat{V}_{n,R,j}(\pi_\beta) - M(\beta)| = o_{\mathbb{P}}(1).$$

根据引理3.1， M 在紧集 $\mathcal{B}$上连续。假设7 意味着，对于每个$\epsilon > 0$ 只要约束集非空,

$$\inf_{\beta\in\mathcal{B}:\,\|\beta-\beta^*\|\ge\epsilon} \{M(\beta^*) - M(\beta)\} > 0.$$

现在，argmax定理（van der Vaart，1998，定理5.7）给出ˆ$\beta^j_n \xrightarrow{\mathbb{P}} \beta^*$。

这就为批次 $j+1$ 建立了归纳假设，并完成了证明。

■

引理 4.2 的证明。 固定一个坐标 $r \in \{1, \dots, q\}$。根据引理 3.1，映射 $\beta \mapsto \theta_0^{\pi_\beta}$ 有界且满足利普希茨条件。逻辑斯蒂概率关于 (a, s)一致满足利普希茨条件，而假设 3–5 保证奖励、特征和重要性权重有界。因此，存在 $L < \infty$，其取值独立于n， t， i以及观测到的数据，使得

$$|\xi_t^i(\beta)_r - \xi_t^i(\beta')_r| \le L\|\beta - \beta'\|.$$

根据（13），每个差值都是鞅差。将条件 Hoeffding 界（Hoeffding，1963年；Azuma，1967年）依次应用于 N_k 归一化增量，可得

$$\mathbb{P}\left\{|\mathbb{G}_{n,k}(\beta)_r - \mathbb{G}_{n,k}(\beta')_r| > x\right\} \le 2\exp\left\{-\frac{x^2}{2L^2\|\beta - \beta'\|^2}\right\}, \qquad x > 0. \tag{25}$$

因此，过程$\mathbb{G}_{n,\,k}(\cdot)_{\,r}$关于半度量 $d(\beta,\ \beta') = L\|\beta - \beta'\|$是次高斯的。其样本路径连续，因而是可分的，因为 $\mathcal{B}$ 是紧的。将 van der Vaart 和 Wellner 的推论 2.2.8（van der Vaart andWellner （1996年），应用于（25），得到

$$\mathbb{E}\sup_{d(\beta,\beta')\le L\delta}|\mathbb{G}_{n,k}(\beta)_r - \mathbb{G}_{n,k}(\beta')_r| \le C\int_0^{L\delta}\sqrt{\log N(\epsilon, \mathcal{B}, d)}\,d\epsilon.$$

记 $d_\beta = p(K-1)$， $\mathcal{B}$的紧致性给出 $N(\epsilon,\ \mathcal{B},\ d) \le (C_0/\epsilon)^{d_\beta}$对于足够小的 ϵ成立。因此，对于每个坐标 r，

$$\mathbb{E}\sup_{\substack{\beta,\beta'\in\mathcal{B}\\ \|\beta-\beta'\|\le\delta}}|\mathbb{G}_{n,k}(\beta)_r - \mathbb{G}_{n,k}(\beta')_r| \le C\int_0^{L\delta}\sqrt{d_\beta\log(C_0/\epsilon)}\,d\epsilon \le C_1\delta\sqrt{\log(C_2/\delta)},$$

其中常数与 n 无关。最后一个不等式由变量替换$\epsilon = L\delta u$ 以及 $\sqrt{\log(1/u)}$ 在（0，1）上的可积性得出。由于 q 是固定的且 $\|x\| \le \sum_{r=1}^{q}|x_r|$，由 Markov 不等式可知，对任意 $\varepsilon > 0$，

$$\sup_{n\ge1}\mathbb{P}\left(\sup_{\substack{\beta,\beta'\in\mathcal{B}\\ \|\beta-\beta'\|\le\delta}}\|\mathbb{G}_{n,k}(\beta) - \mathbb{G}_{n,k}(\beta')\| > \varepsilon\right) \le \frac{qC_1}{\varepsilon}\delta\sqrt{\log(C_2/\delta)} \longrightarrow 0 \qquad \text{as } \delta \downarrow 0,$$

这就证明了该断言。

■

定理证明4.3。 固定 k。

对于任意严格为正的行为策略 μ，令 $g_\mu(\beta, \beta';\ s)$ 表示上文定义 $g_t(\beta, \beta';\ s)$ 的表达式，其中分母中的 $\mu_t(a;\ s)$ 替换为 $\mu(a;\ s)$。特别地，$g_{\mu_t^\circ} = g_t$。

