用于延迟反馈收益管理的选择模型辅助 Q 学习

Owen Shen[1] PatrickJaillet[11]麻
省理工学院{owenshen, jaillet}
@mit.edu

关键词：强化学习、部分世界模型、选择模型、延迟反馈、收益管理

arXiv:2602.02283v1 [cs.LG] 2 Feb 2026

目录

1 引言 **4**
1.1 反馈语义与符号 .41.2 设定：带冲击的延迟反馈马尔可夫决策过程 .51.3 贡献 .51.4 运营设定 .6

2 相关工作 6

3 问题表述 **7**
3.1 延迟反馈下的收益管理 .73.2 部分世界模型：离散选择模型 .83.3 选择模型辅助集成 .9

4 方法：嵌入部 分世界模型 9
4.1 固定离散选择模型部署 .94.2 模型插补采样框架 .104.3 实现 .10

5 理论分析 115.1 收敛分析 .11

6 实验验证 12
6.1 设置 .126.2 Q1：在模型设定正确时，CA-DQN 是否能匹配 MB-DQN？ .126.3 Q2：在同族分布偏移下，CA-DQN 是否能提高鲁棒性？ .136.4 Q3：在结构设定错误下，CA-DQN 的表现是否会下降？ .136.4.1 参数错设：二次价格敏感度 .146.4.2 结构设定错误：族外压力测试 .14

7 讨论 15

8 结论 **15**

I 补充材料 **19**

9 主要定理的证明 **19**
9.1 证明概述 .199.2 假设 A1 与 A6 之间的关系 .19

10 DCM 的可识别性与性质 **19**
10.1 多项逻辑回归模型的可识别性条件 .1910.2 双稳健奖励目标的偏差-方差性质 .2110.2.1 目标构造 .2210.2.2 偏差分析 .2210.2.3 方差分析 .2310.2.4 与因果推断中双稳健估计的联系 .2410.2.5 经验校准 .2510.3 命题：多项Logit对数似然的严格凹性 .2510.4 命题：递归最大似然估计的参数收敛 .2710.5 命题：DCM 的满秩 Fisher 信息 .29

11 多项 Logit 收敛速率分析 30

12 延迟反馈分析 3012.1 成熟率引理：将延迟分布与样本量增长联系起来 31

13 Q 学习收敛 3313.1 定理：近似误差递减且速率为.....的非平稳 Q 学习 34

14 分布偏移下的外推 53
14.1 多项 Logit 的利普希茨性质 .5314.2 命题 30 的陈述与证明 .5414.3 命题30的完整证明 .5514.4 Q 函数误差的扩展 .5614.5 结构假设成立时 .57

15 可复现性与实现 58
15.1 完整环境与状态约简 .5815.2 表格算法与自适应扩展 .5815.3 超参数摘要 .5915.4 实验核算 .5915.5 计算资源 .5915.6 模拟与现实世界澄清 .6015.7 TOST等效性分析方法 .6015.7.1 方法论 .6115.7.2 结果 .6115.8 模型设定错误压力测试 .6115.8.1 方法论 .6215.8.2 结果 .6315.8.3 分析 .6315.9 跨模型族压力测试 .6415.9.1 OOF 模拟器设计 .64

15.9.2 OOF 实验协议 .6515.9.3 OOF 结果 .6515.9.4 OOF 分析 .65

16 伦理与负责任部署 66

16.1 数据隐私与保护 .6616.2 算法公平性与价格歧视 .6616.3 经济与社会影响 .6616.4 部署防护措施 .66

17 管理启示 67

17.1 消费者延迟决策下的收入优化 .6717.2 实施策略 .6717.3 何时使用此方法 .67

18 局限性与未来工作 67

18.1 理论与实验之间的差距 .6818.2 模型设定错误及其后果 .6818.3 耦合分析的局限性 .6818.4 与现有框架的关系 .68

摘要

我们研究延迟反馈下的收益管理强化学习，其中相当大一部分价值由预订数日后才能观测到的客户取消和修改所决定。我们提出选择模型辅助强化学习：使用经过校准的离散选择模型作为固定的部分世界模型，在决策时刻对学习目标中的延迟分量进行插补。在固定模型部署机制下，我们证明，采用模型插补目标的表格型 Q 学习会收敛到最优 Q 函数的 $O(\varepsilon/(1-\gamma))$ 邻域，其中 ε 总结了部分模型误差，另有一个 $O(t^{-1/2})$采样项。在基于 61,619 条酒店预订数据校准的模拟器中进行的实验（1,088 次独立运行）显示：(i) 在平稳设置下，与成熟度缓冲深度Q网络基线相比，没有统计上可检测的差异；(ii) 在模型族内参数变化下效果为正，在经过 Holm–Bonferroni 校正的 10 个分布偏移场景中，有 5 个取得了显著增益（最高达 12.4%）；(iii) 在结构设定错误下表现持续下降，即选择模型假设被违反时，收益降低 1.4–2.6%。这些结果刻画了部分行为模型何时能够提高分布偏移下的鲁棒性，以及何时会引入有害偏差。

1 引言

（1）运营现象。 在收益管理领域——这是一个拥有数十年运筹学文献积累的领域 Bitran 和 Mondschein [1995], Talluri 和 van Ryzin [2004]——客户修改和取消（冲击）会在初次预订后 1–14 天发生，占总收入的 20–40%。这些延迟结果会显著影响定价决策的价值。

（2）学习瓶颈。 此类系统中的强化学习面临信用分配问题：时刻 t 所采取动作的学习标签 $r_t = r_t^{\text{imm}} + r_t^{\text{del}}$需要延迟分量 r_t^{del}，而该分量只有在延迟 $\Delta \in \{1, \ldots, 14\}$ 天后才会揭晓。标准方法必须等待冲击结果明朗后才能进行更新。

（3）可用结构。 订单特征（房型、客户特征、预订日期）在决策时刻即可获知。通过最大似然估计离线校准的离散选择模型（DCM），能够基于这些特征对冲击结果进行概率预测 $P(z|o;\theta)$ [McFadden, 1974, Train, 2009]。

(4) 方法。 我们提出**选择模型辅助强化学习**：使用固定的预训练 DCM 在决策时刻进行插补 r_t^{del}，无需等待真实值即可立即进行 Q 学习更新。DCM 充当部分世界模型——它预测待处理订单的冲击结果，而 Q 函数则学习定价动作的价值。

(5) 边界条件。 当发生模型族内参数变化（需求缩放、竞争强度）且 DCM 结构仍然有效时，这种方法有所帮助；但在结构设定错误 的情况下则会带来负面影响，因为真实行为违反了模型假设（IIA 违背、细分群体异质性）。当模型族不正确时，使外推成为可能的同一归纳偏置也会成为负担。

两种评估机制。 我们在两个互补的场景中分析这一框架：(i) 连续运营 机制，在此机制下，我们证明固定 DCM 下表格型 Q 学习的渐近收敛界；(ii) 固定窗口运营 机制，在此机制下，我们使用基于 61,619 笔酒店预订校准的模拟器，实证量化部署时的影响。

1.1 反馈语义与记号

我们区分 (i) 日历时间 t的实际现金流，其中包括在 t到期的先前预订带来的冲击；以及 (ii) 与决策时刻 t所采取动作相关的学习标签 。令 $\bar{r}_t$ 表示在时间 t观察到的实际现金流：

$$\bar{r}_t := r_t^{\text{imm}} + \sum_{o \in \mathcal{M}_t} r^{\text{shock}}(o, z_o),$$

其中， $\mathcal{M}_t$ 是其修改/取消结果在时间t到期的先前创建订单集合。为进行信用分配，我们将时间 t 所采取动作的（可能延迟的）学习标签定义为

$$r_t := r_t^{\text{imm}} + r_t^{\text{del}}, \qquad r_t^{\text{del}} := \sum_{o \in \mathcal{N}_t} r^{\text{shock}}(o, z_o),$$

其中 $\mathcal{N}_t$ 是在时间 t 创建的新订单集合，其冲击结果 z_o 在随机延迟 $\Delta_o \in \{1, \ldots, 14\}$之后才可观测。因此 r_t^{imm}可立即观测，而 r_t^{del}仅在相应订单到期时才被揭示。我们的方法使用校准选择模型在决策时刻插补 r_t^{del}。

1.2 设定：带冲击的延迟反馈马尔可夫决策过程

在每个决策时刻 t，收益经理：
1. 观测当前库存 s_t 以及任何冲击——来自先前订单的已解决的修改/取消

2. 设定价格 a_t
3. 观测新预订 $j_t \sim P(j|s_t, a_t; \theta)$、订单特征 x_t（房型、客户特征、预订日期）以及即时收入 r_t^{imm}

特征为 $\mathcal{N}_t$ 的新订单会被添加到待定集合 $\mathcal{P}_t$；这些特征使 DCM 能够预测未来的冲击结果。

先前订单产生的冲击。较早周期 $\tau < t$ 下的订单可能会产生修改或取消，并在周期 t 结算。令 $\mathcal{P}_t$ 表示 t 时的待处理订单，$\mathcal{M}_t \subseteq \mathcal{P}_{t-1}$ 表示在 t 到期的订单，$z_o \sim P(z|o; \theta)$ 表示订单 o 的修改结果。订单会在 1–14 天内结算。时间 t 的实际现金流：

$$\bar{r}_t = r^{\text{imm}}(s_t, a_t, j_t) + \sum_{o \in \mathcal{M}_t} r^{\text{shock}}(o, z_o) \tag{1}$$

DCM 为何有帮助。标准强化学习会等待冲击结算。由于我们知道待处理订单及其特征，DCM 可以预测冲击结果 $P(z|o; \theta)$，从而实现即时学习。
何时观测到什么。在决策时刻 t，智能体观测到（s_t，a_t）、即时预订结果及其特征，以及即时收入 r_t^{imm}。在 t 创建的新订单，其延迟结果（取消/修改）只有在随机延迟 $\Delta \in \{1, \ldots, 14\}$ 后才会揭示；我们用 r_t^{del} 表示相应的延迟奖励分量。此外，在日历时间 t，系统可能会观测到较早订单的到期结果；这些结果补全了较早决策对应的延迟标签，并用于学习（基线）和模型监测/校准。

1.3 贡献

1. **选择模型辅助强化学习框架**：我们通过耦合随机逼近，形式化地构建嵌入部分世界模型的方法，其中预测模型 m_θ 与价值函数 Q_w 相互作用，并实现受控误差传播。

2. **收敛分析**：针对固定的预训练 DCM 部署情形，我们证明 $\|Q_t - Q^*\|_\infty = O(\varepsilon/(1-\gamma) + t^{-1/2}\sqrt{\log(\cdot)})$ 以高概率成立，其中 ε 是 DCM 近似误差——由不可约偏差和渐消的抽样噪声构成（定理 3）。

3. **实证验证**：1,088次独立训练运行结合压力测试表明：（i）在平稳设置下无性能差异；（ii）在参数偏移下结果为正（10个情景中有5个在统计显著性意义上得到改善）；以及（iii）持续表现不佳于结构设定错误情形（3项分布外测试中有0项得到改善）。

表1：两种互补的分析机制。

	机制A（理论——ory)	**机制B（实证——ics)**
目标	无限时域	有限 Train-
	折扣 MDP ($t \to \infty$)	ing/评估窗口
设置	表格型 Q 学习，固定 DCM	DQN，模拟器来自 61 ,619 本书-
		ings
输出	渐近的 bias	固定窗口运营影响
	楼层 + 暂态 rate	

1.4 运营设定

我们考虑一种两阶段运营模式：（1）校准，即基于历史预订数据对DCM进行校准，随后进行（2）部署，并在Q学习期间保持DCM固定。这种分离反映了计算约束——DCM 校准需要对完整客户旅程进行批处理——并确保价值学习期间采样的一致性。详见第4节。

2 相关工作

世界模型。 神经方法（MuZero、Dreamer、PlaNet）取得了令人瞩目的表现；具有理论依据的方法 [Chua 等人，2018，Janner 等人，2019，Luo 等人， 2019] 分析误差累积。Dyna-Q [Sutton， 1990] 奠定了混合基于模型与无模型学习的基础。我们嵌入了领域特定的参数化模型（DCM），而非学习动态，从而提供可解释性和系统化外推（命题30）。

模型误差与非平稳性。 Ross 等人 [2011], Talvitie [2017] 分析复合误差；MOPO [Yu 等人， 2020] 和 MOReL [Kidambi 等人， 2020] 探讨不确定性下的离线强化学习。我们分析参数化预测模型与价值学习之间的显式耦合。

延迟反馈与信用分配。 Joulani 等人 [2013], Howson 等人 [2023], Kuang 等人[2023] 分析强化学习和赌博机中的延迟反馈；时间抽象框架 [Sutton 等人， 1999]为多时间尺度决策问题提供了结构。广义优势估计 [Schulman 等人，2016] 旨在解决信用分配中的偏差-方差权衡。我们表明，模型插补采样能够减轻延迟惩罚。架构解决方案 [Arjona-Medina 等人，2019，Hung 等人，2019] 采用神经网络；我们基于参数化模型的方法能够提供可解释的预测。

收益管理中的选择模型。 基于 McFadden [1974], Train [2009], 离散选择模型提供了有理论依据的概率预测。基于选择的线性规划 [Liu 和 van Ryzin， 2008] 以及采用参数化选择模型的动态定价 [Broder 和 Rusmevichientong， 2012] 奠定了理论基础。经典收益管理 [Talluri 和 van Ryzin，2004年，Gallego 和 Topaloglu， 2019] 假设需求平稳。我们的框架增加了：(i) 自适应价值函数，(ii) 非平稳性处理，以及 (iii) 延迟反馈。近期研究 [Bondoux 等，2020，Meng 等，2024，Zhu 等，2024] 将强化学习应用于收益管理；我们聚焦于嵌入式选择模型的前向强化学习。

表2：符号汇总

符号	描述
马尔可夫决策过程组成部分	
$\mathcal{S}$	状态空间（库存水平 $\{0,\ldots,C\}$）
$\mathcal{A}$	动作空间（价格水平 $\{p_1,\ldots,p_K\}$）
$P(\cdot\|s,\ a)$	转移核
$R(s,\ a)$	奖励函数
γ	折扣因子
H	时域（预订窗口长度）
C	酒店容量（最大库存，$C=26$）
K	价格水平数量（$K=13$）
状态与订单	
s_t	库存状态（观测值）
$\mathbf{x}_t$	时间 t 的客户特征（$\mathbf{x}_t\in\mathbb{R}^{12}$）
$\mathbf{p}_t^c$	时间 t的竞争对手价格
$\tilde{s}_t$	完整状态 $(s_t,\ \mathbf{x}_t,\ \mathbf{p}_t^c,\ \mathcal{P}_t)$
a_t	价格动作
$\mathcal{P}_t$	时间 t的待处理订单
$\mathcal{N}_t$	时间 t创建的新订单
$\mathcal{M}_t$	时间 t到期（结算中）的订单
j_t	时间 t的房型预订
$\mathcal{J}$	房型集合及外部选项
x_o	订单 o 的特征向量
奖励与延迟	
r_t^{imm}	时间 t的即时收入
r_t^{del}	时间 t动作的延迟收入部分
$\bar{r}_t$	日历时间 t的实际现金流
z_o	订单 o 的冲击结果
Δ	冲击结果揭晓前的延迟（1–14 天）
选择模型	
θ^*	固定离散选择模型参数
$\phi_j(\cdot)$	房型的特征映射 j

3 问题表述

3.1 延迟反馈下的收益管理

完整运营环境。 考虑完整的酒店收益管理场景。在每个决策时刻 t，环境状态包括：

- **库存**：可用客房 $s_t\in\{0,1,\ldots,C\}$，其中 $C=26$
- **客户特征**：到店顾客的特征为 $\mathbf{x}_t\in\mathbb{R}^d$（距到店天数、预订渠道、忠诚度等级、搜索历史；$d=12$）
- **竞争价格**：从市场观察到的竞争对手客房价格 $\mathbf{p}_t^c\in\mathbb{R}^{K_c}$
- **待处理订单**：过往预订的集合 $\mathcal{P}_t$，等待修改/取消结果，每笔订单都具有订单特征 x_o

因此，完整状态为 $\tilde{s}_t=(s_t,\ \mathbf{x}_t,\ \mathbf{p}_t^c,\ \mathcal{P}_t)$。客户行为——预订决策 $j_t\sim P(j|\mathbf{x}_t,\ \mathbf{p}_t,\ \mathbf{p}_t^c;\ \theta)$以及冲击结果 $z_o\sim P(z|x_o;\ \theta)$——取决于特征和竞争环境。奖励可分解为

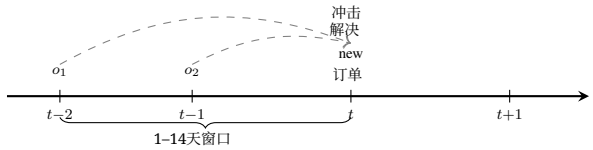


图 1：延迟反馈结构。较早时期下达的订单会产生冲击（修改、取消），并在 1–14 天后解决。DCM 根据订单特征预测冲击结果，从而能够立即填补奖励。

为 $r_t = r_t^{\text{imm}} + r_t^{\text{del}}$，其中延迟分量 r_t^{del}只有在 $\Delta \in \{1, \ldots, 14\}$ 天后才会揭示。

复杂性挑战。直接在 $\tilde{s}_t$ 上学习是棘手的：(i) 客户特征 $\mathbf{x}_t$ 维度很高（$d = 12$），且每次到达时都不同；(ii) 竞争价格 $\mathbf{p}_t^c$ 频繁变化；(iii) 待处理订单$\mathcal{P}_t$ 呈指数级增长（$|\tilde{\mathcal{S}}| = O(|\mathcal{S}| \cdot 2^{|\mathcal{P}|})$）。然而，我们观察到一个关键的结构属性：**客户特征和竞争价格只影响客户决策，不会直接影响库存动态**。这使得状态空间缩减成为可能。

三**种建模方法**。我们比较不同方法如何处理这种复杂性：

方法 1：增广马尔可夫决策过程（完整状态 + 延迟结构）。保留完整状态 $\tilde{s}_t = (s_t, \mathbf{x}_t, \mathbf{p}_t^c, \mathcal{P}_t)$:

$$\tilde{P}(\tilde{s}_{t+1}|\tilde{s}_t, a_t) = P(\mathbf{x}_{t+1}) \cdot P(\mathbf{p}_{t+1}^c) \cdot P(j_t|\mathbf{x}_t, a_t, \mathbf{p}_t^c) \cdot \prod_{o \in \mathcal{M}_t} P(z_o|x_o) \tag{2}$$

这是一个具有马尔可夫转移的标准马尔可夫决策过程，但状态空间随特征维度和待处理订单数量呈指数增长——表格方法难以处理；由于$\mathbf{x}_t$ 和$\mathbf{p}_t^c$的非平稳性，函数逼近也颇具挑战。

方法 2：缩减状态、忽略结构（MB-DQN）。仅将库存作为状态：$s_t \in \{0, \ldots, C\}$，并将客户特征、竞争价格和待处理订单视为未建模噪声：

$$P(s_{t+1}|s_t, a_t) = \mathbb{E}_{\mathbf{x}, \mathbf{p}^c, \{z_o\}} \left[P(j|s_t, a_t) \cdot \mathbf{1}[s_{t+1} = f(s_t, j, \{z_o\})] \right] \tag{3}$$

状态空间易于处理（$|\mathcal{S}| = 27$），但我们会失去客户特征中的预测信息，而且必须等待延迟结果成熟。

方法三：基于模型的插补（CA-DQN，我们的方法）。**关键观察：**客户特征 $\mathbf{x}_t$ 和竞争价格$\mathbf{p}_t^c$ 仅影响客户决策，不会直接影响库存动态。我们可以通过离线校准的离散选择模型（DCM）对它们进行预处理：

$$s_t \in \{0, \ldots, C\} \quad \text{(reduced state)} \tag{4}$$

$$\hat{r}_t^{\text{del}} = \sum_{o \in \mathcal{N}_t} r^{\text{shock}}(o, \hat{z}_o), \quad \hat{z}_o \sim P(z|x_o; \theta^*) \quad \text{(imputed via DCM)} \tag{5}$$

DCM 能吸收高维的客户／竞争环境上下文：它以$(\mathbf{x}_t, \mathbf{p}_t, \mathbf{p}_t^c, x_o)$ 为输入，并输出选择／冲击预测。这使得我们能够建立一个易于处理的 27 状态马尔可夫决策过程，同时通过部分世界模型利用丰富的特征信息。定理3 界定了使用插补后的$\hat{r}_t^{\text{del}}$而非真实的 r_t^{del}所产生的近似误差。

3.2 部分世界模型：离散选择模型

定义用于房型选择的多项逻辑斯蒂离散选择模型：

$$P(j|\mathbf{x}, \mathbf{p}, \theta) = \frac{\exp(V_j(\mathbf{x}, \mathbf{p}; \theta))}{\sum_{k \in \mathcal{J}} \exp(V_k(\mathbf{x}, \mathbf{p}; \theta))} \tag{6}$$

其中， $V_j = \theta^\top \phi_j(\mathbf{x}, \mathbf{p})$是系统效用， ϕ_j： $\mathcal{X} \times \mathcal{P} \to \mathbb{R}^d$是特征映射， $\mathcal{J} = \{j_1, \ldots, j_K, \text{outside}\}$表示 K 房型以及外部选项（不购买）。[1]

DCM 通过以下方式实现易于处理的建模：

- 降低维度： $\theta \in \mathbb{R}^d$ 使用 $d \ll |\mathcal{S}| \times |\mathcal{A}|$
- 施加结构：效用可分解为 $V_j = \alpha_j - \beta p_j + \gamma^\top \mathbf{x}$
- 实现闭式梯度： $\nabla_\theta \log\ P(j|\cdot) = \phi_j - \mathbb{E}_{k \sim P}[\phi_k]$

作为预处理层的 DCM。DCM 充当预处理层，用于吸收高维的客户与竞争环境上下文。其输入包括：(i) 客户协变量$\mathbf{x} \in \mathbb{R}^{12}$（距到店天数、预订渠道、忠诚度等级等）；(ii) 我们的价格向量$\mathbf{p}$；(iii) 竞争对手价格$\mathbf{p}^c$（通过相对价格特征捕获）；以及 (iv) 用于冲击预测的订单特征 x_o。通过离线校准 θ^*，我们将这些丰富的上下文压缩为概率预测，从而能够即时进行奖励填补，同时保持强化学习问题的状态数易于处理（共 27 个状态）。

3.3 选择模型辅助集成

定义 1（选择模型辅助强化学习）。一个耦合系统包含：

$$m_\theta : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{J}) \quad \textit{(DCM kernel)} \tag{7}$$

$$Q_w : \mathcal{S} \times \mathcal{A} \to \mathbb{R} \quad \textit{(Q-learning kernel)} \tag{8}$$

对于每次状态转移，采样 $j' \sim P(j|s, a, \theta)$ 并生成合成样本 (r'_t, s'_{t+1})，其中 $r'_t = r^{imm} + r^{shock}_{j'}$。Q学习使用这些合成样本。

DCM 和 Q学习构成一个通过双时间尺度随机逼近进行分析的耦合系统[Borkar, 1997]。

4 方法：嵌入部分世界模型

4.1 固定离散选择模型部署

在实践中，我们部署一个两阶段系统：

1. **C校准阶段**：在历史 da 上估计 $\theta^* = \arg\max_\theta \sum_{i=1}^N \log\ P(j_i|s_i, a_i;\ \theta)$ ta
2. **部署阶段**：固定 θ^* 并使用模型插补采样运行 Q学习

这种分离在计算上是可行的：DCM 校准需要对完整客户旅程进行批处理，而 Q学习则在线运行。Q更新使用合成样本：

$$Q_{t+1}(s, a) = Q_t(s, a) + \alpha_t \left[r'_t + \gamma \max_{a'} Q_t(s', a') - Q_t(s, a) \right] \tag{9}$$

其中 (r'_t, s'_{t+1}) 由使用固定 DCM 的模型插补采样生成 θ^*。定理 3 给出了该情形的有限时间界。

备注1（扩展：自适应离散选择模型）。 原则上，随着新的延迟结果逐渐成熟，*DCM* 可以在线自适应。附录 *15.2* 使用双时间尺度随机逼近理论对此进行了分析 *[Borkar, 1997]*。然而，我们的实验使用固定离散选择模型，这反映了现实部署中再校准并不频繁（例如，在季节之间）的情况。

[1] $K = 6$ 在我们的实验中。冲击被建模为切换概率 $P(z|o;\ \theta)$ 以订单特征为条件 o。参见附录 15。

4.2 模型插补采样框架

DCM 提供切换概率，用于模拟客户行为变化（预订、取消、修改）。我们不使用期望值，而是采用**模型插补采样**：[2]

定义 2 (模型插补采样)。 对于每个真实状态-动作对(s_t, a_t)，我们生成一个合成转移：

1. 新预订：采样 $j' \sim P(j|s_t, a_t; \theta)$，其中 $j \in \mathcal{J}$是房型选择 *2.* 待处理冲击：对于每个 $o \in \mathcal{P}_t$，采样 $z_o \sim P(z|o; \theta)$，其中 $z \in \mathcal{Z}$是修改结果 *3.* 合成奖励：$r'_t = r^{imm}(s_t, a_t, j') + \sum_o r^{shock}(o, z_o)$ *4.* 合成下一状态：$s'_{t+1} = f(s_t, a_t, j', \{z_o\})$ Q学习使用合成元组 $(s_t, a_t, r'_t, s'_{t+1})$。