根据（13），由 $\mathbb{G}_{n,\,k}(\beta_1)\dots,\ \mathbb{G}_{n,\,k}(\beta_m)$ 构成的每个有限维向量，都是有界鞅差的归一化和。其条件二次变差具有（β，β'）分块

$$\frac{1}{N_k}\sum_{t=1}^{T_k}\sum_{i=1}^{n}\mathbb{E}\{\xi_t^i(\beta)\xi_t^i(\beta')^\top \mid \mathcal{F}_{t,i-1}\}$$
$$= \frac{1}{T_k}\sum_{t=1}^{T_k}\frac{1}{n}\sum_{i=1}^{n}g_{\mu_{n,t}}(\beta, \beta'; S_t^i).$$

$_n$为证明收敛到 $\Gamma_k(\beta,\ \beta')$，固定 t 并写作 $\eta_{n,\ t} = n^{-1}\sum_{i=1}\delta_{S_t^i}$。一致正性与行为策略的收敛性意味着

$$\left\|\eta_{n,t}\{g_{\mu_{n,t}}(\beta,\beta';\cdot) - g_{\mu_t^\circ}(\beta,\beta';\cdot)\}\right\| = o_{\mathbb{P}}(1).$$

定理 4.1 给出 $\eta_{n,\ t} \Rightarrow \eta_t^\circ$依概率成立。由于 $g_{\mu_t^\circ}(\beta,\ \beta';\ \cdot)$ 有界且连续，因此可得

$$\eta_{n,t} g_{\mu_t^\circ}(\beta,\beta';\cdot) \xrightarrow{\mathbb{P}} \int_{\mathcal{S}} g_t(\beta,\beta';s)\, d\eta_t^\circ(s).$$

结合前述两个收敛结果，对每个固定的 $t \le T_k$,

$$\eta_{n,t} g_{\mu_{n,t}}(\beta,\beta';\cdot) \xrightarrow{\mathbb{P}} \int_{\mathcal{S}} g_t(\beta,\beta';s)\, d\eta_t^\circ(s).$$

因此,

$$\begin{aligned}&\left\|\frac{1}{T_k}\sum_{t=1}^{T_k}\eta_{n,t}g_{\mu_{n,t}}(\beta,\beta';\cdot) - \Gamma_k(\beta,\beta')\right\| \\ &\quad\le \frac{1}{T_k}\sum_{t=1}^{T_k}\left\|\eta_{n,t}g_{\mu_{n,t}}(\beta,\beta';\cdot) - \int_{\mathcal{S}} g_t(\beta,\beta';s)\, d\eta_t^\circ(s)\right\| \xrightarrow{\mathbb{P}} 0,\end{aligned}$$

由于 T_k 固定，且每一项都依概率收敛于零。因此，可预测二次变差收敛至 $\Gamma_k(\beta,\ \beta')$。

由 $g_t(\beta,\ \beta';\ \cdot)$ 的连续性可由同一冻结被积函数论证得到，该论证用于在定理 4.1的证明中建立 $K_{\mu_t^\circ} f$ 的连续性；再结合行为策略的连续性和一致正性即可。

由于鞅增量一致有界，且 $N_k \to \infty$，条件林德伯格条件成立。多元鞅中心极限定理（例如参见 Hall and Heyde，1980年，第3章）给出每个有限策略参数集合上的联合高斯收敛，其协方差块为 Γ_k。

引理 4.2 在欧几里得参数度量下给出了渐近随机等度连续性。有限维收敛性和 $\mathcal{B}$ 的紧致性因此给出

$$\mathbb{G}_{n,k} \rightsquigarrow \mathbb{G}_k \quad \text{in } \ell^\infty(\mathcal{B})^q,$$

其中，$\mathbb{G}_k$ 是具有协方差 Γ_k的紧均值为零高斯过程。特别地，$\|\mathbb{G}_{n,\ k}\|_\infty = O_{\mathbb{P}}(1)$。
根据定理 4.1，估计系数一致收敛于总体系数。其总体像是$\mathrm{int}(\Theta)$的一个紧子集；因此，随着概率趋于一，极小值点位于内部。式（3）的一阶条件为

$$2\hat{W}_{\mathrm{cum},k}(\pi)^\top \Omega \Lambda_{n,k}(\pi,\hat{\theta}^k_{\pi,n}) = \lambda_{n,k}\nabla J(\hat{\theta}^k_{\pi,n}).$$