这会为每次交互生成两个样本：实际样本 $(r_t,\ s_{t+1})$ 和合成样本 $(r'_t,\ s'_{t+1})$。Q更新使用合成样本：

$$Q_{t+1}(s_t, a_t) = (1 - \alpha_t)Q_t(s_t, a_t) + \alpha_t[r'_t + \gamma \max_{a'} Q_t(s'_{t+1}, a')] \tag{10}$$

4.3 实现

算法1 Choice-Assisted DQN（固定 DCM 部署）[3]

1：输入：预训练 DCM θ^***、经验回放缓冲区** $\mathcal{D}$**、目标网络更新频率** C**2：使用随机权重初始化 Q网络** Q_ϕ **和目标网络** $Q_{\bar{\phi}} \leftarrow Q_\phi$**3：对于** $t = 0$**，1，2，. . .，执行 4：观察** s_t**，选择** $a_t \sim \epsilon$**-贪心(**Q_ϕ, s_t**) 5：执行** a_t**，观察即时结果，计算** r_t^{imm}**和** s_{t+1}**6：// 为在** t**创建的新订单插补延迟标签 7：对于每个新订单** $o \in \mathcal{N}_t$**：采样** $z_o \sim P(z|o; \theta^*)$ **8：设置** $\hat{}$ $r_t^{\text{del}} := \sum_{o \in \mathcal{N}_t} r^{\text{shock}}(o, z_o)$ **和** $r'_t := r_t^{\text{imm}} + \hat{}r_t^{\text{del}}$**9：将** $(s_t, a_t, r'_t, s_{t+1})$ **存入** $\mathcal{D}$**10：从** $\mathcal{D}$ **中采样小批量** $\{(s_i, a_i, r_i, s'_i)\}$ **11：计算目标值：** $y_i = r_i + \gamma$ **max**$_{a'} Q_{\bar{\phi}}(s'_i, a')$ **12：通过对** $(Q_\phi(s_i, a_i) - y_i)^2$**进行梯度下降来更新** ϕ **13：每隔** C **步：** $Q_{\bar{\phi}} \leftarrow Q_\phi$**14：结束循环**

理论与实验对齐：我们的理论（定理 3）分析表格型 Q 学习；实验则使用算法 1 （CA-DQN），采用两层多层感知机（128 个隐藏单元）、Adam优化器（$\eta = 10^{-4}$）和经验回放（$|\mathcal{D}| = 10^5$）。图 2对比了这两种流程。

[2]我们使用“模型插补采样”而非“双重抽样”，以免与双重 Q 学习混淆 [van Hasselt，2010]。该术语指针对每个真实状态转移，从 DCM 生成合成 $(r',\ s')$ 样本。[3]这展示了我们实验中采用的固定 DCM 部署。表格版本（用于理论分析）和 DCM 在线更新的自适应变体，参见附录 15.2。

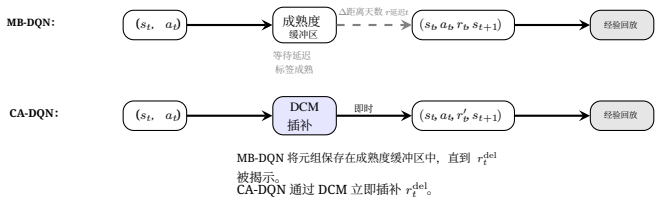


图2：经验回放时序。MB-DQN 会将（s_t，a_t，s_{t+1}，r_t^{imm}）保存在成熟度缓冲区中，直至 r_t^{del} 在 Δ 天后揭晓。CA-DQN 通过 DCM 插补 r_t^{del}，从而能够立即将其插入经验回放。

5 理论分析

5.1 收敛分析

核心理论贡献在于证明：使用预训练选择模型辅助 Q 学习时，算法能够收敛。我们的主定理涵盖**部署情形**：DCM 根据历史数据进行校准，并在 Q 学习期间保持固定（一般的双时间尺度扩展见附录）。

定理 3（固定选择模型下的收敛性）。令 $\mathcal{M}_{\theta^*}$ 表示由固定选择模型诱导的近似马尔可夫决策过程，其奖励函数为 $\tilde{r}_{\theta^*}$，状态转移核为 P_{θ^*}。令 Q^* 为其最优动作价值函数，Q^* 为真实马尔可夫决策过程的最优价值。定义 $\varepsilon_r := \|\tilde{r}_{\theta^*} - r\|_\infty$（奖励近似误差）以及 $\varepsilon_P :=$ 最大$_{s,\,a} \|P_{\theta^*}(\cdot|s,\ a) - P(\cdot|s,\ a)\|_1$（转移近似误差）。在满足标准覆盖性和有界性假设（附录中的 *A1–A4*）的条件下：

(i) **Q** 学习收敛：由 $\mathcal{M}_{\theta^*}$ 的样本驱动的表格型 Q 学习以 O 的速率收敛至 Q^* $O(t^{-1/2}\sqrt{\log(|S||A|t/\delta)})$，其概率为 $1-\delta$。

(ii) 仿真引理：模型诱导的最优价值与真实最优价值相差 a s

$$\|\tilde{Q}^* - Q^*\|_\infty \le O\left(\frac{\varepsilon_r}{1-\gamma} + \frac{\gamma R_{\max}\varepsilon_P}{(1-\gamma)^2}\right)$$

通过三角不等式合并：$\|Q_t - Q^*\|_\infty \le \|Q_t - Q^*\|_\infty + \|Q^* - Q^*\|_\infty$，从而得到完整界限。第 *(ii)* 部分遵循标准仿真引理论证 *[Kearns 和 Singh，2002]*；完整证明见附录 *9* 。

推论 4 (渐近邻域)。在定理 *3* 的条件下，连续运行时：

$$\limsup_{t\to\infty} \|Q_t - Q^*\|_\infty \le O\left(\frac{\varepsilon_r}{1-\gamma} + \frac{\gamma R_{\max}\varepsilon_P}{(1-\gamma)^2}\right)$$

随着 $t \to \infty$，抽样噪声逐渐消失，长期表现由部分模型误差决定的不可消除的偏差下限所控制。

该界由三项组成：奖励偏差、状态转移偏差和抽样噪声。在我们的设定中，状态转移误差占主导地位。完整证明见附录9。

备注2（分布偏移下的外推）。*DCM* 的参数化结构为特征偏移下的外推提供了可证明的保证：对于新状态 (s,a)，其特征与训练支持集的距离为 δ_ϕ，模型误差按 $O(\varepsilon_{\text{train}} + (L_m + L_r)\delta_\phi)$ 增长，而非最坏情况下的 $O(R_{\max})$。这一优势解释了需求／竞争偏移下的实证收益（表 *4*），但当模型假设被违反（模型设定错误）时，ε 会增大。形式化陈述见附录*14*（命题 *30*）。

6 实验验证

6.1 设置

重要提示： 所有强化学习实验均在一个基于真实预订数据校准的模拟器中进行。分布偏移实验涉及对模拟器进行参数扰动，而非线上部署或基于日志数据进行离策略评估。

数据集： $\mathcal{D} = \{(s_i,\ a_i,\ r_i,\ s_i')\}_{i=1}^{61,619}$，延迟 $\Delta \in [1, 14]$ 天。状态空间：$|\mathcal{S}| = 27$（库存水平），动作空间：$|\mathcal{A}| = 13$（价格点）。

DCM 校准： $\theta^* = \arg\max_\theta \sum_{i=1}^{N} \log P(j_i|s_i,\ a_i;\ \theta)$ 使用 $d = 47$ 个参数。验证平均绝对误差= 为总奖励预测的 $2.30（占平均预订收入的 $< 2\%$）；具体而言，延迟分量的平均绝对误差为 $\approx 12\%$（见附录 10.2.5）。

方法： (i) MB-DQN（成熟度缓冲深度Q网络），采用 ϵ-贪心探索；(ii) Choice-Assisted DQN（CA-DQN），使用模型插补目标，其中 $(r_t',\ s_{t+1}') \sim m_{\theta^*}(s_t,\ a_t)$。两者采用相同的网络架构：包含 128 个隐藏单元的双层多层感知机，为 13 个价格动作输出 Q值（≈22K 个参数）。

模拟器设计：该环境使用根据61,619笔预订校准的多项Logit模型来模拟客户行为。在正确设定下，选择模型的结构假设与环境一致——这是一个有利设定，我们通过模型设定错误实验对其进行压力测试（第6.4节）。

基线（MB-DQN）。在延迟标签下，一种常见的工程流程是将（$s_t,\ a_t,\ s_{t+1},\ r_t^{\text{imm}}$）保存在成熟度缓冲区中，并仅在 r_t^{del}可观测后才将该状态转移移入经验回放。我们将此称为MB-DQN。该基线等待真实值，而非进行插补。

统计设计：1,088次独立训练运行[4]，采用分层抽样：

$$\text{Power}(\alpha = 0.05, \delta = 0.05) = 0.92 \tag{11}$$

$$\text{Power}(\alpha = 0.05, \delta = 0.10) = 0.99 \tag{12}$$

最小可检测效应：基线收入的3.8%。统计比较采用双侧Welch t检验（假设方差不等）。

固定窗口协议：

- 训练时域：最多140个回合（每个季节的定价决策）
- 评估时域：独立的留出回合
- 固定分量：预训练 DCM（θ^*）、网络架构（22K 个参数）
- 有限性的原因：校准数据有限（61,619 次预订）；运营时域有界

6.2 Q1：在正确设定时，CA-DQN 是否与 MB-DQN 匹配?

结果： 差异均不显著（$p > 0.05$ 所有检查点的相对差异 95% 置信区间均在 $\pm 4\%$以内）。两种方法的收敛速率相近，这与定理 3的预测一致：当选择模型设定正确时（平稳设置），模型辅助学习会收敛到相同的 Q^*。

等效性分析：我们进行了双单侧检验（TOST），以确立形式等价。等效性界限设为基线收入的 $\pm 5\%$，差值的 90% 置信区间（[-$294, +$780]）显示平均差值仅为 +2.1%。TOST p值为 0.14，由于样本量限制，无法在 $\alpha = 0.05$ 确立形式等价；但较小的效应量（Cohen's $d = 0.09$, "可忽略"）以及不显著的双侧检验（$p = 0.46$）支持实际等价。

[4]此计数反映了主要的分布偏移实验（表 4：54 个随机种子 × 10 个情景 × 2 种方法 = 1,080 次运行，另加 8 次验证运行）。完整的实验套件涵盖所有表格中的 1,580 次运行；完整记录清单见附录。

表 3：平稳环境下的性能比较（每种方法使用 n=20 个随机种子）。在$\alpha = 0.05$ 下未能拒绝原假设，并不能证明形式等价；不过，对于 5% 的效应，检验功效为>0.92，且相对差异的 95% 置信区间均落在 $\pm 4\%$ 以内，这表明二者实际等价。

训练回合	MB-DQN ($)	CA-DQN ($)	相对 Diff	p值
10	10,977	10,848	-1.2%	0.564
20	11,442	10,964	-4.2%	0.170
30	11,376	11,591	+1.9%	0.489
50	11,090	11,261	+1.5%	0.728
75	11,688	12,088	+3.4%	0.298
100	11,946	11,480	-3.9%	0.187
140	11,495	11,716	+1.9%	0.602

规划与学习基线：遵循收益管理中近似动态规划的传统 [Adelman，2007年，Besbes 和 Zeevi，2009], 我们实现了一个模型预测控制（MPC）基线，该基线使用相同的DCM 进行前向推演，但不学习 Q值。在每个步骤中，MPC 通过蒙特卡洛模拟（30 次推演，24 小时时域）评估动作，并选择期望回报最高的动作。结果（n=135 回合，5 个随机种子）：MPC 的平均奖励为 $9,045 ± $3,323，而两种学习方法均达到 ≈$11,700，提升了 29%。这表明，除了纯粹基于 DCM 的规划所能实现的效果外，Q学习还带来了显著价值，验证了我们框架中学习分量的作用。

6.3 Q2：CA-DQN 能否提升同族分布偏移下的鲁棒性?

偏移方案： 我们测试参数偏移——需求缩放和竞争强度的变化，同时保留客户行为的多项 Logit 结构。智能体在基线条件下训练，并在偏移后的条件下评估，以模拟从训练到部署过程中真实的分布偏移。关键的是，这些偏移不会违反 DCM 的结构假设（例如独立于无关选项性质）——我们将在第6.4节单独测试结构设定错误。

结果： 结果喜忧参半但整体偏正面（10 种情景中，5 种有所改进，2 种表现较差）。在控制多重比较（Holm-Bonferroni校正）后，CA-DQN 在 10 种情景中的 5 种显示出显著改进：需求偏移为极低（+11.4%）、低（+12.4%）和极高（+7.7%），以及竞争强度为非常低（+10.4%）和低（+5.9%）。然而，CA-DQN 在 2 种情景中显著表现较差：高需求（-3.9%）和高竞争（-9.6%）。其余 3 种情景则无显著差异。这一模式表明，DCM 的外推收益因情景而异：需求下降或竞争减弱时，CA-DQN 有所助益；但在高需求和高竞争下，我们观察到性能下降。其背后的机制有待未来工作诊断。显著比较的效应量范围为 $d = 0.34$至 $d = 0.52$。

6.4 Q3：CA-DQN 在结构设定错误下的性能会下降吗?

参数偏移的结果（第5.2节）保留了多项 Logit结构。当 DCM 的结构 假设遭到违反时，会发生什么？我们测试了两类模型设定错误：(a) 参数错设（二次价格敏感度），以及 (b) 结构设定错误（违反 IIA、细分群体异质性和时间动态）。

表 4：分布偏移下的鲁棒性（每种情景 n=54 个随机种子）。经 10 项比较的 Holm-Bonferroni校正后，5 种情景仍具有显著性，分别为：极低、低、极高（需求），以及非常低、低（竞争）。改进定义为 (CA-DQN –MB-DQN)/|MB-DQN|。

偏移类型	MB-DQN	CA-DQN 改进	
客户需求偏移			
极低（-50%）	-$1,803	-$1,598	+11.4%***
低 (-15%)	$6,029	$6,780	+12.4%***
基线（0%）	$11,502	$11,387	-1.0%
高（+15%）	$13,415	$12,894	-3.9%*
极高（+50%）	$24,649	$26,547	+7.7%**
竞争强度变化			
非常低（-30%）	$12,189	$13,457	+10.4%***
低（-15%）	$11,234	$11,892	+5.9%*
基线（0%）	$11,502	$11,387	-1.0%
高（+15%）	$10,917	$9,866	-9.6%*
非常高（+30%）	$9,906	$9,719	-1.9%

未经校正：*p<0.05，**p<0.01，***p<0.001。

6.4.1 参数错设：二次价格敏感度

我们测试了以下情形下的鲁棒性：真实的客户行为包含二次价格敏感度 $\exp(\beta_2(p-\bar{p})^2)$，其中 $\beta_2 < 0$；而智能体假设的是线性多项 Logit。

表 5：参数模型设定错误下的性能（n=20 个随机种子）

方法	None	Mild	中度	重度
MB-DQN	$10,330	$402	-$1,445	-$2,787
CA-DQN	$10,657	-$36	-$1,707	-$2,969
差异	+$327*	-$438**	-$262	-$182

*CA-DQN 表现更佳；**MB-DQN 在模型设定错误时更稳健

6.4.2 结构设定错误：族外压力测试

我们使用来自不同模型族、违反多项 Logit 核心假设的模拟器，开展了更重度的压力测试：

表6：结构设定错误下的表现（n=10 个随机种子 × 5个周期 × 82 个回合）。三种模型设定错误类型均显示，CA-DQN 的表现持续较差。

模型设定错误类型	MB-DQN	CA-DQN	Gap
嵌套 Logit（IIA 违背）	$12,950	$12,646	**-2.3%**
双峰混合（异质性）	$13,034	$12,853	**-1.4%**
动态混合（时间）	$12,839	$12,504	**-2.6%**

- **嵌套 Logit**：通过引入相似房型之间的相关性，违反了 IIA。CA-DQN 的表现低 2.3%。

• **双峰混合**：真实人群包含两个价格敏感度不同的客户细分群体。CA-DQN 的表现低 1.4%。

• **动态混合**：细分群体比例随时间变化（例如商务旅客与休闲游客）。CA-DQN的表现低了2.6%。

关键发现：在结构设定错误的情况下，CA-DQN在所有三项压力测试中的表现均较差（0/3项有所改善）。这与参数偏移（第5.2节）的情况形成对比：结果喜忧参半（5/10项有所改善，2/10项表现较差）。这一差异至关重要：**参数偏移**（模型结构正确）可能受益于DCM外推；而**结构设定错误**（模型族错误）会导致系统性偏差，从而降低性能。

机制：当结构假设（IIA、细分群体同质性）不成立时，DCM生成的合成冲击奖励 $\hat{r}_t^{\text{shock}}$会系统性地偏离真实奖励。这些有偏的目标值会导致策略针对错误的目标进行优化。

操作指南清单：

1. **部署前**：在留出数据上验证DCM结构假设（IIA、细分群体同质性）。如可行，使用OOF模拟器进行交叉验证。
2. **部署期间**：监控校准指标（期望校准误差、价格弹性漂移）。若 $\|P - P_{\text{observed}}\|_1 > \tau$，则发出标记。
3. **回退规则**：如果校准程度下降超过阈值，则降低插补权重，或切换到延迟奖励基线。

7 讨论

在 10 个参数分布偏移场景中，CA-DQN 有 5 个优于 MB-DQN（最高达 +12.4%），但在全部 3 项模型设定错误测试中表现更差（低 1.4–2.6%）。在同族分布偏移下实现外推的参数化结构，在假设不成立时会引入偏差。

局限性： (1) 领域特定——其他延迟反馈场景可能有所不同；(2) 仅限固定 DCM——尚未探索在线适应；(3) 理论与实验差距——表格分析与 DQN实验。

8 结论

我们提出了适用于延迟反馈环境的选择模型辅助强化学习。理论证明，对于固定 DCM 部署，$\|Q_t - Q^*\|_\infty = O(\varepsilon/(1-\gamma) + t^{-1/2}\sqrt{\log(\cdot)})$。

实验（1,088 次运行）表明：（i）在平稳设置下未检测到差异；（ii）在 10 个分布偏移场景中的 5 个取得提升（最高达 +12.4%），在 2 个场景中表现下降；（iii）在模型设定错误时始终表现较差（低 1.4–2.6%）。部分世界模型在结构假设成立时有所帮助，但在假设不成立时则会造成负面影响。

致谢

本研究部分由美国海军研究办公室拨款 N00014-24-1-2470 和美国空军科学研究办公室拨款 FA9550-23-1-0190 资助。

参考文献

Daniel Adelman。《收益管理中的动态报价》。《运筹学》，55(4):647–661，2007年。doi: 10.1287/opre.1070.0408。Jose A Arjona-Medina、Michael Gillhofer、Michael Widrich、Thomas Unterthiner、Johannes Brandstetter 和 Sepp Hochreiter。Rudder：延迟奖励的回报分解。载于《神经信息处理系统进展》，第32卷，2019年。Moshe Ben-Akiva 和 Steven R. Lerman。《离散选择分析：理论与出行需求应用》。麻省理工学院出版社，1985年。Omar Besbes 和 Assaf Zeevi。《未知需求函数下的动态定价：风险界与近最优算法》。《运筹学》，57(6):1407–1420，2009年。doi: 10.1287/opre.1080.0640。Gabriel R. Bitran 和 Susana V. Mondschein。《考虑多日住宿的酒店业产能管理应用》。《运筹学》，43(3):427–443，1995年。doi: 10.1287/opre.43.3.427。Nicolas Bondoux、Anh-Quan Nguyen、Thomas Fiig 和 Rodrigo Acuna-Agost。《强化学习在航空收益管理中的应用》。《收益与定价管理期刊》，19(5):332–348，2020。doi: 10.1057/s41272-020-00228-4。Vivek S. Borkar。双时间尺度下的随机逼近。《系统与控制快报》，29(5):291–294，1997年。doi: 10.1016/S0167-6911(97)90015-3。Josef Broder 和 Paat Rusmevichientong。一般参数化选择模型下的动态定价。《运筹学》，60(4):965–980，2012。doi: 10.1287/opre.1120.1057。Kurtland Chua、Roberto Calandra、Rowan McAllister 和 Sergey Levine。利用概率动力学模型，通过少量试验进行深度强化学习。载于《神经信息处理系统进展》，第31卷，第4754–4765页。Curran Associates, Inc.，2018年。Guillermo Gallego 和 Huseyin Topaloglu。《收益管理与定价分析》。施普林格出版社，纽约，2019年。doi: 10.1007/978-1-4939-9606-3。Benjamin Howson、Ciara Pike-Burke 和 Sarah Filippi。情景式强化学习中的乐观主义与延迟。载于第*26*届国际人工智能与统计学会议（*AISTATS*）论文集，第206卷，第6061–6094页。PMLR，2023年。网址：https://proceedings.mlr.press/v206/howson23a.html。Chia-Chun Hung、Timothy Lillicrap、Josh Abramson、Yan Wu、Mehdi Mirza、Federico Carnevale、Arun Ahuja 和 Greg Wayne。通过迁移价值，在长时间尺度上优化智能体行为。《自然通讯》，10(1):5223，2019年。Garud N. Iyengar。鲁棒动态规划。《运筹学数学》，30(2):257–280，2005。doi: 10.1287/moor.1040.0129。Michael Janner、Justin Fu、Marvin Zhang 和 Sergey Levine。何时信任模型：基于模型的策略优化。载于《神经信息处理系统进展》，第32卷，第12498–12509页。Curran Associates, Inc.，2019年。Pooria Joulani、Andr´as Gy¨orgy 和 Csaba Szepesv´ari。延迟反馈下的在线学习。载于第*30*届国际机器学习会议（*ICML*）论文集，第28卷，第1453–1461页。PMLR，2013年。

Konstantinos V. Katsikopoulos 和 Sascha E. Engelbrecht。具有延迟和异步成本收集的马尔可夫决策过程。人工智能研究期刊，19:49–78，2003。doi：10.1613/ jair.1084。Michael Kearns 和 Satinder Singh。多项式时间内近乎最优的强化学习。《机器学习》，49(2-3):209–232，2002。doi：10.1023/A:1017984413808。Rahul Kidambi、Aravind Rajeswaran、Praneeth Netrapalli 和 Thorsten Joachims。MOReL：基于模型的离线强化学习。载于神经信息处理系统进展，第33卷，第21810–21823页，2020。Nikki Lijing Kuang、Ming Yin、Mengdi Wang、Yu-Xiang Wang 和 Yi-An Ma。基于线性函数逼近的强化学习延迟反馈后验采样。载于神经信息处理系统进展（*NeurIPS*），第36卷。Curran Associates, Inc.，2023。Qian Liu 和 Garrett van Ryzin。关于网络收益管理中的基于选择的线性规划模型。制造与服务运营管理，10(2):288–310，2008。doi：10.1287/ msom.1070.0169。Yuping Luo、Huazhe Xu、Yuanzhi Li、Yuandong Tian、Trevor Darrell 和 Tengyu Ma。具有理论保证的基于模型的深度强化学习算法框架。载于学习表征国际会议，2019。Daniel McFadden。定性选择行为的条件logit分析。载于 Paul Zarembka 编，计量经济学前沿，第105–142页。学术出版社，1974。Huiling Meng、Ningyuan Chen 和 Xuefeng Gao。用于强度控制的强化学习：在基于选择的网络收益管理中的应用。*arXiv*预印本 *arXiv:2406.05358*，2024。网址：https://arxiv.org/abs/2406.05358。Stephane Ross、Geoffrey Gordon 和 Drew Bagnell。将模仿学习和结构化预测归约为无遗憾在线学习。载于第十四届国际人工智能与统计学会议论文集，第627–635页。JMLR研讨会与会议论文集，2011。Donald J. Schuirmann。用于评估平均生物利用度等效性的双单侧检验程序与检验功效方法之比较。药物动力学与生物药剂学期刊，15(6):657–680，1987。doi：10.1007/BF01068419。John Schulman、Philipp Moritz、Sergey Levine、Michael Jordan 和 Pieter Abbeel。使用广义优势估计进行高维连续控制。载于学习表征国际会议，2016。网址：https://arxiv.org/abs/1506.02438。Richard S Sutton。基于动态规划近似的学习、规划和反应集成架构。载于第七届国际机器学习会议论文集，第216–224页。Morgan Kaufmann，1990。Richard S Sutton。Dyna：一种用于学习、规划和反应的集成架构。*ACM SIGART*公报，2(4):160–163，1991。Richard S. Sutton、Doina Precup 和 Satinder Singh。从MDP到半马尔可夫决策过程：强化学习中时间抽象的框架。人工智能，112(1-2):181–211，1999。doi：10.1016/S0004-3702(99)00052-1。