由于 $\Lambda_{n,\ k}(\pi,\ \theta)$ 关于 θ 是仿射的，因此精确重排为

$$\begin{aligned}\sqrt{N_k}(\hat{\theta}^k_{\pi,n} - \theta_0^\pi) = \{\hat{W}_{\mathrm{cum},k}(\pi)^\top \Omega \hat{W}_{\mathrm{cum},k}(\pi)\}^{-1}\hat{W}_{\mathrm{cum},k}(\pi)^\top \Omega \sqrt{N_k}\Lambda_{n,k}(\pi,\theta_0^\pi) \\ - \frac{\sqrt{N_k}\lambda_{n,k}}{2}\{\hat{W}_{\mathrm{cum},k}(\pi)^\top \Omega \hat{W}_{\mathrm{cum},k}(\pi)\}^{-1}\nabla J(\hat{\theta}^k_{\pi,n}).\end{aligned}$$

的一致一致性和局部有界性 ∇J 意味着 $\sup_{\pi\in\Pi}\|\nabla J(\hat{\theta}^k_{\pi,\ n})\| = O_{\mathbb{P}}(1)$。假设 5 和 6 使第二项 $o_{\mathbb{P}}(1)$ 具有一致性。由于这些矩阵是方阵，并且以趋近于一的概率可逆，

$$\{\hat{W}^\top\Omega\hat{W}\}^{-1}\hat{W}^\top\Omega = \hat{W}^{-1}.$$

利用（10）和 $\|\mathbb{G}_{n,\ k}\|_\infty = O_{\mathbb{P}}$（1）中的一致矩阵收敛，我们得到一致展开式

$$\sup_{\beta\in\mathcal{B}}\left\|\sqrt{N_k}(\hat{\theta}^k_{\pi_\beta,n} - \theta_\beta) - W_{\mathrm{cum},k}(\pi_\beta)^{-1}\mathbb{G}_{n,k}(\beta)\right\| = o_{\mathbb{P}}(1). \tag{26}$$

泛函连续映射定理和（14）现在给出带有 $\mathbb{Z}_k$（β）$= W_{\mathrm{cum},\ k}$（π_β）$^{-1}\mathbb{G}_k$（β）的（15）。乘以 $\nu^\top$ 即可证明价值过程断言。

■

推论4.4的证明。 在预言机实验下，$\mu_{n,\ t} = \mu^\circ_t$ 对于每个 n 和 t，患者历史均服从 Q_{T_k+1}且相互独立同分布。定理4.3 证明中的一致矩阵大数定律、鞅中心极限定理和随机等度连续性论证因此无需行为策略近似误差即可适用。极限总体矩阵和得分协方差核保持不变，因此一致线性化（26）给出相同的高斯极限 $\mathbb{Z}_k$。乘以$\nu^\top$ 即可证明价值过程断言。■

定理4.5的证明。 根据定理4.1，$\hat{\beta}^k_n \to \beta^*$ 依概率收敛。令

$$Z_{n,k}(\beta) = \sqrt{N_k}\{\hat{V}_{n,R,k}(\pi_\beta) - M(\beta)\}.$$

定理 4.3 给出了 $Z_{n,k}$ 在 $\ell^\infty(\mathcal{B})$ 中弱收敛于一个具有连续样本路径的紧高斯过程。其渐近等度连续性以及 $\hat{\beta}^k_n \to \beta^*$ 意味着

$$Z_{n,k}(\hat{\beta}^k_n) - Z_{n,k}(\beta^*) = o_{\mathbb{P}}(1).$$

因此，$Z_{n,k}(\hat{\beta}^k_n)$ 依分布收敛于 $\nu^\top\mathbb{Z}_k(\beta^*)$，从而证明了（16）。

对于替代中心化，$\hat{\beta}^k_n$ 的定义以及 β^* 的最优性给出

$$0 \le M(\beta^*) - M(\hat{\beta}^k_n) \le \hat{V}_{n,R,k}(\pi_{\hat{\beta}^k_n}) - M(\hat{\beta}^k_n) - \{\hat{V}_{n,R,k}(\pi_{\beta^*}) - M(\beta^*)\}.$$

乘以 $\sqrt{N_k}$ 后，上界为 $Z_{n,k}(\hat{\beta}^k_n) - Z_{n,k}(\beta^*) = o_{\mathbb{P}}(1)$。因此 $\sqrt{N_k}\{M(\hat{\beta}^k_n) - M(\beta^*)\} = o_{\mathbb{P}}(1)$，由斯卢茨基定理得到第二种中心化。■

命题 4.6 的证明。固定 k。对于一个完整历史 $h = (s_1, a_1, \ldots, s_{T_k+1}) \in \mathcal{H}_{T_k+1}$，令

$$m_{k,\beta}(h) = \frac{1}{\sqrt{T_k}}\sum_{t=1}^{T_k}\frac{\pi_\beta(a_t; s_t)}{\mu^\circ_t(a_t; s_t)}\left\{u(s_{t+1}, a_t, s_t) + \gamma\Phi(s_{t+1})^\top\theta_\beta - \Phi(s_t)^\top\theta_\beta\right\}\Phi(s_t).$$

根据紧致性、正性和引理3.1，函数 $m_{k,\ \beta}$ 有界，在 (h、 β) 上联合连续，并且关于 β 一致利普希茨连续。定理 4.1 以及 $\mu_{n,\ t}$ 的一致收敛性表明，对于固定的k,

$$\sup_{1\le i\le n}\sup_{\beta\in\mathcal{B}}\|\hat{\psi}^k_i(\beta) - m_{k,\beta}(H_{i,T_k+1})\| = o_{\mathbb{P}}(1).$$

左侧可由一个确定性常数乘以 $\hat{\theta}^{k}_{\pi_\beta,\,n}$ 以及有限个行为策略的一致误差之和来界定，因此尽管它们依赖于完整数据，该界仍然成立。于是，对 β、β' 一致地，

$$\hat{\Gamma}_k(\beta,\beta') - \frac{1}{n}\sum_{i=1}^{n} m_{k,\beta}(H_{i,T_k+1})m_{k,\beta'}(H_{i,T_k+1})^\top = o_{\mathbb{P}}(1).$$

根据定理 Q_{T_k+1} 4.1，经验历史测度收敛于。因此，紧参数集上的一致连续性和有限网论证可得

$$\sup_{\beta,\beta'}\left\|\frac{1}{n}\sum_{i=1}^{n} m_{k,\beta}(H_{i,T_k+1})m_{k,\beta'}(H_{i,T_k+1})^\top - \int m_{k,\beta}(h)m_{k,\beta'}(h)^\top\, dQ_{T_k+1}(h)\right\| = o_{\mathbb{P}}(1).$$

令 $H^{\circ}_{T_k+1}$ 表示一条一般的预言机轨迹，该轨迹由初始分布 η°_1、行为策略 $\{\mu^{\circ}_t\}$ 和转移核 $\{P_a\}$ 生成；令 $\mathbb{E}^{\circ}$ 表示在其概率律下的期望，因此对于每个可积的 f，都有 $\mathbb{E}^{\circ}[f(H^{\circ}_t)] = \int f(h)\, dQ_t(h)$。令 $\xi^{\circ}_t(\beta)$ 表示$\sqrt{T_k}\, m_{k,\,\beta}(H^{\circ}_{T_k+1})$ 中时刻 t 的加项，并定义 $\mathcal{F}^{\circ}_{t,\,0} = \sigma(H^{\circ}_t)$。条件贝尔曼方程给出

$$\mathbb{E}^{\circ}[\xi^{\circ}_t(\beta) \mid \mathcal{F}^{\circ}_{t,0}] = 0.$$

特别地，对于 $t < s$，$\xi^{\circ}_t(\beta)$ 关于时刻 s 治疗前可获得的信息是可测的，因此

$$\mathbb{E}^{\circ}\left[\xi^{\circ}_t(\beta)\xi^{\circ}_s(\beta')^\top\right] = \mathbb{E}^{\circ}\left[\xi^{\circ}_t(\beta)\mathbb{E}^{\circ}\left[\xi^{\circ}_s(\beta')^\top \mid \mathcal{F}^{\circ}_{s,0}\right]\right] = 0.$$

Si类似地，以 S_t 为条件并对处理分配求和，可得 s

$$\mathbb{E}^{\circ}\left[\xi^{\circ}_t(\beta)\xi^{\circ}_t(\beta')^\top\right] = \int_{\mathcal{S}} g_t(\beta,\beta';s)\, d\eta^{\circ}_t(s).$$

因此，展开定义 $m_{k,\beta}$ 的各项和的乘积，并利用跨时间项为零这一性质，

$$\begin{aligned}&\int m_{k,\beta}(h)m_{k,\beta'}(h)^\top\, dQ_{T_k+1}(h)\\ &\quad= \frac{1}{T_k}\sum_{t=1}^{T_k}\int_{\mathcal{S}} g_t(\beta,\beta';s)\, d\eta^{\circ}_t(s) = \Gamma_k(\beta,\beta').\end{aligned}$$