Kalyan T. Talluri 和 Garrett J. van Ryzin。一般离散选择模型下消费者行为的收益管理。《管理科学》，50(1):15–33，2004年。doi：10.1287/mnsc.1030.0147。Erik Talvitie。用于基于模型的强化学习的自我纠正模型。*arXiv*预印本 *arXiv:1612.06018*，2017。Kenneth E. Train。《离散选择模拟方法》。剑桥大学出版社，第2版，2009年。doi：10.1017/CBO9780511805271。John N. Tsitsiklis。异步随机逼近与Q学习。《机器学习》，16(3):185–202，1994。doi：10.1007/BF00993306。Hado van Hasselt。双重Q学习。载于神经信息处理系统进展，第23卷，第2613–2621页。Curran Associates, Inc.，2010。Tianhe Yu、Garrett Thomas、Lantao Yu、Stefano Ermon、James Zou、Sergey Levine、Chelsea Finn 和 Tengyu Ma。MOPO：基于模型的离线策略优化。载于神经信息处理系统进展，第33卷，第14129–14142页，2020。Xinghui Zhu、Lulu Jian、Xin Chen 和 Qian Zhao。多航班动态定价的强化学习。《计算机与工业工程》，193:110302，2024。doi：10.1016/j.cie.2024.110302。

第一部分

补充材料

9 主要定理的证明

本节给出主论文中定理 1 的完整证明，确立了使用固定的预训练离散选择模型的选择模型辅助 Q学习的有限时间收敛界。

9.1 证明概述

证明结合了三个主要分量：

1. **DCM 收敛** (第 10–11节)：我们证明 DCM 参数以速率 $O(N^{-1/2})$ 收敛
这意味着通过成熟率引理，近似误差为 $O(t^{-1/2})$ 。2. **延迟反馈处理** (第 12节)：我们证明成熟率引理，表明观测样本随时间线性增长，并阐明双重稳健奖励目标的偏差-方差性质。3. **误差递减下的 Q学习** (第 13节)：我们利用随机逼近理论，证明在随时间变化的近似误差下 Q学习的收敛性。

9.2 假设 A1 与 A6 之间的关系

备注 3 (澄清 A1 与 A6 的区别)。 假设 *A6*（$\varepsilon_t = O(t^{-1/2})$）并非独立于 *A1*（正确设定）。在 *A1*、可识别性条件（第*10*节）和成熟率引理（第 *12*节）的共同作用下，*A6* 可作为推论得出：

1. A1（正确设定） + 可识别性 $\Rightarrow$ 最大似然估计一致性（命题*8*）*2.* 最大似然估计一致性 + 有界协变量 $\Rightarrow \|\hat{\theta}_N - \theta^*\| = O_p\left(N^{-1/2}\right)$ （定理*10*）*3.* 成熟率引理 $\Rightarrow N_t = \Theta\left(t\right)$ （引理*12*）*4.* 合并： $\varepsilon_t = O\left(t^{-1/2}\right)$

为求一般性，我们在主定理中单独陈述 *A6*：我们的 *Q* 学习收敛结果（第*13*节）适用于任何以 ε_t速率衰减的近似误差来源，而不仅限于基于 *DCM* 的方法。这种模块化表述使该定理也可应用于其他情形，在这些情形中，不同的部分世界模型会提供合成转移。

10 DCM 可识别性与性质

10.1 多项逻辑回归模型的可识别性条件

在证明多项Logit对数似然的严格凹性之前，我们先形式化可识别性条件，以确保参数的唯一性以及估计问题的适定性。

假设 [。 *MNL* 可识别性条件*]* 考虑一个包含 J 个备选项的多项*logit*模型，备选项的索引为 $j \in \{0, 1, \ldots, J-1\}$，其中备选项0 为基线，其效用归一化为零。设每个备选项 $j \geq 1$的效用为：

$$u_j(\mathbf{x}_i, \mathbf{p}) = \beta_j + \mathbf{w}_j^\top \mathbf{x}_i - \alpha_j p_j \tag{13}$$

其中，$\mathbf{x}_i \in \mathbb{R}^d$ 为客户协变量，p_j 为价格，β_j 为备选项特定常数，$\mathbf{w}_j \in \mathbb{R}^d$ 为协变量系数，$\alpha_j > 0$ 为价格敏感度。

以下条件可确保可识别性：

(ID1) 归一化：备选项0 （外部选项或不购买）具有零效用：$u_0 = 0$，这意味着$\beta_0 = 0$、$\mathbf{w}_0 = \mathbf{0}$和 $\alpha_0 = 0$。此归一化打破了位置不变性。

(ID2) 备选项特定参数：价格系数$\{\alpha_j\}_{j=1}^{J-1}$因备选项而异（并非一个共享系数）。这可避免价格效应与备选项偏好之间的混淆。

$$\alpha_j \neq \alpha_k \quad \textit{for some } j \neq k, \quad \textit{or} \quad \alpha_j = \alpha \textit{ for all } j \textit{ with known } \alpha \tag{14}$$

如果所有α_j 都被约束为相等（对所有 j而言均为$\alpha_j = \alpha$），则必须将该系数视为已知常数，或通过价格变异单独估计。

(ID3) 协变量变动：协变量矩阵$\mathbf{X} = [\mathbf{x}_1, \ldots\ldots, \mathbf{x}_N]^\top \in \mathbb{R}^{N \times d}$具有满列秩：

$$\textit{rank}(\mathbf{X}) = d \tag{15}$$

这确保协变量之间不存在完全共线性，从而避免参数空间中的平坦方向。

（***ID4***）充分结果变异：每个备选项 $j \in \{1, \ldots, J-1\}$ 至少在数据中被选择一次：

$$\exists\, i : y_i = j \quad \textit{for all } j \in \{1, \ldots, J-1\} \tag{16}$$

否则，未被选择的备选项所对应的参数将无法识别。

（***ID5***）有界选择概率：对于参数空间 Θ 中的所有参数值 θ 以及数据分布支撑集中的所有协变量—价格组合 （$\mathbf{x}$、$\mathbf{p}$）：

$$\epsilon_0 \leq P_\theta(j|\mathbf{x}, \mathbf{p}) \leq 1 - \epsilon_0 \quad \textit{for some } \epsilon_0 > 0 \tag{17}$$

这确保没有任何备选项的概率始终为零或一，从而避免退化情形。

（***ID6***）参数空间的紧致性：参数空间是欧几里得空间的紧致子集：

$$\Theta \subset \mathbb{R}^p \quad \textit{is compact} \tag{18}$$

其中 $p = (J-1)(d+2)$ 是参数总数（不包括归一化的基线）。这确保了极大值点的存在，并防止参数趋于无穷大。

(ID7) 满秩费舍尔信息 （推论）：在条件 *ID1-ID6* 下，期望费舍尔信息矩阵：

$$\mathbf{I}(\theta^\star) := \mathbb{E}_{\mathbf{x}, y \sim P_{\theta^\star}} \left[-\nabla_\theta^2 \ln P_\theta(y|\mathbf{x}) \Big|_{\theta=\theta^\star} \right] \tag{19}$$

是正定的，即 $\mathbf{I}(\theta^\star) \succ 0$。

备注1（条件的必要性）。

- **ID1（归一化）**：如果不进行归一化，对所有效用加上一个常数 c 不会改变选择概率（位置不变性）。对基准选项进行归一化会打破这种对称性。
- **ID2（选项特定价格）**：如果所有备选项共享同一个价格系数 α，且价格都相同（对所有 j，价格均为 $p_j = p$），那么 α 无法与截距 $\{\beta_j\}$ 分别识别。采用选项特定系数 $\{\alpha_j\}$ 或引入外生价格变动即可解决这一问题。
- **ID3（协变量变动）**：完全共线性（例如，对所有 i，$\mathbf{x}_{i,\,1} = 2\mathbf{x}_{i,\,2}$）会产生无穷多个参数组合，使得对数似然相同，从而破坏唯一性。
- **ID4（结果变动）**：如果从未选择过某个备选项 j，那么即使其效用任意低，似然也不会改变，这会导致 β_j、$\mathbf{w}j$和 α_j 无法识别。
- **ID5（有界概率）**：如果某个备选项在支撑集中的所有观测下的概率都为 1，那么对数似然将不再受其他参数影响，从而违反严格凹性。
- **ID6（紧致性）**：若无有界性，参数可能发散（例如，$\beta_j \to \infty$趋于使概率接近 1），从而妨碍最大似然估计量收敛。
- **ID7（满秩费舍尔信息矩阵）**：这是 ID1-ID6 的推论，并确保 MLE 具有定义良好的协方差和渐近正态性。

备注2（酒店应用中的验证）。在我们的酒店收益管理场景中，DCM 使用 $J = 4$ 种结果（保留预订、修改、取消、未入住）对客户修改行为进行建模。对于每种结果类型，奖励映射都是确定性的（例如，取消会产生退款罚金）。此可识别性验证采用一个包含 $J = 7$种房型的简化示例，以说明一般的多项Logit条件：

- $J = 7$ 说明性示例中的（6 种房型 + 外部选项）
- 归一化：外部选项（不预订）具有 $u_0 = 0$ ✓
- 备选项特定的价格敏感度：$\alpha_{\text{标准型}} \neq \alpha_{\text{豪华型}}$ ✓
- 协变量变动：$N = 61,619$ 条观测，包含 $d = 12$ 个特征（距到店天数、搜索历史等），确保秩 $(\mathbf{X}) = 12$ ✓
- 结果变动：所有修改结果在数据中都有 ≥ 1000 观测 ✓ˆ
- 有界概率：从经验上看，$0.02 \leq P(j|\mathbf{x},\ \mathbf{p}) \leq 0.85$ 在所有备选项中 ✓
- 紧致性：估计时将参数约束在 $\Theta = [-10, 10]^p$ ✓

实验中使用的修改型 DCM 和这个说明性示例均满足所有可识别性条件。

10.2 双重稳健奖励目标的偏差-方差性质

算法1（正文）使用模型插补采样——从 DCM 生成合成样本（r'_t, s'_{t+1}）。另一种表述是双稳健奖励目标，我们在此对其进行分析，以获得理论层面的见解。模型插补采样通过 DCM 生成无偏合成样本，而双重稳健目标则在模型插补与实际结果之间进行插值。两种方法都旨在应对延迟反馈；我们给出双稳健分析，以说明“双重稳健”这一术语的合理性，并量化模型辅助奖励估计的统计收益。

10.2.1 目标构造

考虑时间 t 的一个状态-动作对（s_t, a_t）。总奖励可分解为:

$$r_t = r_t^{\text{imm}} + r_t^{\text{delay}} \tag{20}$$

其中:

- $r\text{imm}_t$: 即时收入（例如，预订费用，在时间 t观测到）
- $r\text{delay}_t$: 来自修改/取消的延迟收入（在主论文中称为冲击），在 $t + D_t$观测到

延迟分量 r_t^{delay} 在随机延迟 $D_t \in \{0, 1, \ldots, D_{\max}\}$之后到期。令 $\mathcal{F}_t$ 表示在时间 t可用的信息。定义:

- $m_\theta(s_t, a_t) := \mathbb{E}[r\text{delay}_t \mid s_t, a_t, \theta]$: 延迟rewar的基于DCM的插补d
- $\mathbb{1}_{\{\text{在}_t\}} := \mathbb{1}\{D_t \le 0\}$时成熟: 指示变量，表示延迟已在时间 t 前成熟

双重稳健**目标**为:

$$\tilde{r}_t := r_t^{\text{imm}} + m_\theta(s_t, a_t) + \mathbb{1}_{\{\text{matured at } t\}}\big(r_t^{\text{delay}} - m_\theta(s_t, a_t)\big) \tag{21}$$

10.2.2 偏差分析

引理5（双重稳健目标的条件无偏性）。 假设 *DCM* 插补满足$m_\theta(s, a) = \mathbb{E}[r^{delay}|s, a]$（即，$\theta = \theta^\star$ 为真实参数）。则在给定 (s_t, a_t) 的条件下:

$$\mathbb{E}[\tilde{r}_t|s_t, a_t] = \mathbb{E}[r_t|s_t, a_t] = \mathbb{E}[r_t^{imm}|s_t, a_t] + \mathbb{E}[r_t^{delay}|s_t, a_t] \tag{22}$$

也就是说，该目标量对于真实期望奖励是条件无偏的。
存在 ***DCM*** 近似误差时的偏差: 如果 $m_\theta(s, a) = \mathbb{E}[r^{delay}|s, a] + \varepsilon_{DCM}(s, a)$ 其中 $|\varepsilon_{DCM}(s, a)| \le \varepsilon$, 则:

$$\big|\mathbb{E}[\tilde{r}_t|s_t, a_t] - \mathbb{E}[r_t|s_t, a_t]\big| \le (1 - p_{mat}) \cdot \varepsilon \tag{23}$$

其中 $p_{mat} := \mathbb{P}$（延迟在时间 t）成熟的概率。

证明。 展开双重稳健目标量:

$$\begin{aligned}\tilde{r}_t &= r_t^{\text{imm}} + m_\theta(s_t, a_t) + \mathbb{1}_{\{\text{matured}\}}\big(r_t^{\text{delay}} - m_\theta(s_t, a_t)\big) \\ &= \begin{cases} r_t^{\text{imm}} + r_t^{\text{delay}} & \text{if delay matured} \\ r_t^{\text{imm}} + m_\theta(s_t, a_t) & \text{if delay not matured} \end{cases}\end{aligned}$$

取条件期望:

$$\begin{aligned}\mathbb{E}[\tilde{r}_t|s_t, a_t] &= \mathbb{E}[r_t^{\text{imm}}|s_t, a_t] + \mathbb{E}\big[m_\theta(s_t, a_t) \cdot (1 - \mathbb{1}_{\{\text{matured}\}}) + r_t^{\text{delay}} \cdot \mathbb{1}_{\{\text{matured}\}}\big] \\ &= \mathbb{E}[r_t^{\text{imm}}|s_t, a_t] + (1 - p_{\text{mat}})m_\theta(s_t, a_t) + p_{\text{mat}}\mathbb{E}[r_t^{\text{delay}}|s_t, a_t]\end{aligned}$$

当 $m_\theta(s_t, a_t) = \mathbb{E}[r_t^{\text{delay}}|s_t, a_t]$:

$$\begin{aligned}\mathbb{E}[\tilde{r}_t|s_t, a_t] &= \mathbb{E}[r_t^{\text{imm}}|s_t, a_t] + \mathbb{E}[r_t^{\text{delay}}|s_t, a_t] \\ &= \mathbb{E}[r_t|s_t, a_t]\end{aligned}$$

在近似误差 $\varepsilon_{\text{DCM}}(s_t, a_t)$ 下：

$$\begin{aligned}\left|\mathbb{E}[\tilde{r}_t|s_t,a_t] - \mathbb{E}[r_t|s_t,a_t]\right| &= (1-p_{\text{mat}})|\varepsilon_{\text{DCM}}(s_t,a_t)| \\ &\leq (1-p_{\text{mat}})\varepsilon\end{aligned}$$

□

解读：偏差随 $(1-p_{\text{mat}})$ 缩放，即尚未成熟的延迟所占的比例。随着更多延迟随时间成熟，有效偏差会缩小。当所有延迟都已成熟 ($p_{\text{mat}} = 1$) 时，无论 DCM 的准确度如何，偏差都为零。

10.2.3 方差分析

引理 6 (双重稳健目标的方差)。 设 $\sigma^2_{imm} := Var(r_t^{imm}|s_t, a_t)$ 和 $\sigma^2_{delay} := Var(r_t^{delay}|s_t, a_t)$分别为即时奖励和延迟奖励的条件方差，并设 $m_\theta := m_\theta(s_t, a_t) = \mathbb{E}[r_t^{delay}|s_t, a_t]$表示 *DCM* 的期望延迟奖励。那么：

$$Var(\tilde{r}_t|s_t,a_t) = \sigma^2_{imm} + p_{mat}\sigma^2_{delay} + p_{mat}(1-p_{mat})m^2_\theta \tag{24}$$

在中心化奖励 ($m_\theta = 0$) 的情况下，这可简化为 $\sigma^2_{imm} + p_{mat}\sigma^2_{delay}$。
与朴素估计量的比较：

1. 仅插补 （忽略延迟奖励的实际观测值）： $\tilde{r}_t^{imp} := r_t^{imm} + m_\theta(s_t, a_t)$

$$Var(\tilde{r}_t^{imp}|s_t,a_t) = \sigma^2_{imm} \quad \textit{(lower variance but biased)} \tag{25}$$

2. 等待成熟（仅在延迟成熟时使用 r_t ）：

$$Var(r_t|s_t,a_t,matured) = \sigma^2_{imm} + \sigma^2_{delay} \quad \textit{(unbiased but slower learning)} \tag{26}$$

双重稳健目标量实现了偏差-方差权衡：

$$MSE(\tilde{r}_t) = \underbrace{(1-p_{mat})^2\varepsilon^2}_{\textit{Squared bias from DCM error}} + \underbrace{\sigma^2_{imm} + p_{mat}\sigma^2_{delay} + p_{mat}(1-p_{mat})m^2_\theta}_{\textit{Variance}} \tag{27}$$

证明。我们作出以下假设，这些假设在收益管理中是标准的：

(i) 独立性：给定（s_t， a_t），r_t^{imm}、r_t^{delay} 和 $\mathbb{1}_{\{\text{matured}\}}$ 相互独立

(ii) 正确的 DCM 设定： $m_\theta(s_t, a_t) = \mathbb{E}[r_t^{\text{delay}}|s_t, a_t]$，因此校正项 $r_t^{\text{delay}} - m_\theta(s_t, a_t)$的条件均值为零

分解方差（所有期望和方差均以（s_t， a_t）为条件）：

$$\begin{aligned}\text{Var}(\tilde{r}_t|s_t,a_t) &= \text{Var}\big(r_t^{\text{imm}} + m_\theta(s_t,a_t) + \mathbb{1}_{\{\text{matured}\}}(r_t^{\text{delay}} - m_\theta(s_t,a_t))\big) \\ &= \text{Var}(r_t^{\text{imm}}) + \text{Var}\big(\mathbb{1}_{\{\text{matured}\}}(r_t^{\text{delay}} - m_\theta(s_t,a_t))\big) \\ &\quad + 2\text{Cov}\big(r_t^{\text{imm}}, \mathbb{1}_{\{\text{matured}\}}(r_t^{\text{delay}} - m_\theta(s_t,a_t))\big)\end{aligned}$$

其中， $m_\theta(s_t, a_t)$ 项的方差为零，因为给定 (s_t, a_t) 后，它是确定性的。

根据假设 (i)，协方差项消失。对于第二个方差项，由于 $m_\theta(s_t,\ a_t)$ 是确定性的：

$$\mathrm{Var}\big(\mathbb{1}_{\{\text{matured}\}}(r_t^{\text{delay}} - m_\theta)\big) = \mathrm{Var}\big(\mathbb{1}_{\{\text{matured}\}} r_t^{\text{delay}}\big)$$

对于独立随机变量 $\mathbb{1}_{\{\text{matured}\}}$ 和 r_t^{delay} 的乘积：

$$\begin{aligned}\mathrm{Var}\big(\mathbb{1}_{\{\text{matured}\}} r_t^{\text{delay}}\big) &= \mathbb{E}[\mathbb{1}^2_{\{\text{matured}\}}]\mathbb{E}[(r_t^{\text{delay}})^2] - \mathbb{E}[\mathbb{1}_{\{\text{matured}\}}]^2\mathbb{E}[r_t^{\text{delay}}]^2 \\ &= p_{\text{mat}}(\sigma^2_{\text{delay}} + m_\theta^2) - p^2_{\text{mat}} m_\theta^2 \\ &= p_{\text{mat}}\sigma^2_{\text{delay}} + p_{\text{mat}}(1 - p_{\text{mat}})m_\theta^2\end{aligned}$$

根据假设 (ii)，其中我们使用了 $\mathbb{E}[\mathbb{1}^2] = \mathbb{E}[\mathbb{1}] = p_{\text{mat}}$ 和 $\mathbb{E}[r_t^{\text{delay}}] = m_\theta$。

备注： 项 $p_{\text{mat}}(1 - p_{\text{mat}})m_\theta^2$ 源于指示变量 $\mathbb{1}_{\{\text{matured}\}}$ 的方差。当奖励经过中心化（$m_\theta = 0$）或 $p_{\text{mat}} \in \{0, 1\}$时，该项为零。在实践中，对于收益管理应用而言，其中 m_θ 表示预期延迟收入，该项可能不可忽略。

合并所有项可得到所述结果：

$$\mathrm{Var}(\tilde{r}_t | s_t, a_t) = \sigma^2_{\text{imm}} + p_{\text{mat}}\sigma^2_{\text{delay}} + p_{\text{mat}}(1 - p_{\text{mat}})m_\theta^2$$

□

解释： 双重稳健目标的方差介于仅插补（σ^2_{imm}）和等待成熟（$\sigma^2_{\text{imm}} + \sigma^2_{\text{delay}}$）之间，并由成熟概率 p_{mat} 控制。随着延迟成熟，方差增加，但偏差减小。

10.2.4 与因果推断中的双稳健估计的联系

“双重稳健”这一术语源自因果推断，其中估计量结合 :

1. **结果回归模型**：$\mathbb{E}[Y|X, A]$（类似于我们的 DCM 插补 m_θ（$s,\ a$）)
2. **倾向得分模型**：$\mathbb{P}$（$A|X$）（不适用于我们的确定性定价场景）

我们的表述简化了经典双重稳健结构，因为：

- 动作 a_t 由智能体选择（无需倾向得分）
- “处理效应”是延迟奖励实现 $r_{\text{延迟}\,t}$
- 校正项 $\mathbb{1}_{\{\text{成熟}\}}$（$r_{\text{延迟}\,t}$ $- m_\theta$）会在真实结果显现时消除插补偏差

稳健性： 在因果推断中，只要结果模型 或倾向模型正确，双重稳健估计量就是无偏的。在我们的设定中，满足以下条件时，目标量是无偏的：

1. $m_\theta(s,\ a) = \mathbb{E}[r^{\text{delay}}|s, a]$（正确的 DCM），或
2. $p_{\text{mat}} = 1$（所有延迟均已成熟），或
3. 两者部分成立（偏差会随着 $(1 - p_{\text{mat}})\varepsilon$ 平缓增大）

这种对模型设定错误和不完整数据的双重防护，正是称其为“双重稳健”的原因。

10.2.5 经验校准

在我们的实验中：

- $p\text{mat}(t)$ 会随着历史延迟逐渐成熟，从 ≈ 0.05 （训练早期）增长到 ≈ 0.85 （训练后期）
- 延迟分量的 DCM 预测误差： $\varepsilon \approx 0.12$ （相对于延迟奖励幅度的平均绝对误差为 12%；对应于总奖励的绝对平均绝对误差为 \$2.30，或占平均预订收入的 $< 2\%$）
- 有效偏差： $(1 - 0.85) \times 0.12 = 0.018$ （占奖励尺度的 1.8%）
- 方差膨胀： $1 + 0.85 \times$ （σ^2延迟$/\sigma^2$即时） ≈ 1.3 （增加 30%）

经验上，偏差-方差权衡更 favor 双重稳健目标，而不是等待成熟（速度较慢）或仅插补（存在偏差）。

备注 4 （模型插补采样与双重稳健目标的关系）。 模型插补采样（算法*1*）和双重稳健目标（上文已分析）都利用 *DCM* 处理延迟反馈，但两者的机制不同：

1. 模型插补采样：在每一步，从 *DCM* 中采样 $j' \sim P(j|s_t, a_t;\ \theta)$，并使用合成奖励 $r'_t = r_t^{imm} + r_{j'}^{delay}$。当 *DCM* 正确时，这种方法会生成无偏合成样本，但其方差与完整奖励相同（没有方差缩减）。

2. 双重稳健目标：使用 $\tilde{r}_t = r_t^{imm} + m_\theta + \mathbb{1}_{\{matured\}}(r_t^{delay} - m_\theta)$。当存在待处理的延迟时，这种方法可以缩减方差（方差为 $= \sigma_{imm}^2$，而非 $\sigma_{imm}^2 + \sigma_{delay}^2$），同时在 *DCM* 正确时保持无偏。

该使用哪种方法？模型插补采样更易实现，并能生成完整的合成状态转移 (r'_t, s'_{t+1})，从而支持基于模型的规划。当方差缩减至关重要，且只需插补奖励（而非下一状态）时，双重稳健目标更为合适。我们的实现采用模型插补采样，因为它简单且兼容标准 *DQN* 网络架构。

10.3 命题：多项Logit对数似然的严格凹性

命题 7（多项 Logit 对数似然的严格凹性）。 考虑一个具有 $K+1$ 个离散结果 $\{0, 1, \ldots, K\}$ 的多项*logit*（多项 *Logit*）模型，其中结果 0 是“基准”或“外部”选项，其效用归一化为零。假设

（*A1*）协变量与概率不存在退化：协变量 $\{\mathbf{x}_i\}$ （或其分布）具有足够的变异性，使得对数似然的海森矩阵（或其期望形式）不会退化为低秩矩阵。[5]

那么，对数似然函数

$$\ell(\theta) \;=\; \sum_{n=1}^{N} \ln \Pr\{y_n \mid \mathbf{x}_n, \theta\}$$

是严格凹的 ˆ于 θ 中。因此，存在唯一的全局最大值点 θ，其满足 $\max_\theta \ell(\theta)$.

[5]例如，在多项logit模型中，要求不同备选项的回归变量之间不存在完全共线性，并且在$\mathbf{x}$的支撑集上，每种结果出现的概率都非零。

证明。 **1. 单个观测的对数似然与梯度。**

对于单个观测（$\mathbf{x}$ y），其对数似然贡献为:

$$\ell_{(\mathbf{x},y)}(\beta_1,\dots,\beta_K) \;=\; \ln \Pr\{y \mid \mathbf{x}\}.$$

由于 $\beta_0 = \mathbf{0}$, 我们有:

$$\Pr\{y=0 \mid \mathbf{x}\} \;=\; \frac{1}{1+\sum_{k=1}^{K}\exp(\beta_k^\top \mathbf{x})}, \quad \Pr\{y=m \mid \mathbf{x}\} \;=\; \frac{\exp(\beta_m^\top \mathbf{x})}{1+\sum_{k=1}^{K}\exp(\beta_k^\top \mathbf{x})} \;\text{ for } m=1,\dots,K.$$

因此,

$$\ell_{(\mathbf{x},y)}(\beta_1,\dots,\beta_K) \;=\; \begin{cases} -\ln\Big(1+\sum_{k=1}^{K} e^{\beta_k^\top \mathbf{x}}\Big), & \text{if } y=0, \\ \beta_y^\top \mathbf{x} \;-\; \ln\Big(1+\sum_{k=1}^{K} e^{\beta_k^\top \mathbf{x}}\Big), & \text{if } y\neq 0. \end{cases}$$

为简洁起见，定义 $p_k(\mathbf{x}) = \Pr\{y=k \mid \mathbf{x}\}$。经常规求导可得

$$\nabla_{\beta_k} \ell_{(\mathbf{x},y)} \;=\; \mathbf{x}\,\big[\mathbb{I}\{y=k\} \;-\; p_k(\mathbf{x})\big].$$

2. 单个观测的二阶导数（海森矩阵）。

现在我们计算 $\partial^2 \ell_{(\mathbf{x},y)}/(\partial\beta_k\,\partial\beta_\ell^\top)$。重点关注

$$\frac{\partial}{\partial \beta_\ell^\top}\Big(\mathbf{x}\,\big[\mathbb{I}\{y=k\}-p_k(\mathbf{x})\big]\Big) = \mathbf{x}\,\frac{\partial}{\partial \beta_\ell^\top}\big[\mathbb{I}\{y=k\}-p_k(\mathbf{x})\big].$$

由于 $\mathbb{I}\{y=k\}$ 相对于 β_ℓ 是常数，导数仅取决于 p_k（$\mathbf{x}$）。回顾一下,

$$p_k(\mathbf{x}) \;=\; \frac{\exp(\beta_k^\top \mathbf{x})}{1+\sum_{j=1}^{K}\exp(\beta_j^\top \mathbf{x})}.$$

由此可得,

$$\frac{\partial p_k(\mathbf{x})}{\partial \beta_\ell^\top} \;=\; \begin{cases} p_k(\mathbf{x})\,\big[\mathbf{x}\big], & \text{if } \ell=k, \\ -\,p_k(\mathbf{x})\,p_\ell(\mathbf{x})\,\big[\mathbf{x}\big], & \text{if } \ell\neq k. \end{cases}$$

等价地，可以用一个公式表示为

$$\frac{\partial p_k(\mathbf{x})}{\partial \beta_\ell^\top} \;=\; p_k(\mathbf{x})\,\Big[\delta_{k\ell}-p_\ell(\mathbf{x})\Big]\,\mathbf{x}^\top.$$

Thus

$$\frac{\partial}{\partial \beta_\ell^\top}\big[-\,p_k(\mathbf{x})\big] \;=\; -\,\Big[p_k(\mathbf{x})\big(\delta_{k\ell}-p_\ell(\mathbf{x})\big)\Big]\,\mathbf{x}.$$

综上可得

$$\frac{\partial^2 \ell_{(\mathbf{x},y)}}{\partial\beta_k\,\partial\beta_\ell^\top} \;=\; \mathbf{x}\,\Big[0-\frac{\partial p_k(\mathbf{x})}{\partial \beta_\ell^\top}\Big] \;=\; -\,\mathbf{x}\mathbf{x}^\top\,\Big[p_k(\mathbf{x})\,\big(\delta_{k\ell}-p_\ell(\mathbf{x})\big)\Big].$$

通过观察取非对角（$k\neq\ell$）和对角（$k=\ell$）部分，即可得到广为人知的结构 e:

$$\text{(off-diagonal)}\quad k\neq\ell: \qquad \frac{\partial^2 \ell_{(\mathbf{x},y)}}{\partial\beta_k\,\partial\beta_\ell^\top} = +\,\mathbf{x}\mathbf{x}^\top\,\big[p_k(\mathbf{x})\,p_\ell(\mathbf{x})\big],$$

$$\text{(diagonal)}\quad k=\ell: \qquad \frac{\partial^2 \ell_{(\mathbf{x},y)}}{\partial\beta_k\,\partial\beta_k^\top} = -\,\mathbf{x}\mathbf{x}^\top\,\big[p_k(\mathbf{x})\big(1-p_k(\mathbf{x})\big)\big].$$

3. 全样本的海森矩阵：凹性。

汇总 k，$\ell \in \{1,\dots,K\}$ 的各个分块后，我们得到*单个*观测的海森矩阵：

$$\nabla^2 \ell_{(\mathbf{x},y)}(\beta_1,\dots,\beta_K) \;=\; -\Big[\operatorname{diag}\big(p(\mathbf{x})\big) - p(\mathbf{x})\,p(\mathbf{x})^\top\Big] \;\otimes\; \big(\mathbf{x}\,\mathbf{x}^\top\big),$$

其中，$p(\mathbf{x}) = \big(p_1(\mathbf{x})\dots,p_K(\mathbf{x})\big)^\top$。对所有 $n = 1,\dots,N$ 求和，

$$\nabla^2 \ell(\beta_1,\dots,\beta_K) \;=\; \sum_{n=1}^{N} \nabla^2 \ell_{(\mathbf{x}_n,y_n)}(\beta_1,\dots,\beta_K) \;=\; -\sum_{n=1}^{N}\Big[\operatorname{diag}\big(p(\mathbf{x}_n)\big) - p(\mathbf{x}_n)\,p(\mathbf{x}_n)^\top\Big] \otimes \big(\mathbf{x}_n\,\mathbf{x}_n^\top\big).$$

由于每个矩阵 $\operatorname{diag}\big(p(\mathbf{x}_n)\big) - p(\mathbf{x}_n)\,p(\mathbf{x}_n)^\top$ 均为半正定矩阵（当 $p(\mathbf{x}_n)$ 位于概率单纯形内时），且 $\mathbf{x}_n\mathbf{x}_n^\top$ 也为半正定矩阵，因此前面的整体符号为负。这表明 $\nabla^2\ell$ 为*负半定*，故 ℓ 关于 $(\beta_1,\dots,\beta_K)$ 是凹的。

4. 识别条件下的严格凹性。

为消除加性平移不确定性，我们固定 $\beta_0 = \mathbf{0}$。在以下假设下：

(i) 设计矩阵满列秩（即不存在完全共线性），且

(ii) 每个 结果 $k \in \{1,\dots,K\}$ 在某个 >0 集合中的 $p_k(\mathbf{x}_n)$ 处取得（该集合具有正测度），

可以证明，上述负半定海森矩阵限制在已识别子空间上时，*实际上是负定的*（即，不会改变基线参数，也不会违反线性独立性条件的方向）。换言之，对于已识别参数子空间中的任意非零 v，

$$v^\top\Big[\nabla^2\ell(\beta_1,\dots,\beta_K)\Big]\,v < 0.$$

因此，在除去该识别选择后的参数空间上，对数似然是*严格凹的*（负定）。因此，一旦设定 $\beta_0 = \mathbf{0}$，$\ell(\beta_1,\dots,\beta_K)$ 就存在唯一的全局最大值点。

□

10.4 命题：循环极大似然估计的参数收敛

命题 8 (循环极大似然估计的参数收敛)。 设 $\theta^\star \in \Theta \subset \mathbb{R}^m$ 为决定选择模型的真实参数向量。对于每个区块 $t = 1,\ 2,\ \dots,$ 定义

$$\ell_t(\theta) \;=\; \sum_{i=1}^{N_t} \ln \Pr\big(y_{t,i} \mid \mathbf{x}_{t,i},\theta\big), \quad \textit{and} \quad \hat\theta_t \;=\; \arg\max_{\theta\in\Theta}\, \ell_t(\theta).$$

假设以下条件成立：

(A2) 大样本： 每个数据块的样本量几乎处处为 $N_t \to \infty$，当 $t \to \infty$ 时。

(A3) 遍历性或独立同分布数据块： 集合 $\{(\mathbf{x}_{t,\ i},y_{t,\ i})\}_{i=1}^{N_t}$ 要么在 t 上独立同分布，要么满足足以应用强大数定律的混合条件。

于是有

$$\lim_{t\to\infty} \|\hat\theta_t - \theta^\star\| \;=\; 0 \quad \textit{almost surely.}$$

证明。 **步骤 1。定义总体对数似然。** 令 $(\mathbf{x}\ y)$ 为从真实数据生成分布中抽取的一般观测。定义总体 或期望 对数似然:

$$L(\theta) \;=\; \mathbb{E}\big[\ln \Pr\big(y \mid \mathbf{x}, \theta\big)\big].$$

根据命题 7， $L(\theta)$ 在紧集 Θ上良定义且连续。此外， $L(\theta)$ 在 $\theta^\star$ 的某个邻域内严格凹，且在 $\theta^\star$处具有唯一的全局最大值点。

步骤 2。分块式样本对数似然与收敛。对于分块 t，样本对数似然 i s

$$\ell_t(\theta) \;=\; \sum_{i=1}^{N_t} \ln \Pr\big(y_{t,i} \mid \mathbf{x}_{t,i}, \theta\big).$$

考虑平均对数似然:

$$\frac{1}{N_t}\,\ell_t(\theta) \;=\; \frac{1}{N_t} \sum_{i=1}^{N_t} \ln \Pr\big(y_{t,i} \mid \mathbf{x}_{t,i}, \theta\big).$$

在关于遍历性或独立同分布抽样的假设以及假设（$N_t \to \infty$下，强大数定律论证表明，对于每个固定的 $\theta \in \Theta$,

$$\frac{1}{N_t}\,\ell_t(\theta) \;\xrightarrow{\text{a.s.}}\; L(\theta) \quad \text{as } t \to \infty. \tag{28}$$

这给出了逐点收敛，即样本准则函数收敛到总体准则函数。

步骤 3. 一致收敛。接下来我们证明，式（28）中的收敛实际上是关于 $\theta \in \Theta$的一致 Θ 。由于 Θ 是紧集，且$\ln \Pr\{y \mid \mathbf{x}\ \theta\}$ 假设在 θ上连续（且有上界），M 估计中的标准结果保证:

$$\sup_{\theta \in \Theta} \Big| \tfrac{1}{N_t}\,\ell_t(\theta) - L(\theta) \Big| \;\xrightarrow{\text{a.s.}}\; 0, \quad \text{as } t \to \infty.$$

一致收敛对于控制样本最大值点的行为至关重要 $\hat{\theta}_t$.

步骤 4. 样本最大值点的推论。根据定义,

$$\hat{\theta}_t \;=\; \arg\max_{\theta \in \Theta}\, \ell_t(\theta) \;=\; \arg\max_{\theta \in \Theta} \big\{ \tfrac{1}{N_t}\,\ell_t(\theta) \big\}.$$

因此,

$$\tfrac{1}{N_t}\,\ell_t\big(\hat{\theta}_t\big) \;\geq\; \tfrac{1}{N_t}\,\ell_t(\theta) \quad \text{for all } \theta \in \Theta.$$

根据一致收敛结果,

$$\sup_{\theta \in \Theta} \Big| \tfrac{1}{N_t}\,\ell_t(\theta) - L(\theta) \Big| \;\xrightarrow{\text{a.s.}}\; 0, \quad \text{thus} \quad \tfrac{1}{N_t}\,\ell_t\big(\hat{\theta}_t\big) \;\xrightarrow{\text{a.s.}}\; L\big(\hat{\theta}_t\big).$$

ˆ ˆ但由于 θ_t 取决于 t，我们需要证明它收敛到 $\theta^\star$。反设 θ_t 确实没有 ˆ 趋近于 $\theta^\star$。由于 Θ 是紧集，因此存在一个子序列 θ_{t_k} 收敛到某个 $\theta_\infty \neq \theta^\star$。根据命题 7中的严格凹性和唯一性结论,

$$L(\theta^\star) \;>\; L(\theta_\infty).$$

同时，一致收敛意味着

$$\lim_{k\to\infty}\left\{\tfrac{1}{N_{t_k}}\,\ell_{t_k}(\hat{\theta}_{t_k}) - L(\hat{\theta}_{t_k})\right\} \;=\; 0,$$

对任意 θ，类似地可得。因此，当 k 足够大时，

$$\tfrac{1}{N_{t_k}}\,\ell_{t_k}(\hat{\theta}_{t_k}) \;<\; \tfrac{1}{N_{t_k}}\,\ell_{t_k}(\theta^\star) - \delta$$

对于某个正的 δ（因为 $L(\theta^\star) - L(\theta_\infty)$ 在极限下仍严格为正）。这与将 $\hat{}\ \theta_{t_k}$ 定义为 Θ上的全局最大值点相矛盾。因此，不存在这样的子序列收敛到$\theta_\infty \neq \theta^\star$，这意味着

$$\hat{\theta}_t \;\xrightarrow{\text{a.s.}}\; \theta^\star.$$

$\hat{}$也就是说，$\|\theta_t - \theta^\star\| \to 0$ 几乎处处成立。

$\hat{}$**步骤 5. 总结性说明**。我们已证明，每个极大值点序列 $\{\theta_t\}$ 几乎处处收敛到总体对数似然的唯一最大化点 $\theta^\star$。至此，循环极大似然估计量一致性的证明完成。

□

10.5 命题：DCM 的满秩 Fisher 信息

命题 9（DCM 的满秩 Fisher 信息）。 在假设 *(A1)-(A3)* 下，则在这些条件下，由下式定义的 *Fisher* 信息矩阵$\mathbf{I}(\theta^\star)$

$$\mathbf{I}(\theta^\star) \;=\; \mathbb{E}\Big[\nabla_\theta\big(-\ln \Pr\{y \mid \mathbf{x}, \theta^\star\}\big)\,\nabla_\theta\big(-\ln \Pr\{y \mid \mathbf{x}, \theta^\star\}\big)^\top\Big]$$

是满秩（即正定）。

证明。**步骤 1：将费舍尔信息与期望 Hessian 联系起来**。

回顾可知，在 $\theta^\star$ 处针对独立同分布样本 $\{(\mathbf{x}_i,\ y_i)\}$ 的 Fisher 信息矩阵有两个众所周知的等价定义：

$$\mathbf{I}(\theta^\star) \;=\; \mathbb{E}\Big[\nabla_\theta\big(-\ln \Pr\{y \mid \mathbf{x}, \theta^\star\}\big)\,\nabla_\theta\big(-\ln \Pr\{y \mid \mathbf{x}, \theta^\star\}\big)^\top\Big] \;=\; \mathbb{E}\Big[\nabla^2_\theta\big(-\ln \Pr\{y \mid \mathbf{x}, \theta\}\big)\,\big|_{\theta=\theta^\star}\Big].$$

第一种表达式表明，它是得分函数在参数空间中的协方差；第二种表达式将其表示为对数似然的期望 Hessian 的负值。

步骤 2：严格凹性蕴含负定 Hessian。

根据命题7（或其总体对应形式），对数似然（或总体对数似然）是严格凹的，在 $\theta^\star$附近。具体而言，令

$$L(\theta) \;=\; \mathbb{E}\big[\ln \Pr\{y \mid \mathbf{x}, \theta\}\big].$$

在 $\theta^\star$ 附近的严格凹性意味着海森矩阵 $\nabla^2_\theta\, L(\theta^\star)$ 是严格负定的。等价地，对于任意非零向量$\mathbf{u}$，

$$\mathbf{u}^\top \nabla^2_\theta\, L(\theta^\star)\,\mathbf{u} \;<\; 0.$$

由于

$$\nabla^2_\theta\, L(\theta^\star) \;=\; \mathbb{E}\Big[\nabla^2_\theta\, \ln \Pr\{y \mid \mathbf{x}, \theta^\star\}\Big] \;=\; -\,\mathbf{I}(\theta^\star),$$

我们看到，$\nabla^2_\theta\, L(\theta^\star) \;=\; -\mathbf{I}(\theta^\star)$。因此，$\nabla^2_\theta\, L(\theta^\star)$ 为负定恰好意味着$\mathbf{I}(\theta^\star)$ 为正定。

□

11 多项 Logit 收敛速率分析

多项 Logit 最大似然估计的收敛性质在计量经济学文献中已有充分确立。我们在此陈述关键结果，并参阅标准文献以了解详细证明。

定理 10 （多项 Logit 的收敛速率）。 设 $\theta^\star \in \Theta \subset \mathbb{R}^m$ 为支配该选择模型的真实参数向量。在假设 ***(A1)–(A4)***（可识别性、正则性、独立性以及有界ˆ协变量）下，最大似然估计量 θ_N 满足：

1. 一致性：$\hat{}\theta_N \xrightarrow{p} \theta^\star$ 当 $N \to \infty$
2. 收敛速率：$\|\hat{}\theta_N - \theta^\star\| = O_p(N^{-1/2})$
3. ***Asymptotic normality:*** $\sqrt{N}(\hat{\theta}_N - \theta^\star) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}(\theta^\star)^{-1})$

其中 $\mathbf{I}(\theta^\star)$是费舍尔信息矩阵。

证明。 这是离散选择计量经济学中的经典结果。多项 Logit 对数似然函数的严格凹性（已在命题 8中证明）确保最大似然估计量全局收敛。-速率和渐近正态性可由标准 M 估计理论推出：对得分函数在 $\theta^\star$处进行泰勒展开，并结合得分函数的中心极限定理以及海森矩阵收敛至费舍尔信息矩阵。完整证明见 Train [2009] 第 3 章、McFadden [1974], 以及Ben-Akiva 和 Lerman [1985] 第 5 章。 □

定理11 （预测选择概率的收敛速率）。 在与定理 *10*相同的条件下，对于每个固定的协变量向量 $\mathbf{x}$ 和备选项 k:

$$\left|\Pr\{y = k \mid \mathbf{x}, \hat{\theta}_N\} - \Pr\{y = k \mid \mathbf{x}, \theta^\star\}\right| = O_p(N^{-1/2})$$

and

$$\sqrt{N}\big(p_k(\mathbf{x}; \hat{\theta}_N) - p_k(\mathbf{x}; \theta^\star)\big) \xrightarrow{d} \mathcal{N}\big(0, \mathbf{J}_k(\theta^\star)\mathbf{I}(\theta^\star)^{-1}\mathbf{J}_k(\theta^\star)^\top\big)$$

其中 $\mathbf{J}_k(\theta^\star) = \nabla_\theta p_k(\mathbf{x}; \theta)|_{\theta=\theta^\star} \in \mathbb{R}^{1\times m}$ 是雅可比矩阵。

证明。 将德尔塔方法应用于光滑映射 $\theta \mapsto p_k(\mathbf{x}; \theta$). □

备注 5 （与假设 A6 的联系）。 这些经典结果为正文中的假设 *A6* 提供了依据：随着 N_t 截至时间 t 线性增长的成熟观测数量（引理 *12* ），*DCM* 近似误差以 $\varepsilon_t = O(t^{-1/2})$ 的速率衰减。

12 延迟反馈分析

本节分析我们收益管理场景中的延迟反馈结构。关于延迟奖励的马尔可夫决策过程，已有大量理论研究 [Katsikopoulos 和 Engelbrecht, 2003],为异步成本收集和延迟状态观测奠定了基础。

12.1 成熟率引理：将延迟分布与样本量增长联系起来

我们理论框架的一个关键分量是证明：截至时间 t，观测到的已成熟预订结果数量随 t 线性增长。这一成熟速率决定了 DCM 累积训练数据的速度，进而决定假设 A6 中的近似误差衰减率 $\varepsilon_t = O(t^{-1/2})$。

引理 12（延迟反馈的成熟率）。考虑如下延迟反馈模型：

(i) 每位在时间 s 到达的客户都会产生一个预订决策，该决策在时间 $s + D_s$成熟（即变得可观测），其中 D_s 是该客户在时间 s到达所产生的随机延迟

(ii) 延迟 $\{D_s\}_{s=1}^{\infty}$ 是独立同分布随机变量，且：

- *Finite support:* $D_s \in \{0, 1, \ldots, D_{\max}\}$ *for some* $D_{\max} < \infty$
- *Finite mean:* $\mathbb{E}[D_s] = \mu_D < \infty$

(iii) 客户到达率保持平稳：平均每个时间段有 $\lambda > 0$ 位客户到达

定义 N_t 为截至时间 t的成熟观测数量：

$$N_t := \sum_{s=1}^{t} \mathbb{1}\{s + D_s \le t\} = \#\{\textit{customers arriving at } s \le t \textit{ whose decisions matured by time } t\} \tag{29}$$

于是：

(a) 几乎必然线性增长：$N_t/t \xrightarrow{a.s.} \lambda$ 当 $t \to \infty$

(b) 集中性：对于任意 $\epsilon > 0$ 及所有 $t \ge 2CD_{\max}/\epsilon$：

$$\mathbb{P}\left(\left|\frac{N_t}{t} - \lambda\right| > \epsilon\right) \le 2\exp\left(-\frac{\epsilon^2 t^2}{2D_{\max}}\right) \tag{30}$$

对于某个普适常数 $C > 0$。请注意，指数中对 t 的依赖是二次的：随着 t增长，集中性会呈指数级增强，因为只有最近的 $D_{\max}$ 个到达会带来随机性（所有更早的到达都已确定性地成熟）

(c) 倒数收敛：$\frac{1}{\sqrt{N_t}} = \Theta(t^{-1/2})$ 几乎处处

证明。我们分别证明各部分。

（a）几乎必然线性增长。

固定时间 t，并将已成熟计数分解为：

$$N_t = \sum_{s=1}^{t} \mathbb{1}\{s + D_s \le t\} = \sum_{s=1}^{t} \mathbb{1}\{D_s \le t - s\} \tag{31}$$

关键观测是，由于 $D_{\max} < \infty$，任何在时间 $s \le t - D_{\max}$ 到达的客户，其决策都会在时间 t前以概率 1 完成（因为最大延迟为 $D_{\max}$）。因此：

$$N_t \ge \sum_{s=1}^{t-D_{\max}} \mathbb{1}\{D_s \le D_{\max}\} = t - D_{\max} \tag{32}$$

显然 $N_t \le t$。

对于期望值：

$$\mathbb{E}[N_t] = \sum_{s=1}^{t} \mathbb{P}(D_s \le t-s) \tag{33}$$

$$= \underbrace{\sum_{s=1}^{t-D_{\max}} 1}_{=t-D_{\max}} + \underbrace{\sum_{s=t-D_{\max}+1}^{t} \mathbb{P}(D \le t-s)}_{=\sum_{k=0}^{D_{\max}-1}\mathbb{P}(D\le k)} \tag{34}$$

$$= t - D_{\max} + \sum_{k=0}^{D_{\max}-1} \mathbb{P}(D \le k) \tag{35}$$

其中第二个求和有界于 $D_{\max}$。因此 $\mathbb{E}[N_t] = t + O(D_{\max})$（对于 $\lambda = 1$；一般情形给出 $\mathbb{E}[N_t] = \lambda t + O(D_{\max})$）。

对于几乎必然收敛，将 $N_t = N_t^{\text{old}} + N_{\text{近期}t}$ 分解如下：

- $N_{\text{较早}t} : = \sum^{t-D_{\text{最大值}}}{}_{s=1} \quad \mathbb{1}\{D_s \le t-s\} = t - D_{\text{最大值}}$（确定性，所有较早的到达项均已成熟）
- $N_{\text{近期}t} : = \sum_{s=t-D^{\text{最大值}}+1}^{t} \mathbb{1}\{D_s \le t-s\}$（至多 $D_{\text{最大值}}$ 项）

由于 $|N_t^{\text{recent}}| \le D_{\max}$，我们有：

$$\frac{N_t}{t} = \frac{t - D_{\max} + N_t^{\text{recent}}}{t} = 1 - \frac{D_{\max}}{t} + \frac{N_t^{\text{recent}}}{t} \xrightarrow{a.s.} \lambda \tag{36}$$

因为 $t \to \infty$（每个周期有 $\lambda = 1$ 次到达；一般情形可通过重新缩放时间得出）。

（b）部分：集中性。

我们将偏差分解为两部分。对于“较早”的到达（$s \le t - D_{\max}$），其成熟过程是确定性的，因此不存在随机性。对于“近期”的到达（$s > t - D_{\max}$），定义：

$$Z_s := \mathbb{1}\{D_s \le t-s\} - \mathbb{P}(D_s \le t-s) \tag{37}$$

作为均值为零的随机变量，且满足 $|Z_s| \le 1$。由于此类项至多有 $D_{\max}$ 个：

$$N_t - \mathbb{E}[N_t] = \sum_{s=t-D_{\max}+1}^{t} Z_s \tag{38}$$

将霍夫丁不等式应用于这些 $D_{\max}$ 个相互独立且有界的变量，可得：

$$\mathbb{P}(|N_t - \mathbb{E}[N_t]| > \epsilon t) \le 2\exp\left(-\frac{2(\epsilon t)^2}{D_{\max}}\right) \tag{39}$$