我们已证明 $\sup_{\beta,\beta'}\|\hat{\Gamma}_k(\beta,\beta') - \Gamma_k(\beta,\beta')\| = o_{\mathbb{P}}(1)$。矩阵一致性假设5以及逆矩阵恒等式，推出相应协方差的一致收敛。β 和 $\hat{\beta}^{k}_{n} \to \beta^*$ 的连续性证明了标量方差结论。当 $\sigma^2_{V,k} > 0$时，由斯卢茨基定理可得学生化结论。

■

C 第5节的证明

引理的证明5.1。固定 $\beta \in \mathcal{B}$和一个坐标 $r \in \{1,\dots,d\}$。根据条件均值为零的性质，$\{X_t(\beta)_r\}_{t\ge1}$ 是一个增量有界的鞅差序列，增量上界为 C。Azuma–

Hoeffding 不等式表明，对每个 $\epsilon > 0$,

$$\mathbb{P}\left(\left|\frac{1}{T}\sum_{t=1}^{T} X_t(\beta)_r\right| > \epsilon\right) \le 2\exp\left(-\frac{T\epsilon^2}{2C^2}\right).$$

右侧关于 T 可求和。因此，根据博雷尔–坎泰利引理,

$$\frac{1}{T}\sum_{t=1}^{T} X_t(\beta)_r \xrightarrow{\text{a.s.}} 0.$$

由于 d 固定不变，对向量 $X_t(\beta)$ 同样可以得到这一结论。
为获得一致收敛，固定 $\delta > 0$，并令 $\mathcal{B}_\delta$ 为 $\mathcal{B}$ 的一个有限 δ-网。由利普希茨界,

$$\sup_{\beta\in\mathcal{B}}\left\|\frac{1}{T}\sum_{t=1}^{T} X_t(\beta)\right\| \le \max_{\beta\in\mathcal{B}_\delta}\left\|\frac{1}{T}\sum_{t=1}^{T} X_t(\beta)\right\| + L\delta.$$

T由于 $\mathcal{B}_\delta$ 是有限的，该最大值几乎必然收敛于零。因此,

$$\limsup_{T\to\infty}\sup_{\beta\in\mathcal{B}}\left\|\frac{1}{T}\sum_{t=1}^{T} X_t(\beta)\right\| \le L\delta \qquad \text{almost surely.}$$

将这一论证应用于可数序列 $\delta \downarrow 0$，即可证明该断言。

■

定理的证明 5.2。 固定 $n \ge 1$，并始终记 $\theta_\beta = \theta_0^{\pi_\beta}$。
命题 *(i)* 的证明。对于每个 $\beta \in \mathcal{B}$，定义

$$A_t(\beta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_\beta(A_t^i; S_t^i)}{\mu_{n,t}(A_t^i; S_t^i)}\Phi(S_t^i)\{\Phi(S_t^i) - \gamma\Phi(S_{t+1}^i)\}^\top.$$

由于时刻 t的状态和行为策略 $\mathcal{F}_t$是$\mathcal{F}_{t,\,0}$-可测的，假设 2 给出

$$\mathbb{E}[A_t(\beta) \mid \mathcal{F}_{t,0}] = \frac{1}{n}\sum_{i=1}^{n}\sum_{a\in\mathcal{A}} \pi_\beta(a; S_t^i)\Phi(S_t^i)\{\Phi(S_t^i) - \gamma(P_a\Phi)(S_t^i)\}^\top.$$

因此，记 $X_t(\beta) = A_t(\beta) - \mathbb{E}[A_t(\beta) \mid \mathcal{F}_{t,\,0}]$，我们有

$$\hat{W}_{\text{cum},k}(\pi_\beta) - \overline{W}_{n,k}(\pi_\beta) = \frac{1}{T_k}\sum_{t=1}^{T_k} X_t(\beta).$$

的每个坐标 $X_t(\beta)$ 都是鞅差。此外， Φ的有界性、行为策略的一致正性以及π_β 的利普希茨性质意味着， $X_t(\beta)$ 关于 β一致有界且满足利普希茨性质，其常数与 t无关。将引理 5.1 应用于向量化矩阵，并注意到 $T_k \to \infty$，可得

$$\sup_{\beta\in\mathcal{B}}\|\hat{W}_{\text{cum},k}(\pi_\beta) - \overline{W}_{n,k}(\pi_\beta)\| \xrightarrow{\text{a.s.}} 0,$$

这证明了（i）。

（*ii*）的证明。 我们首先证明在总体系数处，估计函数一致收敛。对于 $\beta \in \mathcal{B}$，令

$$D_t(\beta) = \frac{1}{n}\sum_{i=1}^{n} \frac{\pi_\beta(A_t^i; S_t^i)}{\mu_{n,t}(A_t^i; S_t^i)} e_t^i(\theta_\beta)\Phi(S_t^i).$$