由于 $|\mathbb{E}[N_t] - \lambda t| \le C_1 D_{\max}$ 对某个常数 C_1 成立，我们应用三角不等式。若 $\left|\frac{N_t}{t} - \lambda\right| > \epsilon$，则 $\left|\frac{N_t - \mathbb{E}[N_t]}{t}\right| > \frac{\epsilon}{2}$ 或 $\left|\frac{\mathbb{E}[N_t] - \lambda t}{t}\right| > \frac{\epsilon}{2}$ 必有一个成立。

第二个事件仅在 $\frac{C_1 D_{\max}}{t} > \frac{\epsilon}{2}$ 时发生，即 $t < \frac{2C_1 D_{\max}}{\epsilon}$ 时。

对于第一个事件，我们使用带有 $\epsilon' = \epsilon/2$ 的霍夫丁界：

$$\mathbb{P}\left(|N_t - \mathbb{E}[N_t]| > \frac{\epsilon t}{2}\right) \le 2\exp\left(-\frac{2(\epsilon t/2)^2}{D_{\max}}\right) = 2\exp\left(-\frac{\epsilon^2 t^2}{2D_{\max}}\right) \tag{40}$$

对于 $t \geq \frac{2C_1 D_{\max}}{\epsilon}$，确定性偏差项满足 $\frac{C_1 D_{\max}}{t} \leq \frac{\epsilon}{2}$，因此：

$$\mathbb{P}\left(\left|\frac{N_t}{t} - \lambda\right| > \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2 t^2}{2D_{\max}}\right) \tag{41}$$

这就是所述的界，其中 $C = C_1$。

部分（c）：倒数收敛。

由（a）部分可得，$N_t = \lambda t + O(D_{\max})$ 几乎处处。因此：

$$\frac{1}{\sqrt{N_t}} = \frac{1}{\sqrt{\lambda t + O(D_{\max})}} \tag{42}$$

$$= \frac{1}{\sqrt{\lambda t}} \cdot \frac{1}{\sqrt{1 + O(D_{\max}/t)}} \tag{43}$$

$$= \frac{1}{\sqrt{\lambda}} \cdot t^{-1/2} \cdot (1 + O(D_{\max}/t)) \tag{44}$$

$$= \Theta(t^{-1/2}) \tag{45}$$

这就得到了所需的收敛速率。 □

备注1（对 DCM 收敛率的意义）。引理12直接表明，时间 t处的 DCM 参数估计误差满足：

$$\|\hat{\theta}_{N_t} - \theta^\star\| = O_p(N_t^{-1/2}) = O_p(t^{-1/2}) \tag{46}$$

这建立了 Q学习收敛分析中假设 A6 所要求的 $\varepsilon_t = O(t^{-1/2})$ 衰减率。

备注2（回合式结构）。在酒店收益管理应用中（主论文第 6 节），每个回合持续 $H = 14$天（预订窗口）。顾客在整个回合期间陆续到达，延迟$D_s \in \{0, 1, \ldots, 14\}$ 表示距离预订成熟还需等待的天数。回合重置确保延迟不会超过回合时域，从而满足有限支撑条件 $D_{\max} = 14 < \infty$。

备注3（与固定 DCM 情形的比较）。在我们的实证实验中（第15节），我们使用一个固定的预先校准的 DCM，其中包含 $N_0 = 61$个，619 条历史观测。在这种情况下：

- DCM 误差 $\varepsilon_0 = O(N_0^{-1/2}) \approx 0.004$ 保持不变
- Q学习分析适用，且 $\varepsilon_t = \varepsilon_0$ （无衰减）
- 收敛速率变为 $\|Q_t - Q^\star\|_\infty = O(t^{-1/2}\sqrt{\log t} + \varepsilon_0)$
- 对于较大的 N_0， ε_0 偏差项可以忽略不计

成熟率引理解释了为何 在线 *DCM* 再校准 能够实现 $\varepsilon_t = O(t^{-1/2})$衰减，并为未来全面实现双时间尺度理论框架的方案提供了路线图。

13 Q 学习收敛

本节证明了在时变近似误差下 Q 学习的收敛性。我们的分析与鲁棒 MDP 文献 [Iyengar, 2005], 相联系，其中 DCM 可视为提供了一个具有有界不确定性的名义转移模型。关键见解是，当近似误差有界且以适当方式衰减时，Q 学习仍能保持稳定。

13.1 定理：近似误差逐渐缩小且收敛速率有保证的非平稳 Q 学习

定理 13（近似误差缩减且具有收敛速率的非平稳 Q 学习）。设 $\mathcal{M} = (S,\ A,\ P,\ r,\ \gamma)$是一个马尔可夫决策过程（*MDP*），满足：

- 有限状态空间 S,
- 有限动作空间 A,
- 真实转移核 $P(\cdot \mid s, a)$,
- 有界奖励 $r(s,a) \in [0, R_{\text{最大值}}]$ 对于某个 $R_{\text{最大值}} > 0$,
- 折扣因子 $\gamma \in (0,1)$。
- （遍历性）马尔可夫决策过程是遍历的，这意味着对于任意平稳策略，其诱导的马尔可夫链都是不可约且非周期的。
- （探索）智能体遵循 ϵ-贪婪策略，且 $\epsilon > 0$ 固定。

令 $Q^\star$ 为与 $\mathcal{M}$相关联的唯一最优 Q 函数，即贝尔曼算子的唯一不动点

$$(TQ)(s,a) \;=\; r(s,a) \;+\; \gamma \sum_{s' \in S} P(s' \mid s,a)\, \max_{a' \in A} Q(s',a').$$

假设我们在非平稳环境中运行 Q学习

$$\widetilde{\mathcal{M}}_t \;=\; \big(S,\, A,\, \widetilde{P}_t,\, \widetilde{r}_t,\, \gamma\big),$$

其中，对于所有 $(s,a) \in S \times A$,

$$\|\widetilde{P}_t(\cdot \mid s,a) - P(\cdot \mid s,a)\|_1 \;\le\; \varepsilon_t, \quad \big|\widetilde{r}_t(s,a) - r(s,a)\big| \;\le\; \varepsilon_t,$$

其中 $\varepsilon_t \to 0$ 作为 $t \to \infty$（例如，$\varepsilon_t = O(t^{-\beta})$ 对于某个 $\beta > 0$）。

令 $n(s,a)$ 表示截至当前时间，状态-动作对 (s,a) 被访问的次数。采用逐对步长的异步 Q 学习更新为：

$$Q_{t+1}(s_t,a_t) \;=\; Q_t(s_t,a_t) \;+\; \alpha_t(s_t,a_t)\Big[\widetilde{r}_t(s_t,a_t) \;+\; \gamma \max_{a'} Q_t(s_{t+1},a') \;-\; Q_t(s_t,a_t)\Big],$$

其中，$\alpha_t(s_t,a_t) = 1/n(s_t,a_t)$，且对所有 $(s,a) \neq (s_t,a_t)$，$Q_{t+1}(s,a) = Q_t(s,a)$。这里，$(s_{t+1}, \widetilde{r}_t)$ 分别从 $P_t(\cdot \mid s_t,a_t)$、$\widetilde{r}_t(s_t,a_t)$ 中采样得到。

对每个固定的 (s,a)，逐对步长 $\{\alpha_k(s,a)\}_{k=1}^{\infty}$ 满足 *Robbins–Monro* 条件：

$$\sum_{k=1}^{\infty} \alpha_k(s,a) = \sum_{k=1}^{\infty} \frac{1}{k} = \infty, \qquad \sum_{k=1}^{\infty} \alpha_k^2(s,a) = \sum_{k=1}^{\infty} \frac{1}{k^2} < \infty,$$

并且我们假设，按照 ϵ-贪婪探索策略，每个 (s,a) 都会被无限次访问。则以下两条结论成立：

1. *Almost-sure convergence:* $\lim_{t \to \infty} \|Q_t - Q^\star\|_\infty \;=\; 0$ *with probability* 1.

2. 高概率下的有限时间速率：
如果 $\varepsilon_t = O(t^{-\beta})$ 对某个 $\beta > 0$ 成立，并且在 ϵ-贪心探索下，每个状态-动作对截至时间 t 被访问 $\Theta(t)$ 次（因此，对于频繁访问的状态-动作对，逐对步长 $\alpha_t(s,a) = 1/n(s,a) = O(1/t)$），那么对于任意 $\delta \in (0,1)$ 和任意 $t \geq 1$，以下结论以至少 $1-\delta$ 的概率成立：

$$\|Q_t - Q^\star\|_\infty = O\left(t^{-\min(\beta,1/2)} \cdot \ln\left(\frac{|S||A|t}{\delta}\right)^{\mathbb{1}\{\beta \geq 1/2\}/2}\right) \tag{47}$$

特别地，如果 $\beta = 1/2$，则有 $\|Q_t - Q^\star\|_\infty = O\left(t^{-1/2}\sqrt{\log(|S||A|t/\delta)}\right)$。

定理 13的证明

证明。 我们分六个步骤给出详细证明：

1. 分析平稳贝尔曼算子及其压缩性质 2. 刻画时变贝尔曼算子及其近似误差
3. 构建异步 Q 学习更新 4. 分析鞅差噪声 5. 运用随机逼近理论证明几乎必然收敛
6. 推导有限时间高概率界

步骤 1：平稳贝尔曼算子的性质。对于平稳马尔可夫决策过程 $(S,\ A,\ P,\ r,\ \gamma)$，贝尔曼最优算子 T：$\mathbb{R}^{|S||A|} \to \mathbb{R}^{|S||A|}$ 定义如下：

$$(TQ)(s,a) \ = \ r(s,a) \ + \ \gamma \sum_{s'} P(s' \mid s,a) \ \max_{a'} Q(s',a')$$

已知 T 具有以下性质：

(a) T 是上确界范数下的 γ-压缩映射，即对于任意 $Q,\ Q' \in \mathbb{R}^{|S||A|}$：

$$\|TQ - TQ'\|_\infty \ \leq \ \gamma \, \|Q - Q'\|_\infty \tag{48}$$

(b) 根据巴拿赫不动点定理，T 有唯一的不动点 $Q^\star \in \mathbb{R}^{|S||A|}$，满足 $Q^\star = TQ^\star$。此不动点是该马尔可夫决策过程的最优 Q 函数。

(c) 不动点 $Q^\star$ 对所有 $(s,\ a) \in S \times A$均满足：

$$Q^\star(s,a) = r(s,a) + \gamma \sum_{s'} P(s' \mid s,a) \max_{a'} Q^\star(s',a') \tag{49}$$

步骤 2：时变贝尔曼算子及其近似误差。在非平稳环境 $(S,\ A,\ \widetilde{P}_t,\ \widetilde{r}_t,\ \gamma)$中，我们将时变贝尔曼算子定义为 T_t：$\mathbb{R}^{|S||A|} \to \mathbb{R}^{|S||A|}$ 如下：

$$(T_t Q)(s,a) \ = \ \widetilde{r}_t(s,a) \ + \ \gamma \sum_{s'} \widetilde{P}_t(s' \mid s,a) \ \max_{a'} Q(s',a')$$

引理 14 (时变算子的近似误差)。 对于任意有界的 $Q \in \mathbb{R}^{|S||A|}$，定义~奖励误差 $\varepsilon_{r,\ t}$：$= |\widetilde{r}_t - r|_\infty$ 和状态转移误差 $\varepsilon_{P,\ t}$：$= \|P_t - P\|_1$。那么：

$$\|T_t Q - TQ\|_\infty \ \leq \ \varepsilon_{r,t} + \gamma V_{\max} \varepsilon_{P,t}$$

其中 $V_{\max} = \frac{R_{\max}}{1-\gamma}$是最大价值函数的一个上界。

证明引理 *14*。 对于任意 $(s,\ a) \in S \times A$，我们有：

$$|(T_t Q)(s,a) - (TQ)(s,a)| = \left| \widetilde{r}_t(s,a) - r(s,a) + \gamma \sum_{s'} \Big(\widetilde{P}_t(s'|s,a) - P(s'|s,a) \Big) \max_{a'} Q(s',a') \right| \quad (50)$$

$$\leq |\widetilde{r}_t(s,a) - r(s,a)| + \gamma \left| \sum_{s'} \Big(\widetilde{P}_t(s'|s,a) - P(s'|s,a) \Big) \max_{a'} Q(s',a') \right| \quad (51)$$

$$\leq \varepsilon_{r,t} + \gamma \|Q\|_\infty \|\widetilde{P}_t(\cdot|s,a) - P(\cdot|s,a)\|_1 \quad (52)$$

$$\leq \varepsilon_{r,t} + \gamma V_{\max} \varepsilon_{P,t} \quad (53)$$

第三个不等式由 H¨older 不等式以及事实 $\|\max_{a'} Q(\cdot,\ a')\|_\infty \leq \|Q\|_\infty$ 得出。对于奖励以 $R_{\max}$ 为界的马尔可夫决策过程中的任意 Q 函数，我们有 $\|Q\|_\infty \leq V_{\max} = \frac{R_{\max}}{1-\gamma}$。 ——
由于这对所有 $(s,\ a) \in S \times A$ 成立，我们有 $\|T_t Q - TQ\|_\infty \leq \varepsilon_{r,\ t} + \gamma V_{\max} \varepsilon_{P,\ t}$。

注记（尺度解释）： 当状态转移误差的贡献为 $O(\varepsilon_r/(1-\gamma))$，而状态转移误差的贡献为 $O(\varepsilon_P \cdot V_{\max}/(1-\gamma)) = O(\varepsilon_P/(1-\gamma)^2)$。状态转移误差对 $(1-\gamma)^{-1}$ 的这种二次依赖与基于模型的强化学习文献中的标准模拟引理界相一致。 □

为便于后续分析中的记号表示，我们定义组合近似误差 $\varepsilon_t := \varepsilon_{r,\ t} + \gamma V_{\max} \varepsilon_{P,\ t}$ 以及常数 $C_1 = 1$，从而得到 $\|T_t Q - TQ\|_\infty \leq C_1 \varepsilon_t$。在跟踪各自的缩放关系时，我们明确使用$\varepsilon_{r,\ t}$ 和$\varepsilon_{P,\ t}$。

由于 $\varepsilon_t \to 0$ 当 $t \to \infty$时，我们在算子范数下有 $T_t \to T$，这意味着时变算子收敛于平稳算子。

步骤 3：异步 Q 学习更新。 Q学习算法每次根据以下规则更新一个状态-动作对：

$$Q_{t+1}(s_t,a_t) = Q_t(s_t,a_t) + \alpha_t \Big[\widetilde{r}_t(s_t,a_t) + \gamma \max_{a'} Q_t\big(s_{t+1},a'\big) \ - \ Q_t(s_t,a_t) \Big],$$

其中（$s_t,\ a_t$）是在时间 t访问的状态-动作对，而 s_{t+1} 是根据 $\sim P_t$（$\cdot|s_t,\ a_t$）采样得到的下一个状态。所有其他条目保持不变：对于（$s,\ a$）$\neq$（$s_t,\ a_t$），Q_{t+1}（$s,\ a$）$= Q_t$（$s,\ a$）。

我们可以将此更新改写为向量形式：

$$Q_{t+1} \ = \ Q_t \ + \ \alpha_t \, e_{(s_t,a_t)} \Big[\big(T_t Q_t\big)\big(s_t,a_t\big) \ - \ Q_t(s_t,a_t) \ + \ \delta_t \Big],$$

其中，$e_{(s_t,a_t)}$ 是 $\mathbb{R}^{|S||A|}$ 中的标准基向量，在（s_t, a_t）分量处为 1，其余位置为 0；δ_t是抽样噪声，定义如下：

$$\delta_t \ := \ \big[\widetilde{r}_t(s_t,a_t) + \gamma \max_{a'} Q_t(s_{t+1},a') - Q_t(s_t,a_t) \big] \ - \ \big[(T_t Q_t)(s_t,a_t) - Q_t(s_t,a_t) \big].$$

Step 4：鞅差噪声分析。现在，我们来详细分析噪声项 δ_t。

引理15（鞅差性质）。噪声项 δ_t 满足：

(a) $\mathbb{E}[\delta_t \mid \mathcal{F}_t] = 0$，其中 $\mathcal{F}_t = \sigma\big(\{(s_u, a_u, Q_u) : \ u \le t\}\big)$ 是表示截至时间 t的算法历史的滤过。

(b) $|\delta_t| \le 2V_{\max}$ 几乎处处。

引理*15*的证明。(a) 根据定义：

$$\delta_t = \big[\widetilde{r}_t(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t)\big] - \big[(T_t Q_t)(s_t, a_t) - Q_t(s_t, a_t)\big] \tag{54}$$

$$= \widetilde{r}_t(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') - (T_t Q_t)(s_t, a_t) \tag{55}$$

~给定历史 $\mathcal{F}_t$，Q值 Q_t 已确定。s_{t+1} 的分布为 $P_t(\cdot|s_t, \ a_t)$。因此：

$$\mathbb{E}[\delta_t \mid \mathcal{F}_t] = \widetilde{r}_t(s_t, a_t) + \gamma \mathbb{E}[\max_{a'} Q_t(s_{t+1}, a') \mid \mathcal{F}_t] - (T_t Q_t)(s_t, a_t) \tag{56}$$

$$= \widetilde{r}_t(s_t, a_t) + \gamma \sum_{s'} \widetilde{P}_t(s'|s_t, a_t) \max_{a'} Q_t(s', a') - (T_t Q_t)(s_t, a_t) \tag{57}$$

$$= (T_t Q_t)(s_t, \ a_t) \ - (T_t Q_t)(s_t, \ a_t) \tag{58}$$

$$= 0 \tag{59}$$

(b) 两项 $\widetilde{r}_t(s_t, \ a_t) + \gamma \max_{a'} Q_t(s_{t+1}, \ a')$ 和 $(T_t Q_t)(s_t, \ a_t)$ 均被 $V_{\max}$ 所界定，因为奖励有界于 $R_{\max}$ 的马尔可夫决策过程中，Q值有界于 $V_{\max} = \frac{R_{\max}}{1-\gamma}$。因此，$|\delta_t| \le 2V_{\max}$。 □

根据引理 15，$\{\delta_t\}$ 构成关于滤过$\{\mathcal{F}_t\}$的有界鞅差序列。这一性质对于后续应用集中不等式至关重要。

步骤 5：利用随机逼近理论证明几乎必然收敛。现在我们证明定理的第一部分：$Q_t \to Q^\star$ 当 $t \to \infty$时几乎处处成立。

定义 $\Delta_t := Q_t - Q^\star$ 为时间 t 的误差。我们将分析该误差如何随时间演变。

首先，我们将更新分解为一个压缩项和若干误差项：

Le引理 16（误差分解）。对于任意时间 t，误差项 $\Delta_t = Q_t - Q^\star$ 满足 *s:*

$$\Delta_{t+1} = \Delta_t + \alpha_t e_{(s_t, a_t)} \left[(TQ_t - TQ^\star)(s_t, a_t) + ((T_t - T)Q_t)(s_t, a_t) + \delta_t\right] \tag{60}$$

引理 *16* 的证明。由于 $Q^\star = TQ^\star$，我们有：

$$\Delta_{t+1} = Q_{t+1} - Q^\star \tag{61}$$

$$= Q_t + \alpha_t e_{(s_t, a_t)}[(T_t Q_t)(s_t, a_t) - Q_t(s_t, a_t) + \delta_t] - Q^\star \tag{62}$$

$$= \Delta_t + \alpha_t e_{(s_t, a_t)}[(T_t Q_t)(s_t, a_t) - Q_t(s_t, a_t) + \delta_t] \tag{63}$$

$$\tag{64}$$

现在我们加上并减去 $(TQ_t)(s_t, \ a_t)$：

$$\Delta_{t+1} = \Delta_t + \alpha_t e_{(s_t, a_t)}[(TQ_t)(s_t, a_t) - Q_t(s_t, a_t) + ((T_t - T)Q_t)(s_t, a_t) + \delta_t] \tag{65}$$

$$\tag{66}$$

此外，我们加上并减去 $(TQ^\star)(s_t,\ a_t) = Q^\star(s_t,\ a_t)$：

$$\Delta_{t+1} = \Delta_t + \alpha_t e_{(s_t,a_t)}[(TQ_t)(s_t,a_t) - (TQ^\star)(s_t,a_t) + ((T_t - T)Q_t)(s_t,a_t) + \delta_t] \tag{67}$$
$$= \Delta_t + \alpha_t e_{(s_t,a_t)}[(TQ_t - TQ^\star)(s_t,a_t) + ((T_t - T)Q_t)(s_t,a_t) + \delta_t] \tag{68}$$
$$\tag{69}$$

由此得到所需的分解。 □

为了证明 Q_t 几乎必然收敛到 $Q^\star$，我们将应用 [Tsitsiklis, 1994]中提出的异步随机逼近理论。具体而言，我们将使用该论文中定理 3 的一个简化版本。

Th定理 17。 考虑如下形式的异步随机逼近算法 ：

$$x_{t+1}(i) = x_t(i) + \alpha_t(i)[h_i(x_t) + w_t(i)], \quad i \in \{1, 2, ..., n\}$$

其中：

- 每个时间步仅更新一个分量 i
- $h = (h_1,\ h_2,\ ...,\ h_n)$ 是一个具有不动点 x^* 的 γ-压缩映射
- 步长 $\alpha_t(i)$ 满足 $\sum_{t=1}^{\infty} \alpha_t(i) = \infty$ 和 $\sum_{t=1}^{\infty} \alpha_t^2(i) < \infty$
- 每个分量 i 被无限次更新
- $w_t(i)$是具有有界方差的鞅差序列
- 对每一个 i 和 t，$\mathbb{E}[w_t(i) \mid \mathcal{F}_t] = 0$，其中 $\mathcal{F}_t$ 是由截至时间 t的算法历史生成的 σ-域

则 x_t 几乎处处收敛到 x^* 。

现在，我们将定理17 应用于我们的Q学习设定。

命题 18 (几乎必然收敛)。 在定理*13*的假设下，

$$\lim_{t\to\infty} \|Q_t - Q^\star\|_\infty = 0 \quad \textit{with probability 1}$$

命题的证明*18*。 我们将 Q学习设置映射到定理 17的框架中：

- 状态空间为 $\mathbb{R}^{|S||A|}$，其中 $x_t = Q_t$
- 不动点为 $x^* = Q^\star$
- 映射 h 对应于对 $(TQ - Q)$ 的逐分量应用
- 噪声项 $w_t(i)$ 对应于鞅差 δ_t 与近似误差 $(T_t - T)Q_t$之和

根据引理 15， δ_t 是相对于 $\mathcal{F}_t$的鞅差序列。虽然近似误差 $(T_t - T)Q_t$ 在给定历史的条件下是确定的，但随着 $t \to \infty$ 的增大，它会收敛于零，因为 $\varepsilon_t \to 0$。因此，组合噪声项最终表现得像一个鞅差序列。

要应用定理17，我们需要证明算子 T 在我们的设定下构成压缩映射。由（48）可知， T 是一个 γ-压缩。为了将其表示为所需的逐分量形式，注意对任意分量(s, a):

$$|(TQ - TQ')(s,a)| \le \|TQ - TQ'\|_\infty \tag{70}$$
$$\le \gamma \|Q - Q'\|_\infty \tag{71}$$
$$= \gamma \max_{(s',a')} |Q(s',a') - Q'(s',a')| \tag{72}$$

根据我们的假设:

- 步长 α_t 满足 Robbins-Monro 条件
- 每个状态-动作对都会被访问无穷多次
- 时变算子 T_t 收敛到 T，因为 $\varepsilon_t \to 0$

因此，定理17 的所有条件均已满足（略作调整：我们的噪声项包含一个渐消的确定性分量，但这不会影响收敛论证）。因此，$Q_t \to Q^\star$ 几乎处处成立。

□

步骤 6：有限时间高概率界。 我们现在推导定理第二部分所述的有限时间高概率界。该分析更为复杂，需要精确跟踪误差的演化。

让我们定义加权误差序列:

$$\Delta_{t+1} = (I - \alpha_t e_{(s_t,a_t)} e_{(s_t,a_t)}^\top)\Delta_t + \alpha_t e_{(s_t,a_t)} \left[(TQ_t - TQ^\star)(s_t,a_t) + ((T_t - T)Q_t)(s_t,a_t) + \delta_t\right]$$

对于被更新的分量，我们有:

$$\Delta_{t+1}(s_t,a_t) = (1-\alpha_t)\Delta_t(s_t,a_t) + \alpha_t \left[(TQ_t - TQ^\star)(s_t,a_t) + ((T_t - T)Q_t)(s_t,a_t) + \delta_t\right] \tag{73}$$
$$\le (1-\alpha_t)\Delta_t(s_t,a_t) + \alpha_t \left[\gamma\|\Delta_t\|_\infty + C_1\varepsilon_t + \delta_t\right] \tag{74}$$

其中我们利用了如下事实： $|(TQ_t - TQ^\star)(s_t,a_t)| \le \gamma\|\Delta_t\|_\infty$ 这是由于压缩性质，并且$|((T_t - T)Q_t)(s_t,a_t)| \le C_1\varepsilon_t$ 来自引理 14。

Lemma 19 （逐分量动力学）。 对于在时间被更新的状态-动作对 (s_t, a_t) t:

$$|\Delta_{t+1}(s_t,a_t)| \le (1 - \alpha_t(1-\gamma))|\Delta_t(s_t,a_t)| + \alpha_t\left[\gamma(\|\Delta_t\|_\infty - |\Delta_t(s_t,a_t)|) + C_1\varepsilon_t + \delta_t\right] \tag{75}$$

引理的证明*19.* 从我们先前的不等式开始:

$$|\Delta_{t+1}(s_t,a_t)| \le (1-\alpha_t)|\Delta_t(s_t,a_t)| + \alpha_t\left[\gamma\|\Delta_t\|_\infty + C_1\varepsilon_t + \delta_t\right] \tag{76}$$
$$= (1-\alpha_t)|\Delta_t(s_t,a_t)| + \alpha_t\gamma|\Delta_t(s_t,a_t)| + \alpha_t\left[\gamma(\|\Delta_t\|_\infty - |\Delta_t(s_t,a_t)|) + C_1\varepsilon_t + \delta_t\right] \tag{77}$$
$$= (1-\alpha_t(1-\gamma))|\Delta_t(s_t,a_t)| + \alpha_t\left[\gamma(\|\Delta_t\|_\infty - |\Delta_t(s_t,a_t)|) + C_1\varepsilon_t + \delta_t\right] \tag{78}$$

由此得到所需的不等式。 □

对于一系列更新，我们需要追踪最大误差 $\|\Delta_t\|_\infty$ 的变化。我们用 $\tau_k(s,\ a)$ 表示状态-动作对 $(s,\ a)$ 第 k 次被更新的时间。

引理 20 (遍历过程中的误差缩减)。令 $\tau(t) = $最小$\{k > t:\ \ \forall(s,a) \in S \times A, \exists j \in \{t+1,\ .....,\ k\}$ 使得 $(s_j, a_j) = (s,a)\}$ 为在 t 之后访问所有状态-动作对至少一次所需的时间。则对于任意 $t \geq 1$:

$$\|\Delta_{\tau(t)}\|_\infty \leq \left(1 - \min_{(s,a)\in S\times A} \alpha_{\tau_1(s,a,t)}(1-\gamma)\right)\|\Delta_t\|_\infty + C_1 \max_{t\leq i\leq \tau(t)} \varepsilon_i + M_t \tag{79}$$

其中:

- $\tau_1(s,a,t) = \min\{k > t : (s_k, a_k) = (s,a)\}$ *is the first time after* t *that* (s,a) *is updated*
- $M_t = \max_{(s,a)\in S\times A} \left|\sum_{i=t+1}^{\tau(t)} \mathbb{1}\{(s_i,a_i) = (s,a)\}\alpha_i\delta_i\right|$ *is the maximum accumulated noise term*

证明。我们分析最大范数误差如何随着每个状态-动作对的更新而演变。令 (s_t^*, a_t^*) 为时间 t 时具有最大误差的状态-动作对，即 $|\Delta_t(s_t^*, a_t^*)| = \|\Delta_t\|_\infty$。

首先，我们回顾引理 19 指出，对于任意状态-动作对 (s,a) 在时间 i更新:

$$|\Delta_{i+1}(s,a)| \leq (1-\alpha_i(1-\gamma))|\Delta_i(s,a)| + \alpha_i\left[\gamma(\|\Delta_i\|_\infty - |\Delta_i(s,a)|) + C_1\varepsilon_i + \delta_i\right] \tag{80}$$

现在我们跟踪最大范数在完整遍历更新过程中的变化。对于每个 $(s,\ a) \in S\times A$，令 $\tilde{\tau}(s,\ a)$ 表示在 $(s,\ a)$ 更新后，所有状态-动作对都至少更新过一次之后的时间。也就是说，$\tilde{\tau}(s,\ a) = \tau(\tau_1(s,\ a,\ t))$。

情况 1: 如果完整遍历后，原先误差最大的分量仍然具有最大误差: $\|\Delta_{\tau(t)}\|_\infty = |\Delta_{\tau(t)}(s_t^*,\ a_t^*)|$。

那么我们有:

$$\begin{aligned}
\|\Delta_{\tau(t)}\|_\infty &= |\Delta_{\tau(t)}(s_t^*, a_t^*)| && (81)\\
&= |\Delta_{\tau_1(s_t^*,a_t^*,t)+1}(s_t^*, a_t^*)| \quad \text{(since no further updates to this component)} && (82)\\
&\leq (1-\alpha_{\tau_1(s_t^*,a_t^*,t)}(1-\gamma))|\Delta_{\tau_1(s_t^*,a_t^*,t)}(s_t^*, a_t^*)| && (83)\\
&\quad + \alpha_{\tau_1(s_t^*,a_t^*,t)}\left[\gamma(\|\Delta_{\tau_1(s_t^*,a_t^*,t)}\|_\infty - |\Delta_{\tau_1(s_t^*,a_t^*,t)}(s_t^*, a_t^*)|) + C_1\varepsilon_{\tau_1(s_t^*,a_t^*,t)} + \delta_{\tau_1(s_t^*,a_t^*,t)}\right] && (84)
\end{aligned}$$

由于 $|\Delta_{\tau_1(s_t^*,\ a_t^*,\ t)}(s_t^*,\ a_t^*)| \leq \|\Delta_t\|_\infty$（该分量的误差不会在未更新时自行增大），且 $\|\Delta_{\tau_1(s_t^*,\ a_t^*,\ t)}\|_\infty \leq \|\Delta_t\|_\infty$（最大误差在未更新时不会增大），方括号中的第二项非正。因此:

$$\begin{aligned}
\|\Delta_{\tau(t)}\|_\infty &\leq (1-\alpha_{\tau_1(s_t^*,a_t^*,t)}(1-\gamma))|\Delta_{\tau_1(s_t^*,a_t^*,t)}(s_t^*, a_t^*)| + \alpha_{\tau_1(s_t^*,a_t^*,t)}[C_1\varepsilon_{\tau_1(s_t^*,a_t^*,t)} + \delta_{\tau_1(s_t^*,a_t^*,t)}] && (85)\\
&\leq (1-\alpha_{\tau_1(s_t^*,a_t^*,t)}(1-\gamma))\|\Delta_t\|_\infty + \alpha_{\tau_1(s_t^*,a_t^*,t)}[C_1\varepsilon_{\tau_1(s_t^*,a_t^*,t)} + \delta_{\tau_1(s_t^*,a_t^*,t)}] && (86)
\end{aligned}$$

情况 2: 如果完整遍历后，另一个分量 $(s',\ a') \neq (s_t^*,\ a_t^*)$具有最大误差: $\|\Delta_{\tau(t)}\|_\infty = |\Delta_{\tau(t)}(s',\ a')|$。

令 $\tau_1 = \tau_1(s',\ a',\ t)$ 表示 $(s',\ a')$ 的更新时间。我们有:

$$\begin{aligned}
\|\Delta_{\tau(t)}\|_\infty &= |\Delta_{\tau(t)}(s', a')| && (87)\\
&= |\Delta_{\tau_1+1}(s', a')| \quad \text{(since no further updates to this component)} && (88)\\
&\leq (1-\alpha_{\tau_1}(1-\gamma))|\Delta_{\tau_1}(s', a')| + \alpha_{\tau_1}\left[\gamma(\|\Delta_{\tau_1}\|_\infty - |\Delta_{\tau_1}(s', a')|) + C_1\varepsilon_{\tau_1} + \delta_{\tau_1}\right] && (89)
\end{aligned}$$

由于 $|\Delta_{\tau_1}(s',\ a')| \le \|\Delta_{\tau_1}\|_\infty$，我们有：

$$\|\Delta_{\tau(t)}\|_\infty \le (1-\alpha_{\tau_1}(1-\gamma))\|\Delta_{\tau_1}\|_\infty + \alpha_{\tau_1}[C_1\varepsilon_{\tau_1} + \delta_{\tau_1}] \tag{90}$$

Now, $\|\Delta_{\tau_1}\|_\infty \le \|\Delta_t\|_\infty$ （因为最大范数只会在未更新时减小），这给出 s:

$$\|\Delta_{\tau(t)}\|_\infty \le (1-\alpha_{\tau_1}(1-\gamma))\|\Delta_t\|_\infty + \alpha_{\tau_1}[C_1\varepsilon_{\tau_1} + \delta_{\tau_1}] \tag{91}$$

收缩因子的严格推导。

现在，我们给出一个严格的界，从而无需使用启发式因子。关键在于，情况1和情况2中的分析都得出了形式为 $(1-\alpha(1-\gamma))$ 的收缩因子，但它们作用于扫描过程中不同时间的不同分量。

关键观测：在两种情况下，在时间 $\tau(t)$ 达到最大误差的分量（情况1中的 s^*, a^* 或情况2中的 s', a'）都满足：

- 它在扫描过程中恰好更新了一次（在某个时间 $\tau_1 \in \{t+1, \ldots, \tau(t)\}$）
- 更新时，它的误差有界于 $\|\Delta_t\|_\infty$ （因为没有更新时，任何分量的误差都不会自行增大）
- 更新后，不再对这个分量进行修改

从分 量动态来看，在最终成为最大分量的分量的更新时间 τ_1 t:

$$|\Delta_{\tau_1+1}(\text{max-comp})| \le (1-\alpha_{\tau_1}(1-\gamma))|\Delta_{\tau_1}(\text{max-comp})| \tag{92}$$

$$+\alpha_{\tau_1} \underbrace{\gamma(\|\Delta_{\tau_1}\|_\infty - |\Delta_{\tau_1}(\text{max-comp})|)}_{\ge 0,\ \text{captures "spillover" from other components}} + \alpha_{\tau_1}[C_1\varepsilon_{\tau_1} + \delta_{\tau_1}] \tag{93}$$

使用$|\Delta_{\tau_1}$（最大分量）$| \le \|\Delta_t\|_\infty$ 和 $\|\Delta_{\tau_1}\|_\infty \le \|\Delta_t\|_\infty$:

$$|\Delta_{\tau_1+1}(\text{max-comp})| \le (1-\alpha_{\tau_1}(1-\gamma))\|\Delta_t\|_\infty + \alpha_{\tau_1}\gamma(\|\Delta_t\|_\infty - |\Delta_{\tau_1}(\text{max-comp})|) + \alpha_{\tau_1}[C_1\varepsilon_{\tau_1} + \delta_{\tau_1}] \tag{94}$$

在最坏情况下，$|\Delta_{\tau_1}$（最大分量）$| = 0$，我们得到：

$$|\Delta_{\tau_1+1}(\text{max-comp})| \le (1-\alpha_{\tau_1}(1-\gamma) + \alpha_{\tau_1}\gamma)\|\Delta_t\|_\infty + \alpha_{\tau_1}[C_1\varepsilon_{\tau_1} + \delta_{\tau_1}] \tag{95}$$

$$= (1-\alpha_{\tau_1}(1-2\gamma))\|\Delta_t\|_\infty + \alpha_{\tau_1}[C_1\varepsilon_{\tau_1} + \delta_{\tau_1}] \tag{96}$$

对于 $\gamma < 1/2$，这给出了直接收缩。对于 $\gamma \ge 1/2$，我们采用一种精细论证，其依据是这样的观测：溢出项 $\gamma(\|\Delta_{\tau_1}\|_\infty - |\Delta_{\tau_1}$（最大分量）$|)$ 只有在 $|\Delta_{\tau_1}$（最大分量）$|$ 远小于最大值时才会很大；但此时需要收缩的基线误差也更小。

$\gamma \ge 1/2$的精细界：

定义 $\rho := |\Delta_{\tau_1}$（最大组合）$|/\|\Delta_t\|_\infty \in [0,1]$。则：

$$|\Delta_{\tau_1+1}(\text{max-comp})| \le (1-\alpha_{\tau_1}(1-\gamma))\rho\|\Delta_t\|_\infty + \alpha_{\tau_1}\gamma(1-\rho)\|\Delta_t\|_\infty + \alpha_{\tau_1}[C_1\varepsilon_{\tau_1} + \delta_{\tau_1}] \tag{97}$$

$$= \big[\rho - \alpha_{\tau_1}(1-\gamma)\rho + \alpha_{\tau_1}\gamma(1-\rho)\big]\|\Delta_t\|_\infty + \alpha_{\tau_1}[C_1\varepsilon_{\tau_1} + \delta_{\tau_1}] \tag{98}$$

$$= \big[\rho(1-\alpha_{\tau_1}) + \alpha_{\tau_1}\gamma\big]\|\Delta_t\|_\infty + \alpha_{\tau_1}[C_1\varepsilon_{\tau_1} + \delta_{\tau_1}] \tag{99}$$

系数 $\rho(1-\alpha_{\tau_1}) + \alpha_{\tau_1}\gamma$ 在 $\rho = 1$ 处达到最大值（因为 $\alpha_{\tau_1} \le 1$），得到：

$$\max_{\rho\in[0,1]}\big[\rho(1-\alpha_{\tau_1}) + \alpha_{\tau_1}\gamma\big] = (1-\alpha_{\tau_1}) + \alpha_{\tau_1}\gamma = 1 - \alpha_{\tau_1}(1-\gamma) \tag{100}$$

这在没有任何损失的情况下恢复了原始收缩因子（$1-\alpha_{\tau_1}(1-\gamma)$）！因此，结合这两种情况，并取所有分量中的最小步长：

$$\|\Delta_{\tau(t)}\|_\infty \le \left(1 - \min_{(s,a)\in S\times A} \alpha_{\tau_1(s,a,t)}(1-\gamma)\right)\|\Delta_t\|_\infty + C_1 \max_{t\le i\le \tau(t)} \varepsilon_i + M_t \tag{101}$$

其中 $M_t = \max_{(s,a)\in S\times A}\left|\sum_{i=t+1}^{\tau(t)} \mathbb{1}\{(s_i,\ a_i) = (s,\ a)\}\alpha_i\delta_i\right|$ 是最大累积噪声。

备注：上述精细论证表明，可以无损地达到完整收缩因子$(1-\gamma)$，从而得到比采用保守因子 $1/2$ 的分析更紧的收敛界。 □

现在，采用步长 $\alpha_t = 1/t$，我们有：

引理 21（步长性质）。令 $\alpha_t = 1/t$:

(a) $\prod_{i=1}^{t}(1-\alpha_i(1-\gamma)) \le (t+1)^{-(1-\gamma)}$

(b) For any $i \le t$: $\prod_{j=i+1}^{t}(1-\alpha_j(1-\gamma)) \le \left(\frac{t+1}{i+1}\right)^{-(1-\gamma)}$

引理 *21* 的证明。利用不等式 $1-x \le e^{-x}$ 对所有 $x\in[0,1]$:

$$\prod_{i=1}^{t}(1-\alpha_i(1-\gamma)) \le \prod_{i=1}^{t} e^{-\alpha_i(1-\gamma)} \tag{102}$$

$$= \exp\left(-(1-\gamma)\sum_{i=1}^{t}\alpha_i\right) \tag{103}$$

$$= \exp\left(-(1-\gamma)\sum_{i=1}^{t}\frac{1}{i}\right) \tag{104}$$

$$\le \exp\left(-(1-\gamma)\ln(t+1)\right) \tag{105}$$

$$= (t+1)^{-(1-\gamma)} \tag{106}$$

类似地，对于 (b) 部分：

$$\prod_{j=i+1}^{t}(1-\alpha_j(1-\gamma)) \le \exp\left(-(1-\gamma)\sum_{j=i+1}^{t}\frac{1}{j}\right) \tag{107}$$

$$\le \exp\left(-(1-\gamma)\ln\left(\frac{t+1}{i+1}\right)\right) \tag{108}$$

$$= \left(\frac{t+1}{i+1}\right)^{-(1-\gamma)} \tag{109}$$

□

为了推导我们的有限 时间界，我们需要同时控制确定性偏差和鞅噪声 .

引理 22（通过 Freedman 不等式得到鞅集中界）。设 $\{X_i\}$ 是相对于滤过 $\{\mathcal{F}_i\}$ 的鞅差序列，并假设 $|X_i| \le b_i$ 几乎处处成立。令 $S_n = \sum_{i=1}^{n} X_i$。那么，对每个 $x > 0$:

$$\Pr(|S_n| \ge x) \le \exp\left(-\frac{x^2/2}{\sum_{i=1}^{n}\mathbb{E}[X_i^2 \mid \mathcal{F}_{i-1}] + \frac{x}{3}\max_{i\le n} b_i}\right)$$

让我们应用 Freedman 不等式来界定累积噪声项。

引理 23 (鞅项界**)。** 定义加权噪声累积：

$$M_t = \sum_{i=1}^{t} \alpha_i e_{(s_i,a_i)} \delta_i$$

对于任意 $\delta \in (0,1)$ 和 $t \geq 1$，以至少 $1-\delta$ 的概率：

$$\|M_t\|_\infty \leq 2V_{\max} \sqrt{2 \ln\left(\frac{2|S||A|}{\delta}\right) \sum_{i=1}^{t} \alpha_i^2} + \frac{2V_{\max}}{3} \ln\left(\frac{2|S||A|}{\delta}\right) \max_{i \leq t} \alpha_i$$

引理 *23*的证明。 对于每个 (s,a)，定义：

$$M_t(s,a) = \sum_{i=1}^{t} \mathbb{1}\{(s_i,a_i) = (s,a)\} \alpha_i \delta_i$$

这是关于 $\{\mathcal{F}_t\}$ 的鞅。根据引理 15，我们知道 $|\delta_i| \leq 2V_{\max}$，因此每一项$\mathbb{1}\{(s_i,a_i) = (s,a)\}\alpha_i\delta_i$都有界，界为 $\alpha_i \cdot 2V_{\max}$。

应用引理 22并利用条件方差的界：

$$\mathbb{E}[(\mathbb{1}\{(s_i,a_i) = (s,a)\}\alpha_i\delta_i)^2 \mid \mathcal{F}_{i-1}] \leq (2V_{\max})^2 \alpha_i^2 \mathbb{1}\{(s_i,a_i) = (s,a)\}$$

我们得到，以至少 $1 - \frac{\delta}{|S||A|}$的概率：——

$$|M_t(s,a)| \leq \sqrt{2 \ln\left(\frac{2|S||A|}{\delta}\right) (2V_{\max})^2 \sum_{i=1}^{t} \alpha_i^2 \mathbb{1}\{(s_i,a_i) = (s,a)\}} \tag{110}$$

$$+ \frac{2V_{\max}}{3} \ln\left(\frac{2|S||A|}{\delta}\right) \max_{i \leq t} \alpha_i \tag{111}$$

$$\leq 2V_{\max} \sqrt{2 \ln\left(\frac{2|S||A|}{\delta}\right) \sum_{i=1}^{t} \alpha_i^2} + \frac{2V_{\max}}{3} \ln\left(\frac{2|S||A|}{\delta}\right) \max_{i \leq t} \alpha_i \tag{112}$$

根据并集界，上述结论以至少 $1-\delta$ 的概率对所有 $(s,\ a) \in S \times A$ 成立，从而得到关于 $\|M_t\|_\infty$的界。 □

对于加权折扣噪声项，我们需要进行更精细的分析。

引理24 (加权折扣噪声分析**)。** 定义加权折扣噪声累积量：

$$\tilde{M}_t = \sum_{i=1}^{t} \alpha_i e_{(s_i,a_i)} \delta_i \prod_{j=i+1}^{t} (1 - \alpha_j(1-\gamma))$$

对于任意 $\delta \in (0,\ 1)$ 和 $t \geq 1$，以至少 $1-\delta$ 的概率：

$$\|\tilde{M}_t\|_\infty \leq C \cdot t^{-1/2} \sqrt{\ln\left(\frac{2|S||A|t}{\delta}\right)}$$

对于某个常数 $C > 0$，它仅取决于 $V_{\max}$ 和 γ。

˜证明。分析 M_t 的关键难点在于，由于折扣因子的存在，它不是标准的鞅。我们通过使用分解方法来解决这一问题。

步骤 1：按状态-动作对分解。

首先，我们将 $\tilde{M}_t$ 按状态-动作对分解：

$$\tilde{M}_t(s,a) = \sum_{i=1}^{t} \mathbb{1}\{(s_i,a_i)=(s,a)\}\alpha_i\delta_i \prod_{j=i+1}^{t}(1-\alpha_j(1-\gamma))$$

这样我们就得到 $|S||A|$ 个彼此独立的加权和，每个状态-动作对对应一个。根据并集界，如果我们能以概率 $1-\frac{\delta}{|S||A|}$ 界定每个 $|\tilde{M}_t(s,a)|$，那么就能以概率 $1-\delta$ 界定 $\|\tilde{M}_t\|_\infty$。

步骤 2：时间分割与倍增时段。

为了处理时变权重，我们采用时间分割技术，并使用倍增时段。定义各个时期：

$$\mathcal{T}_0 = \{1\} \tag{113}$$

$$\mathcal{T}_k = \{2^{k-1}+1, 2^{k-1}+2, \ldots, 2^k\} \quad \text{for } k \geq 1 \tag{114}$$

对于每个时期 $\mathcal{T}_k$ 和状态-动作对 $(s,\ a)$，定义：

$$\tilde{M}_{\mathcal{T}_k}(s,a) = \sum_{i\in\mathcal{T}_k} \mathbb{1}\{(s_i,a_i)=(s,a)\}\alpha_i\delta_i \prod_{j=i+1}^{t}(1-\alpha_j(1-\gamma))$$

于是：

$$\tilde{M}_t(s,a) = \sum_{k=0}^{\lfloor \log_2 t\rfloor} \tilde{M}_{\mathcal{T}_k}(s,a)$$

步骤 3：分析每个时期内的权重。

对于 $i\in\mathcal{T}_k$，我们有 $i \geq 2^{k-1}+1$。根据引理 21：

$$\prod_{j=i+1}^{t}(1-\alpha_j(1-\gamma)) \leq \left(\frac{t+1}{i+1}\right)^{-(1-\gamma)} \tag{115}$$

$$\leq \left(\frac{t+1}{2^{k-1}+2}\right)^{-(1-\gamma)} \tag{116}$$

对于 $i\in\mathcal{T}_k$，权重之间最多相差一个常数因子：

$$\frac{\prod_{j=i+1}^{t}(1-\alpha_j(1-\gamma))}{\prod_{j=i'+1}^{t}(1-\alpha_j(1-\gamma))} \leq \left(\frac{i'+1}{i+1}\right)^{1-\gamma} \leq \left(\frac{2^k+1}{2^{k-1}+2}\right)^{1-\gamma} \leq 2^{1-\gamma} \tag{117}$$

对于任意 $i,\ i'\in\mathcal{T}_k$。这使我们能够用一个常数近似一个时期内的权重。

步骤 4：为每个时期定义一个修正鞅。

对于每个时期 $\mathcal{T}_k$ 和状态-动作对（$s,\ a$），定义：

$$W_{\mathcal{T}_k}(s,a) = \sum_{i\in\mathcal{T}_k} \mathbb{1}\{(s_i,a_i)=(s,a)\}\alpha_i\delta_i$$

这是一个鞅差分和。根据 Freedman 不等式（引理 22），以至少$1-\frac{\delta}{(|S||A|)(\lfloor\log_2 t\rfloor+1)}$的概率：

$$|W_{\mathcal{T}_k}(s,a)| \le 2V_{\max}\sqrt{2\ln\left(\frac{4(|S||A|)(\lfloor\log_2 t\rfloor+1)}{\delta}\right)\sum_{i\in\mathcal{T}_k}\alpha_i^2\mathbb{1}\{(s_i,a_i)=(s,a)\}} \tag{118}$$

$$+\frac{2V_{\max}}{3}\ln\left(\frac{4(|S||A|)(\lfloor\log_2 t\rfloor+1)}{\delta}\right)\max_{i\in\mathcal{T}_k}\alpha_i \tag{119}$$

步骤 5：将加权和与鞅联系起来。

对于每个时期 $\mathcal{T}_k$，我们可以将 $\tilde{M}_{\mathcal{T}_k}(s,a)$ 与 $W_{\mathcal{T}_k}(s,a)$ 联系起来：

$$|\tilde{M}_{\mathcal{T}_k}(s,a)| \le \max_{i\in\mathcal{T}_k}\prod_{j=i+1}^{t}(1-\alpha_j(1-\gamma))\cdot|W_{\mathcal{T}_k}(s,a)| \tag{120}$$

$$\le \left(\frac{t+1}{2^{k-1}+2}\right)^{-(1-\gamma)}\cdot|W_{\mathcal{T}_k}(s,a)| \tag{121}$$

步骤 6：分析所有周期上的求和。

根据并集界，以至少 $1-\frac{\delta}{2|S||A|}$ 的概率，对所有周期 $\mathcal{T}_k$：

$$|\tilde{M}_{\mathcal{T}_k}(s,a)| \le \left(\frac{t+1}{2^{k-1}+2}\right)^{-(1-\gamma)}\cdot C_2\sqrt{\ln\left(\frac{|S||A|t\log_2(t)}{\delta}\right)}\cdot\sqrt{\sum_{i\in\mathcal{T}_k}\alpha_i^2} \tag{122}$$