根据条件治疗分配机制，

$$\begin{aligned}\mathbb{E}[D_t(\beta) \mid \mathcal{F}_{t,0}] = \frac{1}{n}\sum_{i=1}^{n} \Phi(S_t^i) \sum_{a\in\mathcal{A}} \pi_\beta(a; S_t^i) \\ \times \left\{ r_a(S_t^i) + \gamma(P_a\Phi)(S_t^i)^\top \theta_\beta - \Phi(S_t^i)^\top \theta_\beta \right\} = 0,\end{aligned}$$

其中，最后一个等号由假设4下的贝尔曼方程推出。因此， $D_t(\beta)$ 是一个鞅差。由引理3.1， θ_β 在 β 上有界且满足利普希茨条件。结合有界性和正性条件，这意味着 $D_t(\beta)$ 满足引理5.1的假设。因此，

$$\sup_{\beta\in\mathcal{B}} \left\| \frac{1}{T}\sum_{t=1}^{T} D_t(\beta) \right\| \xrightarrow{\text{a.s.}} 0.$$

由于 $N_k = nT_k$，根据$\Lambda_{n的k}$ 定义可得

$$\Lambda_{n,k}(\pi_\beta, \theta_\beta) = \frac{1}{T_k}\sum_{t=1}^{T_k} D_t(\beta).$$

因此，上述收敛性沿着 $\{T_k\}$成立。特别地，

$$R_k := \sup_{\beta\in\mathcal{B}} \|\Lambda_{n,k}(\pi_\beta, \theta_\beta)\| \xrightarrow{\text{a.s.}} 0. \tag{27}$$

根据假设8和第 (i) 部分，

$$\begin{aligned}\liminf_{k\to\infty} \inf_{\beta\in\mathcal{B}} \sigma_{\min}\{\hat{W}_{\text{cum},k}(\pi_\beta)\} \geq \liminf_{k\to\infty} \inf_{\beta\in\mathcal{B}} \sigma_{\min}\{\overline{W}_{n,k}(\pi_\beta)\} \\ - \limsup_{k\to\infty} \sup_{\beta\in\mathcal{B}} \|\hat{W}_{\text{cum},k}(\pi_\beta) - \overline{W}_{n,k}(\pi_\beta)\| > 0\end{aligned}$$

几乎必然地，其中不等式由 $\sigma_{\min}(A) \geq \sigma_{\min}(B) - \|A - B\|$推出。因此，经验矩阵最终一致可逆，且

$$\sup_{\beta\in\mathcal{B}} \|\hat{W}_{\text{cum},k}(\pi_\beta)^{-1}\| = O_{\text{a.s.}}(1). \tag{28}$$

此处的界可能依赖于样本路径，但对于所有充分大的 k，它都是有限的，且与 k 无关。

接下来，我们利用 $\hat{\theta}^k_{\pi_\beta,\ n}$ 的极小化性质。由于 $\theta_\beta \in \Theta$ 和 $J \geq 0$，

$$\begin{aligned}\Lambda_{n,k}(\pi_\beta, \hat{\theta}^k_{\pi_\beta,n})^\top \Omega \Lambda_{n,k}(\pi_\beta, \hat{\theta}^k_{\pi_\beta,n}) \\ \leq \Lambda_{n,k}(\pi_\beta, \theta_\beta)^\top \Omega \Lambda_{n,k}(\pi_\beta, \theta_\beta) + \lambda_{n,k} J(\theta_\beta).\end{aligned}$$

根据引理 3.1，集合 $\{\theta_\beta : \beta \in \mathcal{B}\}$ 是紧的。因此， J的连续性给出 $C_J :=$上确界$_{\beta\in\mathcal{B}}\, J(\theta_\beta) < \infty$。利用 Ω 的正定性，并对 β取上确界，可得

$$\sup_{\beta\in\mathcal{B}} \|\Lambda_{n,k}(\pi_\beta, \hat{\theta}^k_{\pi_\beta,n})\|^2 \leq \frac{\lambda_{\max}(\Omega) R_k^2 + \lambda_{n,k} C_J}{\lambda_{\min}(\Omega)} \xrightarrow{\text{a.s.}} 0,$$

由（27）和$\lambda_{n,\ k} \to 0$。

最后，估计函数的仿射表示给出

$$\hat{W}_{\mathrm{cum},k}(\pi_\beta)(\hat{\theta}^k_{\pi_\beta,n} - \theta_\beta) = \Lambda_{n,k}(\pi_\beta, \theta_\beta) - \Lambda_{n,k}(\pi_\beta, \hat{\theta}^k_{\pi_\beta,n}).$$