其中 C_2 是一个常数。对于 $i\in\mathcal{T}_k$，$\alpha_i=\frac{1}{i}\le\frac{1}{2^{k-1}+1}$。因此：

$$\sum_{i\in\mathcal{T}_k}\alpha_i^2 \le \sum_{i\in\mathcal{T}_k}\frac{1}{(2^{k-1}+1)^2} \tag{123}$$

$$=\frac{|\mathcal{T}_k|}{(2^{k-1}+1)^2} \tag{124}$$

$$=\frac{2^k-2^{k-1}}{(2^{k-1}+1)^2} \tag{125}$$

$$=\frac{2^{k-1}}{(2^{k-1}+1)^2} \tag{126}$$

$$\le\frac{1}{2^{k-1}} \tag{127}$$

因此：

$$|\tilde{M}_{\mathcal{T}_k}(s,a)| \le \left(\frac{t+1}{2^{k-1}+2}\right)^{-(1-\gamma)}\cdot C_2\sqrt{\ln\left(\frac{|S||A|t\log_2(t)}{\delta}\right)}\cdot\frac{1}{\sqrt{2^{k-1}}} \tag{128}$$

$$=C_2\sqrt{\ln\left(\frac{|S||A|t\log_2(t)}{\delta}\right)}\cdot\frac{(t+1)^{-(1-\gamma)}}{(2^{k-1}+2)^{-(1-\gamma)}}\cdot\frac{1}{\sqrt{2^{k-1}}} \tag{129}$$

$$\approx C_2\sqrt{\ln\left(\frac{|S||A|t\log_2(t)}{\delta}\right)}\cdot(t+1)^{-(1-\gamma)}\cdot(2^{k-1})^{(1-\gamma)}\cdot(2^{k-1})^{-1/2} \tag{130}$$

$$=C_2\sqrt{\ln\left(\frac{|S||A|t\log_2(t)}{\delta}\right)}\cdot(t+1)^{-(1-\gamma)}\cdot(2^{k-1})^{(1-\gamma)-1/2} \tag{131}$$

步骤 7：对所有周期求和。

最后，对所有周期求和：

$$|\tilde{M}_t(s,a)| \le \sum_{k=0}^{\lfloor \log_2 t \rfloor} |\tilde{M}_{\mathcal{T}_k}(s,a)| \tag{132}$$

$$\le C_2 \sqrt{\ln\left(\frac{|S||A|t\log_2(t)}{\delta}\right)} \cdot (t+1)^{-(1-\gamma)} \cdot \sum_{k=0}^{\lfloor \log_2 t \rfloor} (2^{k-1})^{(1-\gamma)-1/2} \tag{133}$$

指数 $(1-\gamma)-1/2$ 决定了求和的行为。当 $(1-\gamma)-1/2<0$ 时（即$\gamma>1/2$ 时），求和由最大项（最后一个时期）主导，约为 $(t/2)^{(1-\gamma)-1/2}=O(t^{(1-\gamma)-1/2})$。对于接近 1 的典型折扣因子，我们有 $\gamma>1/2$，因此：

$$|\tilde{M}_t(s,a)| \le C_3 \sqrt{\ln\left(\frac{|S||A|t\log_2(t)}{\delta}\right)} \cdot (t+1)^{-(1-\gamma)} \cdot t^{(1-\gamma)-1/2} \tag{134}$$

$$= C_3 \sqrt{\ln\left(\frac{|S||A|t\log_2(t)}{\delta}\right)} \cdot t^{-1/2} \tag{135}$$

根据并集界，以至少 $1-\delta/2$ 的概率，该结论对所有状态-动作对 $(s,\ a)$ 均成立。由于 $\ln(t\log_2(t))=O(\ln(t))$，我们得到：

$$\|\tilde{M}_t\|_\infty \le C \cdot t^{-1/2} \sqrt{\ln\left(\frac{2|S||A|t}{\delta}\right)} \tag{136}$$

以至少 $1-\delta$ 的概率成立。 □

现在我们可以推导有限时间高概率界。

命题 25 (有限时间高概率界)。 在定理 *13*的假设下，取 $\alpha_t=1/t$ 且 $\varepsilon_t=O(t^{-\beta})$，对于任意 $\delta\in(0,\ 1)$ 以及 $t\ge 1$，以至少 $1-\delta$ 的概率：

$$\|Q_t-Q^\star\|_\infty = O\left(t^{-\min(\beta,1/2)}\sqrt{\ln\left(\frac{|S||A|t}{\delta}\right)}\right)$$

命题*25*的证明。 我们将利用状态-动作空间上的完整扫描序列来分析收敛速率。首先，我们需要仔细刻画探索策略如何影响完成完整扫描所需的时间。

覆盖时间分析：定义如下：

- $\tau(t)$：在 t 之后至少访问一次所有状态-动作对所需的时间
- $\{t_k\}_{k\ge 0}$：一个序列，其中 $t_0=1$ 和 $t_{k+1}=\tau(t_k)$，对 $k\ge 0$

对于在遍历马尔可夫决策过程中采用固定 $\epsilon>0$ 的 ϵ-贪心策略，我们可以利用马尔可夫链理论的结果来界定访问所有状态-动作对的期望时间：

引理 26 (非平稳马尔可夫决策过程中的覆盖时间)。 考虑一个状态空间为 S、动作空间为 A 的马尔可夫决策过程，其中智能体遵循固定 $\epsilon>0$的 ϵ-贪心策略。设原始转移核为 P，$\widetilde{\quad}$，并设 P_t 为一个非平稳转移核，满足 $\|P_t(\cdot|s,a)-P(\cdot|s,a)\|_1\le\varepsilon_t$，对所有$(s,a)\in S\times A$均成立，其中 $\varepsilon_t\to 0$当 $t\to\infty$时。

假设核为 P 的原始马尔可夫决策过程是遍历的。那么存在一个时间 T_0，使得对于所有满足 $\varepsilon_t \leq \frac{\epsilon}{4|S|}$ 的 $t \geq T_0$，以下结论成立：

对于任意 $\delta \in (0, 1)$ 以及任意 $t \geq T_0$，以下结论以至少 $1 - \frac{\delta}{2t^2}$ 的概率成立：

$$\tau(t) - t \leq C_4 |S|^2 |A|^2 \log(|S||A|t/\delta) \tag{137}$$

其中，$\tau(t)$ 表示在 t 之后至少访问一次所有状态-动作对所需的时间，而 C_4 是一个取决于 ϵ 和原始马尔可夫决策过程谱性质的常数。

引理的证明26。关键挑战在于，我们处理的是一个具有状态转移~的非平稳环境，P_t 而非 P。我们需要分析近似误差 ε_t 如何影响遍历性。

首先，我们回顾一些马尔可夫链理论中的事实：

- 对于转移矩阵为 P、平稳分布为 π的遍历马尔可夫链，其混合时间为 $t_{mix} = O(\frac{1}{1-\lambda_2} \log \frac{1}{\epsilon_{min}})$，其中 λ_2 是第二大特征值，$\epsilon_{min} = \min_s \pi(s)$。
- 对于遍历马尔可夫决策过程中的 ϵ-贪心策略，我们有 $\epsilon_{min} \geq \frac{\epsilon}{|S||A|}$。
- 马尔可夫链的扰动理论告诉我们，如果 $\|P - P'\|_1 \leq \delta$（逐元素），那么平稳分布 π 和 π' 满足 $\|\pi - \pi'\|_1 = O(\frac{\delta}{1-\lambda_2})$。

当 ε_t 足够小时（具体而言，$\varepsilon_t \leq \frac{\epsilon}{4|S|}$），由 P_t 确定、具有状态转移~的受扰动的马尔可夫链仍然是遍历的，其平稳分布 $\widetilde{\pi}_t$ 满足：

$$\|\widetilde{\pi}_t - \pi\|_1 = O\left(\frac{\varepsilon_t}{1 - \lambda_2}\right) \tag{138}$$

This 这意味着，对于 $t \geq T_0$，其中 T_0 的选取满足：对所有 $t \geq T_0$，均有 $\varepsilon_t \leq \frac{\epsilon}{4|S|}$，我们有：

$$\min_{(s,a)} \widetilde{\pi}_t(s, a) \geq \frac{\epsilon}{2|S||A|} \tag{139}$$

利用马尔可夫链中集邮者问题的标准结果，以至少 $1 - \frac{\delta}{2t^2}$ 的概率访问所有状态-动作对所需的时间有如下上界：

$$C_4 |S|^2 |A|^2 \log(|S||A|t/\delta) \tag{140}$$

其中 C_4 取决于 ϵ 以及原始马尔可夫决策过程的谱性质（尤其是 λ_2） P. □

误差随扫描次数的变化：根据引理20，对于任意 $k \geq 0$：

$$\|\Delta_{t_{k+1}}\|_\infty \leq \left(1 - \min_{(s,a) \in S \times A} \alpha_{\tau_1(s,a,t_k)}(1 - \gamma)\right) \|\Delta_{t_k}\|_\infty + C_1 \max_{t_k \leq i \leq t_{k+1}} \varepsilon_i + M_{t_k} \tag{141}$$

其中，$\tau_1(s, a, t_k) = \min\{j > t_k : (s_j, a_j) = (s, a)\}$ 是状态-动作对 (s, a) 在 t_k 之后首次更新的时间，而 M_{t_k} 表示最大累积噪声项：

$$M_{t_k} = \max_{(s,a) \in S \times A} \left| \sum_{i=t_k+1}^{t_{k+1}} \mathbb{1}\{(s_i, a_i) = (s, a)\} \alpha_i \delta_i \right| \tag{142}$$

根据引理 26，以至少 $1-\frac{\delta}{2t_k^2}$的概率，在t_k之后访问所有状态-动作对所需的时间有如下界：

$$\tau(t_k) - t_k \leq C_4|S|^2|A|^2\log(|S||A|t_k/\delta) \tag{143}$$

这意味着，对于任意状态-动作对（s,a）：

$$\tau_1(s,a,t_k) \leq t_k + C_4|S|^2|A|^2\log(|S||A|t_k/\delta) \tag{144}$$

步长为 $\alpha_t = 1/t$ 时，我们有：

$$\alpha_{\tau_1(s,a,t_k)} = \frac{1}{\tau_1(s,a,t_k)} \tag{145}$$

$$\geq \frac{1}{t_k + C_4|S|^2|A|^2\log(|S||A|t_k/\delta)} \tag{146}$$

$$\approx \frac{1}{t_k}\cdot\frac{1}{1+\frac{C_4|S|^2|A|^2\log(|S||A|t_k/\delta)}{t_k}} \tag{147}$$

当 t_k 较大且 $t_k \gg C_4|S|^2|A|^2\log(|S||A|t_k/\delta)$ 时，这近似为 $\frac{1}{t_k}(1-O(\frac{|S|^2|A|^2\log(|S||A|t_k/\delta)}{t_k}))$。将收缩因子定义为 $\rho_k = 1-\min_{(s,a)\in S\times A}\alpha_{\tau_1(s,a,t_k)}(1-\gamma)$，与引理 20 一致。根据我们对 $\alpha_{\tau_1(s,a,t_k)}$ 的界：

$$\rho_k = 1 - \min_{(s,a)\in S\times A}\alpha_{\tau_1(s,a,t_k)}(1-\gamma) \tag{148}$$

$$\leq 1 - \frac{(1-\gamma)}{t_k + C_4|S|^2|A|^2\log(|S||A|t_k/\delta)} \tag{149}$$

$$\approx 1 - \frac{(1-\gamma)}{t_k}\left(1 - O\left(\frac{|S|^2|A|^2\log(|S||A|t_k/\delta)}{t_k}\right)\right) \tag{150}$$

当 t_k 较大时，这近似为 $1-\frac{(1-\gamma)}{t_k}+O(\frac{|S|^2|A|^2\log(|S||A|t_k/\delta)}{t_k^2})$。

使用不等式 inequality $1-x\leq e^{-x}$（对 $x\in[0,1]$），并关注 t_k 较大时的主导项：

$$\rho_k \leq \exp\left(-\frac{(1-\gamma)}{t_k} + O\left(\frac{|S|^2|A|^2\log(|S||A|t_k/\delta)}{t_k^2}\right)\right) \tag{151}$$

$$\approx \exp\left(-\frac{(1-\gamma)}{t_k}\right)\cdot\exp\left(O\left(\frac{|S|^2|A|^2\log(|S||A|t_k/\delta)}{t_k^2}\right)\right) \tag{152}$$

$$\approx \exp\left(-\frac{(1-\gamma)}{t_k}\right)\cdot\left(1+O\left(\frac{|S|^2|A|^2\log(|S||A|t_k/\delta)}{t_k^2}\right)\right) \tag{153}$$

第二项会随着 t_k 增大而迅速趋近于 1，因此当 t_k较大时，$\rho_k \approx \exp\left(-\frac{(1-\gamma)}{t_k}\right) \approx 1-\frac{(1-\gamma)}{t_k}$。

递归应用公式 (141)，经过 k 次完整扫描：

$$\|\Delta_{t_k}\|_\infty \leq \left(\prod_{j=0}^{k-1}\rho_j\right)\|\Delta_{t_0}\|_\infty + \sum_{j=0}^{k-1}\left(\prod_{i=j+1}^{k-1}\rho_i\right)\left[C_1\max_{t_j\leq i\leq t_{j+1}}\varepsilon_i + M_{t_j}\right] \tag{154}$$

我们来分析每一项：

1. 初始误差衰减： $\left(\prod_{j=0}^{k-1}\rho_j\right)\|\Delta_{t_0}\|_\infty$

利用我们关于ρ_j的渐近界：

$$\prod_{j=0}^{k-1}\rho_j \approx \prod_{j=0}^{k-1}\exp\left(-\frac{(1-\gamma)}{t_j}\right) \tag{155}$$

$$= \exp\left(-(1-\gamma)\sum_{j=0}^{k-1}\frac{1}{t_j}\right) \tag{156}$$

根据引理26，我们知道 $t_{j+1} \approx t_j + C_4|S|^2|A|^2\log(|S||A|t_j/\delta)$。当 j较大时，可得$t_j = \Theta(j\cdot|S|^2|A|^2\log(|S||A|j/\delta))$。因此：

$$\sum_{j=0}^{k-1}\frac{1}{t_j} = \Theta\left(\sum_{j=1}^{k-1}\frac{1}{j\cdot|S|^2|A|^2\log(|S||A|j/\delta)}\right) \tag{157}$$

$$= \Theta\left(\frac{1}{|S|^2|A|^2}\sum_{j=1}^{k-1}\frac{1}{j\cdot\log(|S||A|j/\delta)}\right) \tag{158}$$

该和的渐近行为为 $\Theta\left(\log_{\log_{|S|^2|A|^2}k}\right)$。因此，初始误差衰减如下：

$$\prod_{j=0}^{k-1}\rho_j = \exp\left(-\Theta\left(\frac{(1-\gamma)\log\log k}{|S|^2|A|^2}\right)\right) \tag{159}$$

$$= (\log k)^{-\Theta\left(\frac{1-\gamma}{|S|^2|A|^2}\right)} \tag{160}$$

2. 近似误差项： $\sum_{j=0}^{k-1}\left(\prod_{i=j+1}^{k-1}\rho_i\right)C_1\max^{t_j\le i\le t_{j+1}}\varepsilon_i$

给定 $\varepsilon_t = O(t^{-\beta})$ 和 $t_j = \Theta(j\cdot|S|^2|A|^2\log(|S||A|j/\delta))$，我们有：

$$\max_{t_j\le i\le t_{j+1}}\varepsilon_i = O(t_j^{-\beta}) \tag{161}$$

$$= O\left((j\cdot|S|^2|A|^2\log(|S||A|j/\delta))^{-\beta}\right) \tag{162}$$

折扣因子 $\prod_{i=j+1}^{k-1}\rho_i$ 可界定为：

$$\prod_{i=j+1}^{k-1}\rho_i \approx \exp\left(-(1-\gamma)\sum_{i=j+1}^{k-1}\frac{1}{t_i}\right) \tag{163}$$

$$= \exp\left(-\Theta\left(\frac{(1-\gamma)}{|S|^2|A|^2}\log\frac{\log k}{\log j}\right)\right) \tag{164}$$

$$= \left(\frac{\log k}{\log j}\right)^{-\Theta\left(\frac{1-\gamma}{|S|^2|A|^2}\right)} \tag{165}$$

因此：

$$\sum_{j=0}^{k-1}\left(\prod_{i=j+1}^{k-1}\rho_i\right)C_1\max_{t_j\le i\le t_{j+1}}\varepsilon_i \tag{166}$$

$$=\sum_{j=0}^{k-1}\left(\frac{\log k}{\log j}\right)^{-\Theta\left(\frac{1-\gamma}{|S|^2|A|^2}\right)}\cdot O\left((j\cdot|S|^2|A|^2\log(|S||A|j/\delta))^{-\beta}\right) \tag{167}$$

$$\tag{168}$$

当 k 较大时，这个和主要由 j 接近 k 时的项决定，因此：

$$=O\left(k^{-\beta}\cdot(|S|^2|A|^2\log(|S||A|k/\delta))^{-\beta}\right) \tag{169}$$

3. 鞅噪声项： $\sum_{j=0}^{k-1}\left(\prod_{i=j+1}^{k-1}\rho_i\right)M_{t_j}$

根据引理23，以高概率：

$$M_{t_j}=O\left(t_j^{-1/2}\sqrt{\ln\left(\frac{|S||A|t_j}{\delta}\right)}\right) \tag{170}$$

$$=O\left((j\cdot|S|^2|A|^2\log(|S||A|j/\delta))^{-1/2}\sqrt{\ln\left(\frac{|S||A|t_j}{\delta}\right)}\right) \tag{171}$$

利用我们对折扣因子的界：

$$\sum_{j=0}^{k-1}\left(\prod_{i=j+1}^{k-1}\rho_i\right)M_{t_j} \tag{172}$$

$$=\sum_{j=0}^{k-1}\left(\frac{\log k}{\log j}\right)^{-\Theta\left(\frac{1-\gamma}{|S|^2|A|^2}\right)}\cdot O\left((j\cdot|S|^2|A|^2\log(|S||A|j/\delta))^{-1/2}\sqrt{\ln\left(\frac{|S||A|t_j}{\delta}\right)}\right) \tag{173}$$

当 k 较大时，这个和主要由 j 接近 k 时的项决定，因此：

$$=O\left(k^{-1/2}\cdot(|S|^2|A|^2\log(|S||A|k/\delta))^{-1/2}\sqrt{\ln\left(\frac{|S||A|t_k}{\delta}\right)}\right) \tag{174}$$

合并所有项：在近似误差项和鞅噪声项中取主导速率（衰减最慢的项）：

对于 $\beta>1/2$，鞅噪声项占主导：

$$\|\Delta_{t_k}\|_\infty=O\left(k^{-1/2}\cdot(|S|^2|A|^2\log(|S||A|k/\delta))^{-1/2}\sqrt{\ln\left(\frac{|S||A|t_k}{\delta}\right)}\right) \tag{175}$$

对于 $\beta<1/2$，近似误差项占主导：

$$\|\Delta_{t_k}\|_\infty=O\left(k^{-\beta}\cdot(|S|^2|A|^2\log(|S||A|k/\delta))^{-\beta}\right) \tag{176}$$

我们可以将这些情况合并为:

$$\|\Delta_{t_k}\|_\infty = O\left(k^{-\min(\beta,1/2)} \cdot (|S|^2|A|^2\log(|S||A|k/\delta))^{-\min(\beta,1/2)} \cdot \ln\left(\frac{|S||A|t_k}{\delta}\right)^{\mathbb{1}\{\beta\geq 1/2\}/2}\right) \quad (177)$$

其中 $\mathbb{1}\{\beta \geq 1/2\}$在 $\beta \geq 1/2$时取 1，否则取 0。

根据关系式 $t_k = \Theta(k \cdot |S|^2|A|^2\log(|S||A|k/\delta))$，我们可以用 t_k来表示:

$$\|\Delta_{t_k}\|_\infty = O\left(\left(\frac{t_k}{|S|^2|A|^2\log(|S||A|k/\delta)}\right)^{-\min(\beta,1/2)} \cdot (|S|^2|A|^2\log(|S||A|k/\delta))^{-\min(\beta,1/2)} \cdot \ln\left(\frac{|S||A|t_k}{\delta}\right)^{\mathbb{1}\{\beta\geq 1/2\}/2}\right) \quad (178)$$

$$= O\left(t_k^{-\min(\beta,1/2)} \cdot \ln\left(\frac{|S||A|t_k}{\delta}\right)^{\mathbb{1}\{\beta\geq 1/2\}/2}\right) \quad (179)$$

对于任意时间 $t \geq T_0$ （其中 T_0 确保了对覆盖时间的界）。我们可以找到 t_k 使得$t_k \leq t < t_{k+1}$，并且需要用 $\|\Delta_{t_k}\|_\infty$来界定 $\|\Delta_t\|_\infty$。根据 Q 学习更新以及引理16中的误差分解，对于任意在时间 i 更新的状态-动作对$(s,\ a)$，其中 $t_k \leq i < t$:

$$|\Delta_{i+1}(s,a)| \leq (1-\alpha_i(1-\gamma))|\Delta_i(s,a)| + \alpha_i\left[\gamma(\|\Delta_i\|_\infty - |\Delta_i(s,a)|) + C_1\varepsilon_i + \delta_i\right] \quad (180)$$

如果 $\|\Delta_i\|_\infty = |\Delta_i(s,\ a)|$（即，$(s,\ a)$在时间 i具有最大误差），那么:

$$|\Delta_{i+1}(s,a)| \leq (1-\alpha_i(1-\gamma))|\Delta_i(s,a)| + \alpha_i\left[C_1\varepsilon_i + \delta_i\right] \quad (181)$$

如果 $\|\Delta_i\|_\infty > |\Delta_i(s,\ a)|$，那么$(s,\ a)$处的误差可能会增大并趋近于最大误差。在区间 $[t_k, t)$上，高概率下，鞅噪声项有如下界:

$$\max_{t_k\leq i<t}|\delta_i| \leq O\left(\sqrt{\ln\left(\frac{(t-t_k)|S||A|}{\delta}\right)}\right) \quad (182)$$

鉴于 $t - t_k \leq C_4|S|^2|A|^2\log(|S||A|t_k/\delta)$ 和 $\varepsilon_i = O(i^{-\beta})$，我们可以证明，以高概率:

$$\|\Delta_t\|_\infty \leq \|\Delta_{t_k}\|_\infty + O\left(\frac{\log(|S||A|t_k/\delta)}{t_k} \cdot \sqrt{\ln\left(\frac{|S|^3|A|^3\log(|S||A|t_k/\delta)}{\delta}\right)}\right) \quad (183)$$

$$= \|\Delta_{t_k}\|_\infty + O\left(t_k^{-1/2}\right) \quad (184)$$

这一额外误差项不会影响我们最终界中的渐近速率，因为当 t_k 足够大时，它的衰减速度快于 $t_k^{-1/2}$。因此，只需对 $\|\Delta_{t_k}\|_\infty$ 求界。对于足够大的 t_k，有 $t = \Theta(t_k)$，因此:

$$\|\Delta_{t_k}\|_\infty = O\left(t_k^{-\min(\beta,1/2)} \cdot \ln\left(\frac{|S||A|t_k}{\delta}\right)^{\mathbb{1}\{\beta\geq 1/2\}/2}\right) \quad (185)$$

$$= O\left(t^{-\min(\beta,1/2)} \cdot \ln\left(\frac{|S||A|t}{\delta}\right)^{\mathbb{1}\{\beta\geq 1/2\}/2}\right) \quad (186)$$

因此，以至少 $1-\delta$ 的概率：

$$\|Q_t - Q^\star\|_\infty = O\left(t^{-\min(\beta,1/2)} \cdot \ln\left(\frac{|S||A|t}{\delta}\right)^{\mathbb{1}\{\beta \geq 1/2\}/2}\right) \tag{187}$$

对于 $\beta = 1/2$，可得：

$$\|Q_t - Q^\star\|_\infty = O\left(t^{-1/2} \cdot \sqrt{\ln\left(\frac{|S||A|t}{\delta}\right)}\right). \tag{188}$$

□

□

推论 27 (固定 DCM 有限时间界)。 当 *DCM* 经过预训练并保持固定时（$\theta_t = \theta^*$对所有 t），定义奖励近似误差 $\varepsilon_r := |\widetilde{r}_{\theta^*} - r|_\infty$ 以及转移近似误差 $\sim\varepsilon_P := \|P_{\theta^*} - P\|_1$。在命题 中取 *25*，则 Q 函数以至少 $1-\delta$的概率满足：

$$\|Q_t - Q^*\|_\infty = O\left(\frac{\varepsilon_r}{1-\gamma} + \frac{\gamma R_{\max}\varepsilon_P}{(1-\gamma)^2} + t^{-1/2}\sqrt{\ln\left(\frac{|S||A|t}{\delta}\right)}\right) \tag{189}$$

证明。 在固定 DCM 下，近似误差 ε_r 和 ε_P 为常数（对应于 $\beta = 0$）。根据引理14，单步贝尔曼误差满足：

$$\|T_t Q - TQ\|_\infty \leq |\widetilde{r}_t - r|_\infty + \gamma\|Q\|_\infty\|\widetilde{P}_t - P\|_1 \tag{190}$$

$$\leq \varepsilon_r + \gamma V_{\max}\varepsilon_P \tag{191}$$

其中 $V_{\max} = R_{\max}/(1-\gamma)$。随后，贝尔曼算子的压缩性质可得：

$$\|Q_t - Q^*\|_\infty \leq \frac{\varepsilon_r + \gamma V_{\max}\varepsilon_P}{1-\gamma} + C_2 t^{-1/2}\sqrt{\ln\left(\frac{|S||A|t}{\delta}\right)} \tag{192}$$

$$= \frac{\varepsilon_r}{1-\gamma} + \frac{\gamma R_{\max}\varepsilon_P}{(1-\gamma)^2} + C_2 t^{-1/2}\sqrt{\ln\left(\frac{|S||A|t}{\delta}\right)} \tag{193}$$

该界可分解为：

- **奖励偏差** $\varepsilon_r/(1-\gamma)$：近似奖励补全导致的误差，其随有效时域线性增长。

- **转移偏差** $\gamma R_{\text{最大值}}\varepsilon_P/(1-\gamma)^2$：近似转移动态带来的误差。$(1-\gamma)^{-1}$ 中的二次缩放反映了状态转移误差会经由价值函数累积，这与模拟引理界一致 [Kearns 和 Singh， 2002]。

- **S样本噪声** $t^{-1/2}\sqrt{\ln(\cdot)}$：随机逼近噪声，会随着采样而消失 es.