因此，对所有充分大的 k,

$$\sup_{\beta\in\mathcal{B}} \|\hat{\theta}^k_{\pi_\beta,n} - \theta_\beta\| \le \sup_{\beta\in\mathcal{B}} \|\hat{W}_{\mathrm{cum},k}(\pi_\beta)^{-1}\| \left\{ R_k + \sup_{\beta\in\mathcal{B}} \|\Lambda_{n,k}(\pi_\beta, \hat{\theta}^k_{\pi_\beta,n})\| \right\} \xrightarrow{\text{a.s.}} 0,$$

由（28）以及前两个收敛结果可得。这证明了（ii）。

ˆ（*iii*）的证明。由（ii）以及$V_{n,\ R,\ k}$和 M的定义，

$$\sup_{\beta\in\mathcal{B}} |\hat{V}_{n,R,k}(\pi_\beta) - M(\beta)| \le \|\nu\| \sup_{\beta\in\mathcal{B}} \|\hat{\theta}^k_{\pi_\beta,n} - \theta_\beta\| \xrightarrow{\text{a.s.}} 0.$$

由于ˆβ^k_n 在 $\mathcal{B}$ 上使ˆ$V_{n,\ R,\ k}$达到最大值，而 β^* 使 M达到最大值，因此

$$0 \le M(\beta^*) - M(\hat{\beta}^k_n) \le 2 \sup_{\beta\in\mathcal{B}} |\hat{V}_{n,R,k}(\pi_\beta) - M(\beta)| \xrightarrow{\text{a.s.}} 0.$$

对于任意 $\epsilon > 0$， $\mathcal{B}$的紧致性、 M的连续性以及 β^* 的唯一性意味着

$$\sup_{\substack{\beta\in\mathcal{B} \\ \|\beta-\beta^*\|\ge\epsilon}} M(\beta) < M(\beta^*)$$

只要约束集非空。由此可知，前述收敛性最终几乎必然迫使 $\|\hat{}\beta^k_n - \beta^*\| < \epsilon$。由于 ϵ 是任意的，ˆβ^k_n $\xrightarrow{\text{a.s.}} \beta^*$，从而证明了 (iii)。

■

参考文献

Kazuoki Azuma。某些相依随机变量的加权和。东北数学杂志，19(3):357–367，1967年。

Bibhas Chakraborty、Eric B. Laber 和 Ying-Qi Zhao。关于数据驱动的动态治疗方案预期表现的推断。《临床试验》，11(4):408–417，2014年。

Raaz Dwivedi、Katherine Tian、Sabina Tomkins、Predrag Klasnja、Susan Murphy 和 Devavrat Shah。序贯实验中的反事实推断。《统计年刊》，53(6):2380–2406，2025年。doi: 10.1214/25-AOS2519。

Ashkan Ertefaie 和 Robert L. Strawderman。构建无限时间范围内的动态治疗方案。《生物统计学》，105(4)：963–977，2018 年。doi：10.1093/biomet/asy043。

Gene H. Golub 和 Charles F. Van Loan。《矩阵计算》。约翰斯·霍普金斯大学出版社，第4版，2013年。

Vitor Hadad、David A. Hirshberg、Ruohan Zhan、Stefan Wager 和 Susan Athey。自适应实验中的策略评估置信区间。美国国家科学院院刊，118 (15):e2014602118，2021年。doi: 10.1073/pnas.2014602118。Peter Hall 和 Christopher C. Heyde。鞅极限定理及其应用。学术出版社，1980年。Wassily Hoeffding。有界随机变量和的概率不等式。《美国统计学会杂志》，58(301):13–30，1963年。Nathan Kallus 和 Masatoshi Uehara。在马尔可夫决策过程中的高效离策略评估中采用双重强化学习。机器学习研究杂志，21(167):1–63，2020年。网址：https://www.jmlr.org/papers/v21/19-827.html。

Nathan Kallus 和 Masatoshi Uehara。通过双重强化学习高效破解离策略评估中的时域诅咒。《运筹学》，70(6):3282–3302，2022。doi: 10.1287/opre. 2021.2249。

Agnan Kessy、Alex Lewin 和 Korbinian Strimmer。最优白化与去相关。美国统计学家，72(4):309–314，2018。doi: 10.1080/00031305.2016.1277159。

Michail G. Lagoudakis 和 Ronald Parr。最小二乘策略迭代。机器学习研究期刊，4:1107–1149，2003。网址https://www.jmlr.org/papers/v4/lagoudakis03a.html。

Tze Leung Lai 和 Ching Zong Wei。随机回归模型中的最小二乘估计及其在动态系统辨识与控制中的应用。《统计年鉴》，10(1):154–166，1982。

Peng Liao、Zhengling Qi、Runzhe Wan、Predrag Klasnja 和 Susan A. Murphy。平均奖励马尔可夫决策过程中的批量策略学习。《统计年鉴》，50(6)：3364–3387，2022 年。doi：10.1214/22-AOS2231。

Weidong Liu、Jiyuan Tu、Xi Chen 和 Yichen Zhang。强化学习中的稳健策略评估的在线估计与推断。《统计年鉴》，53(5)：2128–2152，2025 年。doi：10.1214/25-AOS2533。

Nan Lu、Ethan Lee、Ruicheng Ao、David Simchi-Levi 和 Junwei Lu。强化学习的最优价值推断。*arXiv* 预印本 *arXiv:2609.09981*，2026 年。URL：https://arxiv.org/abs/2609.09981v4。版本 4。

Daniel J. Luckett、Eric B. Laber、Anna R. Kahkoska、David M. Maahs、Elizabeth Mayer-Davis 和 Michael R. Kosorok。使用 V 学习估计移动健康中的动态治疗方案。《美国统计学会杂志》，115(530):692–706，2020 年。doi：10.1080/01621459.2018. 1537919。

Sunil Madhow、Dan Qiao、Ming Yin 和 Yu-Xiang Wang。利用自适应收集数据进行离线强化学习的速率。在 第 7 届年度动力学与控制学习会议论文集，第 283 卷，载于 《机器学习研究论文集》，第 259–271 页。PMLR，2025 年。网址：https://proceedings.mlr.press/v283/madhow25a.html。

Chengchun Shi、Sheng Zhang、Wenbin Lu 和 Rui Song。无限时域环境下强化学习价值函数的统计推断。《皇家统计学会杂志》：*B* 辑（统计方法学），84(3):765–793，2022 年。doi：10.1111/rssb.12465。

Vasilis Syrgkanis 和 Ruohan Zhan。强化学习后推断。《运筹学》，74(2)：917–957，2026 年。doi：10.1287/opre.2024.1019。

Anna L. Trella、Kelly W. Zhang、Hinal Jajal、Inbal Nahum-Shani、Vivek Shetty、Finale Doshi-Velez 和 Susan A. Murphy。口腔健康临床试验中已部署的在线强化学习算法。载于*AAAI*人工智能会议论文集，第39卷，28792–28800页，2025年。doi：10.1609/aaai.v39i28.35143。

Aad W. van der Vaart。《渐近统计》。剑桥大学出版社，19 98.

Aad W. van der Vaart 和 Jon A. Wellner。弱收敛与经验过程：及其在统计学中的应用。施普林格，1996年。

Weichen Wu、Gen Li、Yuting Wei 和 Alessandro Rinaldo。基于时序差分学习的策略评估统计推断。《强化学习期刊》，第7卷，2026年。网址：https://rlj.org/papers/wu2026statistical。

Huizhen Yu. 最小二乘时序差分方法：一般条件下的分析。《*SIAM*控制与优化期刊》，50(6):3310–3343，2012。doi: 10.1137/100807879。

Ruohan Zhan、Zhimei Ren、Susan Athey 和 Zhengyuan Zhou。利用自适应收集数据进行策略学习。《管理科学》，70(8):5270–5297，2024。doi: 10.1287/mnsc.2023.4921。

Kelly W. Zhang、Lucas Janson 和 Susan A. Murphy。批处理多臂老虎机的推断。载于《神经信息处理系统进展》，第33卷，第9818–9829页。Curran Associates, Inc.，2020年。

Kelly W. Zhang、Lucas Janson 和 Susan A. Murphy。基于自适应收集数据的 M 估计量进行统计推断。载于《神经信息处理系统进展》，第34卷，第7460–7471页。Curran Associates, Inc.，2021年。

Kelly W. Zhang、Lucas Janson 和 Susan A. Murphy。针对纵向数据自适应抽样后的统计推断。*arXiv* 预印本 *arXiv:2202.07098*，2022 年。网址：https://arxiv.org/abs/2202.07098。

Kelly W. Zhang、Nowell Closser、Anna L. Trella 和 Susan A. Murphy。用于数字健康干预的可复现多臂老虎机。《统计科学》，40(4):671–691，2025 年。doi：10.1214/25-STS1017。