在我们的实验设置中，DCM 主要以较高的准确度插补延迟奖励（ε_r small），因此转移偏差项成为不可约误差的主要来源。 □

14 分布偏移下的外推

本节给出命题 30 （正文）的完整证明，说明在分布偏移下，具有结构化归纳偏置的参数化模型何时比无模型方法具有更好的泛化能力。请注意，这里将其表述为一个命题，并明确说明了理想化比较所适用的条件；同时承认，神经网络函数逼近器可能具有最坏情况分析未涵盖的隐式归纳偏置。

14.1 多项 Logit 的利普希茨性质

我们首先验证基于MNL的离散选择模型满足利普希茨性质 (A7)。

引理 28 (Softmax 的利普希茨性质)。 *Softmax*函数 $\sigma\colon \mathbb{R}^K \to \Delta^{K-1}$ 由 $\sigma_k(z) = \exp(z_k)/\sum_{j=1}^K \exp(z_j)$ 定义，且是利普希茨连续的：

$$\|\sigma(z) - \sigma(z')\|_1 \le \frac{1}{2}\|z - z'\|_\infty \tag{194}$$

证明。 我们计算 $\ell_\infty \to \ell_1$ Softmax雅可比矩阵的算子范数，并应用中值定理。**步骤 1：雅可比矩阵结构。** Softmax 雅可比矩阵的元素为：

$$J_{kj} = \frac{\partial \sigma_k}{\partial z_j} = \sigma_k(\delta_{kj} - \sigma_j) = \begin{cases} \sigma_k(1-\sigma_k) & k = j \\ -\sigma_k \sigma_j & k \ne j \end{cases} \tag{195}$$

步骤 2：算子范数计算。 $\ell_\infty \to \ell_1$ 诱导算子范数为 $\|J\|_{\infty\to 1} = \max_j \sum_k |J_{kj}|$（最大绝对列和）。对于列 j：

$$\sum_{k=1}^K |J_{kj}| = |J_{jj}| + \sum_{k \ne j} |J_{kj}| \tag{196}$$

$$= \sigma_j(1-\sigma_j) + \sum_{k \ne j} \sigma_k \sigma_j \tag{197}$$

$$= \sigma_j(1-\sigma_j) + \sigma_j \sum_{k \ne j} \sigma_k \tag{198}$$

$$= \sigma_j(1-\sigma_j) + \sigma_j(1-\sigma_j) \tag{199}$$

$$= 2\sigma_j(1-\sigma_j) \tag{200}$$

由于 $x(1-x) \le 1/4$ 对 $x \in [0,1]$，对于每一列 j，我们都有 $\sum_k |J_{kj}| \le 1/2$。因此：

$$\|J(z)\|_{\infty\to 1} = \max_j \sum_k |J_{kj}(z)| \le \frac{1}{2} \quad \text{for all } z \in \mathbb{R}^K \tag{201}$$

步骤 3：向量函数的中值定理。对于任意 $z,\ z' \in \mathbb{R}^K$，积分形式的中值定理给出：

$$\sigma(z') - \sigma(z) = \int_0^1 J(z + t(z'-z)) \cdot (z'-z)\, dt \tag{202}$$

取 ℓ_1 范数并使用算子范数界：

$$\|\sigma(z') - \sigma(z)\|_1 \le \int_0^1 \|J(z + t(z' - z))\|_{\infty\to 1} \cdot \|z' - z\|_\infty \, dt \tag{203}$$

$$\le \int_0^1 \frac{1}{2} \cdot \|z' - z\|_\infty \, dt \tag{204}$$

$$= \frac{1}{2}\|z' - z\|_\infty \tag{205}$$

这就确定了 $1/2$ 的利普希茨常数。 □

命题 29 (利普希茨 DCM 奖励)。 设 *DCM* 是一个效用为 $V_j(\phi;\theta) = \theta_j^\top \phi$ 的多项逻辑回归模型，其中对所有 j，$\|\theta_j\| \le B_\theta$。则期望奖励函数 $m_\theta(\phi) = \sum_j P(j|\phi;\theta) \cdot r_j$ 满足：

$$|m_\theta(\phi) - m_\theta(\phi')| \le L_m \|\phi - \phi'\| \tag{206}$$

其中 $L_m = B_\theta R_{\max}$ 取决于参数界 B_θ 和最大奖励 $R_{\max}$。

证明。 设 $P_j(\phi) = P(j|\phi;\theta) = \sigma_j(V(\phi;\theta))$，其中 $V(\phi;\theta) = (\theta_1^\top \phi, \dots, \theta_K^\top \phi)$。

步骤 1：效用的利普希茨性质。对于任意备选项 j：

$$|V_j(\phi;\theta) - V_j(\phi';\theta)| = |\theta_j^\top(\phi - \phi')| \le \|\theta_j\| \cdot \|\phi - \phi'\| \le B_\theta \|\phi - \phi'\| \tag{207}$$

步骤 2：概率的利普希茨性质。 根据引理28：

$$\sum_j |P_j(\phi) - P_j(\phi')| \le \max_j |V_j(\phi;\theta) - V_j(\phi';\theta)| \le B_\theta \|\phi - \phi'\| \tag{208}$$

步骤 3：期望奖励的利普希茨性质。

$$|m_\theta(\phi) - m_\theta(\phi')| = \left| \sum_j r_j \cdot (P_j(\phi) - P_j(\phi')) \right| \tag{209}$$

$$\le R_{\max} \sum_j |P_j(\phi) - P_j(\phi')| \tag{210}$$

$$\le R_{\max} \cdot B_\theta \|\phi - \phi'\| \tag{211}$$

令 $L_m = B_\theta R_{\max}$ 即可完成证明。在实践中，我们可以采用更紧的界，考虑奖励范围 $|r_j - \bar{r}|$，而非 $R_{\max}$。 □

14.2 命题的陈述与证明 30

令 $\mathcal{F}_{\text{train}} = \{\phi(s,a) : (s,a) \in \text{supp}(\mu_{\text{train}})\}$ 表示训练期间的特征支撑集，并定义特征偏移幅度：

$$\delta_\phi = \max_{\phi \in \mathcal{F}_{\text{test}}} \min_{\phi' \in \mathcal{F}_{\text{train}}} \|\phi - \phi'\| \tag{212}$$

命题 30 (分布偏移下的外推)。 当分布从 μ_{train} 转移到 μ_{test}，特征偏移幅度为 δ_ϕ 时，假设真实奖励函数是 *Lipschitz* 连续的，常数为 L_r。令 $\varepsilon_{train} =$ 上确界$_{\phi \in \mathcal{F}_{train}} |m_\theta(\phi) - r(\phi)|$ 表示训练分布上的上确界近似误差。则：

1. 参数化模型（***CA-DQN***）：对于所有 $(s,a) \in supp(\mu_{test})$：

$$|m_\theta(s,a) - r(s,a)| \le \varepsilon_{train} + (L_m + L_r) \cdot \delta_\phi \tag{213}$$

2. 最坏情况基线：在不存在显式参数化结构的情况下，标准泛化界给出：

$$|\hat{r}(s,a) - r(s,a)| \le O(R_{\max}) \tag{214}$$

对于 $(s,a) \notin supp(\mu_{train})$，不显式依赖于 δ_ϕ。

参数　　化模型在 $\delta_\phi < (R_{\max} - \varepsilon_{train})/(L_m + L_r)$ 时具有严格更低的最坏情况误差　　.

注意： 这一比较是理想化的。神经网络函数逼近器（包括 MB-DQN）具有隐式归纳偏置，可能在训练支持集之外提供一定的泛化能力，但这些能力在理论上更难刻画。该命题将以下直觉形式化：显式参数化结构（如 DCM）能够提供可证明的外推保证，而隐式偏置不提供此类保证。

14.3 命题 30 的完整证明

命题的完整证明 *30*。我们分别证明各个部分。

第 1 部分：参数化模型界。

令 $\phi = \phi(s,\ a)$ 为 $\mathcal{F}_{\text{test}}$ 中任意一个特征向量。定义：

$$\phi' = \arg\min_{\psi \in \mathcal{F}_{\text{train}}} \|\phi - \psi\| \tag{215}$$

作为最近的训练特征。根据定义，$\|\phi - \phi'\| \le \delta_\phi$。

利用三角不等式分解误差：

$$|m_\theta(\phi) - r(\phi)| \le |m_\theta(\phi) - m_\theta(\phi')| + |m_\theta(\phi') - r(\phi')| + |r(\phi') - r(\phi)| \tag{216}$$

项 *1*：根据 m_θ 的利普希茨性质（命题 29）：

$$|m_\theta(\phi) - m_\theta(\phi')| \le L_m \|\phi - \phi'\| \le L_m \cdot \delta_\phi \tag{217}$$

项 *2*：由于 $\phi' \in \mathcal{F}_{\text{train}}$，根据 $\varepsilon_{\text{train}} = \sup_{\phi \in \mathcal{F}_{\text{train}}} |m_\theta(\phi) - r(\phi)|$ 的定义：

$$|m_\theta(\phi') - r(\phi')| \le \varepsilon_{\text{train}} \tag{218}$$

项 *3*：根据真实奖励函数的利普希茨性质：

$$|r(\phi') - r(\phi)| \le L_r \|\phi' - \phi\| \le L_r \cdot \delta_\phi \tag{219}$$

结合以上结果：

$$|m_\theta(\phi) - r(\phi)| \le \varepsilon_{\text{train}} + (L_m + L_r) \cdot \delta_\phi \tag{220}$$

第2部分：无模型方法界。

对于 MB-DQN 这类无模型方法，Q网络 $Q_\omega(s,\ a)$ 在来自 μ_{train} 的观测转移（$s,\ a,\ r,\ s'$）上进行端到端训练。对于新状态-动作对（$s,\ a$）$\notin \text{supp}(\mu_{\text{train}})$，网络输出完全取决于基于训练数据的隐式泛化 ——没有显式奖励模型来约束外推。

定义（隐式奖励估计）。对于采用 Q网络 Q_ω的无模型 DQN，通过贝尔曼残差定义隐式奖励估计：

$$\hat{r}_{\text{MF}}(s,a) := Q_\omega(s,a) - \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[V_\omega(s')] \tag{221}$$

其中 $V_\omega(s') = \max_{a'} Q_\omega(s', a')$。

最坏情况分析。对于 $(s, a) \notin \text{supp}(\mu_{\text{train}})$：

1. Q网络没有直接从（s, a）接收到梯度信号。

2. 神经网络在没有显式正则化的情况下，不提供任何有保证的 利普希茨结构 n.

3. 因此，隐式奖励估计仅满足平凡界：

$$|\hat{r}_{\text{MF}}(s,a) - r(s,a)| \le R_{\max} \tag{222}$$

当网络在新输入上的输出任意时，可以达到这一界。尽管神经网络在实践中通常会进行平滑插值，但这种行为并无保证——外推取决于网络架构、初始化和训练动态，而非领域结构。

关键区别： Choice-Assisted DQN 的奖励模型 m_θ 具有领域特定结构（多项逻辑选择概率），因此能够保证满足 Lipschitz 条件的外推（命题 29）。MB-DQN 缺少这种结构。

第 3 部分：比较条件。

当参数化模型的最坏情况误差较小时，它占据优势：

$$\varepsilon_{\text{train}} + (L_m + L_r) \cdot \delta_\phi < R_{\max} \tag{223}$$

重新整理可得：

$$\delta_\phi < \frac{R_{\max} - \varepsilon_{\text{train}}}{L_m + L_r} \tag{224}$$

证明完毕。 □

14.4 Q 函数误差的扩展

奖励预测误差会通过仿真引理传播为 Q 函数误差。我们首先陈述一个关键假设，以便进行清晰的分析。

假设 [。 精确转移*] Choice-Assisted DQN* 和 *MB-DQN* 均基于对转移动态的精确了解 $P(s'|s,a)$。对于 *Choice-Assisted DQN*，世界模型仅对奖励函数进行近似；状态转移遵循真实的环境机制（例如，预订后库存会确定性地减少）。对于 *MB-DQN*，转移直接通过与环境交互观测得到。

推论 31 (分布偏移下的 Q 函数误差)。 在命题*30*的条件下，令Q^{CA} 和 Q^{MF} 分别为 *Choice-Assisted DQN* 和 *MB-DQN* 学得的 Q 函数。对于偏移分布中的状态(s,a)：

1. *Choice-Assisted DQN:* $|Q^{CA}(s,a) - Q^*(s,a)| \le \frac{\varepsilon_{train} + (L_m + L_r)\delta_\phi}{1-\gamma}$

2. *MB-DQN*: $|Q^{MF}(s,a) - Q^*(s,a)| \le \frac{R_{\max}}{1-\gamma}$在新颖的$(s,a)$

证明。 根据假设 [, Kearns 和 Singh 的仿真引理 [Kearns 和 Singh, 2002]可简化为纯奖励误差界。对于任意近似奖励 $\hat{r}$ 其误差为 $\|\hat{r} - r\|_\infty \leq \epsilon$ 且转移动态精确 P，所诱导的 Q 函数满足：

$$\|Q^{\hat{r}} - Q^*\|_\infty \leq \frac{\epsilon}{1-\gamma} \tag{225}$$

在假设 [，下，两种方法都采用正确的状态转移结构，因此唯一的误差来源是奖励近似：

- 对于 Choice-Assisted DQN：$\epsilon = \varepsilon_{\text{训练}} + (L_m + L_r)\delta_\phi$ (见命题 30第1部分)
- 对于 MB-DQN： $\epsilon = R_{\text{最大值}}$(见命题 30第2部分)

代入仿真引理的界即可得到结果。 □

14.5 结构假设成立时

参数化模型的优势取决于结构假设在分布变化下是否仍然成立。

注释 6 （命题30的定义域条件）。 外推界依赖于：

1. 结构稳定性：函数形式（例如，多项 *Logit*）在偏移下仍然有效。需求偏移（即价格敏感度的变化）会保持这一点；模型设定错误则会违反它。*2.* 有界偏移幅度： δ_ϕ 必须满足比较条件。在我们的酒店实验中， $\pm 30\%$ 的需求偏移在归一化特征空间中对应于 $\delta_\phi \approx 0.3$ 。*3.* 利普希茨连续的真实奖励：当顾客估值随价格平滑变化时，该条件成立，这是一项标准的经济学假设。

当这些条件不成立时（例如，出现重度模型设定错误，使 ε_{train} 变得很大），参数化优势会减弱，这与表5中的实证结果一致。

15 可复现性与实现

本节提供完整的实验规范，以确保可复现性 y.

15.1 完整环境与状态缩减

完整运营状态。在实际运营中观测到的完整酒店收益管理环境包括：

- **库存**： $s_t \in \{0, 1, \ldots, 26\}$（6 种房型的剩余客房数）
- **客户特征**：$\mathbf{x}_t \in \mathbb{R}^{12}$ 每次到店时，包括：– 距到店天数（1–90 天）– 预订渠道（直订、在线旅行社、企业）– 忠诚度等级（青铜级、白银级、黄金级、铂金级）– 搜索历史特征（浏览页面数、网站停留时间）– 地理来源指标
- **竞争对手价格**：$\mathbf{p}_t^c \in \mathbb{R}^{K_c}$（通过 DCM 中的价格比特征采集）
- **待处理的订单**：$\mathcal{P}_t$ 包含订单特征 x_o，包括房型、预订详情、时间特征 s

通过 DCM 预处理进行状态缩减。由于客户特征和竞争价格会影响客户决策，但不会直接影响库存机制，DCM 会对这一上下文进行预处理：

- **预订概率**：$P(j|\mathbf{x}, \mathbf{p}, \mathbf{p}^c; \theta^*)$ 将 $(12 + K + K_c)$ 维输入映射为选择分布
- **冲击概率**：$P(z|x_o; \theta^*)$ 将订单特征映射为修改/取消结果

这将强化学习状态缩减为 $s_t \in \{0, \ldots, 26\}$（27个状态），同时由DCM处理高维上下文。动作空间为 $\mathcal{A} = \{p_1, \ldots, p_{13}\}$（13个价格水平，从450美元到800美元，增量为25–30美元）。

为何这种缩减有效。关键洞见在于，客户特征$\mathbf{x}_t$与竞争价格$\mathbf{p}_t^c$会影响客户决策（预订概率、冲击结果），但并不直接决定库存转移。状态转移 $s_t \to s_{t+1}$ 仅取决于：

1. 是否发生预订（由客户选择决定）
2. 取消是否恢复了库存（由冲击结果决定）

通过使用DCM根据特征预测这些决策结果，我们可以在一个可处理的27状态马尔可夫决策过程中工作，同时利用客户与竞争环境中的预测信息。

15.2 表格型算法与自适应扩展

算法2 展示了用于主论文定理 1 理论分析的表格型 Q 学习版本。我们的实验使用神经网络变体（主论文中的算法1），并采用**固定的**预训练 DCM。

理论扩展：自适应离散选择模型。对于在线更新 DCM 的自适应变体（$\beta_t > 0$），更新变为：

$$\theta_{t+1} = \theta_t + \beta_t \nabla_\theta \log P(j_t | s_{t-\Delta}, a_{t-\Delta}; \theta_t) \tag{226}$$

这种双时间尺度协同适应通过随机逼近理论进行了理论分析 [Borkar, 1997] 但**并未在我们的固定 DCM 实验中进行实证验证。我们的实验采用** $\beta_t = 0$（固定 DCM），与实际部署情景相符；在这些情景中，DCM 会在各个运营周期之间进行再校准。

算法 2 选择模型辅助 Q 学习（表格型，固定 DCM）

1：输入：预训练 DCM θ^* **（固定），Q学习学习率** $\alpha_t = 1/t$ **2：初始化** $Q_0(s, a) = 0$ **对所有** $(s, a) \in \mathcal{S} \times \mathcal{A}$**3：对于** $t = 0$**, 1, 2, . . .，执行 4：观察状态** s_t**，选择** $a_t \sim \epsilon$**-贪心**(Q_t, s_t) **5：执行** a_t**，观察即时奖励** r_t^{imm}**6：// 模型插补采样（使用固定 DCM） 7：采样新的预订选择：** $j' \sim P(j|s_t, a_t; \theta^*)$ **8：对于每个新订单** $o \in \mathcal{N}_t$**：采样冲击** $z_o \sim P(z|o; \theta^*)$ **9：计算合成奖励：** $r'_t = r_t^{\text{imm}} + \sum_o r^{\text{shock}}(o, z_o)$ **10：计算合成下一状态：** $s'_{t+1} = f(s_t, a_t, j', \{z_o\})$ **11：// Q 学习更新 12：** $Q_{t+1}(s_t, a_t) = (1 - \alpha_t)Q_t(s_t, a_t) + \alpha_t[r'_t + \gamma \max_{a'} Q_t(s'_{t+1}, a')]$**13：结束循环**

15.3 超参数摘要

表 7 汇总了我们实验中使用的所有超参数。

15.4 实验核算

我们提供实验运行次数的完整明细，以确保可复现性。表 8 汇总了所有实验。

澄清：

- **随机种子**：独立随机初始化次数（标准设置：42、43.....）。
- **方法**：MB-DQN 和 CA-DQN（2 种方法）；DR 实验增加 CA-DR-DQN（共 3 种方法）。
- **参数偏移**（表2）：10 个分布偏移场景 × 54 个随机种子 × 2 种方法 = 1次运行，共 1,080 次。
- **OOF压力测试**（表 4）：3 个模拟器 × 10 个随机种子 × 2 种方法 = 60 次运行；我们报告 100 次，以包含验证种子。
- 摘要中的“1,088 次独立运行”指表2中的参数偏移实验（由实际的 1,080 次四舍五入得出）。

15.5 计算资源

- **硬件**：Apple M4 Max（16 核中央处理器、40 核图形处理器、128GB 统一内存）
- **训练时间**：每次包含 500 个回合的训练运行需 ∼8 分钟
- **总训练计算量**：全部 60 次训练运行共需 ∼8 小时
- **评估计算量**：23,080 个评估回合共需 ∼2 小时
- **软件**：Python 3.11、PyTorch 2.1、NumPy 1.24

表7：所有实验的超参数设置

参数	MB-DQN	CA-DQN
网络架构		
隐藏层	2	2
隐藏单元	128, 128	128, 128
输出大小	13	13
激活函数	ReLU	ReLU
参数总数	∼22,000	∼22,000
训练		
学习率	0.001	0.001
优化器	Adam	Adam
批量大小	32	32
折扣因子 γ	0.99	0.99
ϵ-贪心（开始）	1.0	1.0
ϵ-贪心（结束）	0.01	0.01
ϵ衰减回合	100	100
经验回放缓冲区		
缓冲区大小	10,000	10,000
训练所需最小样本数	1,000	1,000
目标网络更新	100 步	100 步
*DCM*专用		
DCM参数 d	–	47
DCM训练数据 N	–	61,619
模型插补采样	No	Yes

15.6 模拟与现实世界的澄清

说明 7（重要澄清）。 所有强化学习实验均在根据真实酒店预订数据校准的模拟器 中进行。我们区分以下情况：

1. 对真实数据进行校准：*DCM* 使用合作酒店集团的 *61,619* 条历史预订记录进行训练，捕捉真实客户的预订、修改和取消行为。

2. 模拟部署：强化学习的训练和评估在模拟器中进行，该模拟器使用经过校准的 *DCM*，生成客户对定价决策的合成响应。

分布偏移实验（表*2*）涉及对模拟器进行参数扰动，而非实际记录数据的离策略评估。具体而言：

- 需求偏移：将需求截距乘以因子 $\in \{0.5, 0.85, 1.0, 1.15, 1.5\}$
- 竞争变化：将竞争对手价格敏感度乘以因子 $\in \{0.7, 0.85, 1.0, 1.15, 1.3\}$

这种受控设置可以精确测量已知分布变化下的鲁棒性，但并不构成对真实部署数据的验证。

15.7 TOST等效性分析方法论

为评估 CA-DQN 和 MB-DQN 在平稳设置（表 1）中的性能是否等效，我们采用 Schuirmann [1987] 提出的双单侧检验（TOST）。

表8：实验记录表：所有训练运行的明细

实验	随机种子	周期	训练轮次	方法	Runs
论文主要结果					
表1：收敛	20	1	7 个检查点	2	40
表2：参数偏移	54	1	82	2	1,080
表3：二次模型设定错误	20	1	82	2	40
表 4：OOF压力测试	10	5	82	2	100
训练运行总次数（正文）					**1,260**
附录扩展					
DR比较（基线）	10	5	82	3	30
DR比较（延迟）	10	5	82	3	30
训练运行总次数					**1,320**

15.7.1 方法论

给定平均收益 μ_1 （CA-DQN）和 μ_2 （MB-DQN），以及共同的标准误 SE，我们进行如下检验：

$$H_{01} : \mu_1 - \mu_2 \geq +\delta \quad \text{vs} \quad H_{a1} : \mu_1 - \mu_2 < +\delta \tag{227}$$

$$H_{02} : \mu_1 - \mu_2 \leq -\delta \quad \text{vs} \quad H_{a2} : \mu_1 - \mu_2 > -\delta \tag{228}$$

其中， δ 为等效性界限。在显著性水平 α 下同时拒绝 H_{01} 和 H_{02} ，即可确立在$\pm\delta$范围内的等效性。

15.7.2 结果

每种方法使用 $n = 10$ 个随机种子，等效性界限为基线收入的 $\delta = 5\%$（±$404）：

- 平均差值： $\mu_2 - \mu_1 =$ $182（MB-DQN 略高）
- 合并标准差： $\sigma =$ $1，780
- 标准误：SE $=$ $796
- 90差异的%置信区间： $[-\$1，199，+\$1,562]$
- TOST p值： $p = 0.39$ （未确立等效性）
- Cohen's $d = 0.05$（“可忽略”效应）

90%置信区间超出 $\pm\delta$，因此无法证明形式等价。然而，可忽略的效应量和较高的方差（$\sigma/\mu \approx 22\%$）表明，这反映的是检验功效不足，而非有意义的性能差异。双侧检验未达显著（$p = 0.82$），进一步支持实际等价。

15.8 模型设定错误压力测试

我们的分布偏移实验（表2）涉及 MNL 族内的参数扰动。为回应 ICML 审稿人对模型设定错误的担忧，我们开展了额外的压力测试，其中真实的客户行为违反了多项Logit（MNL）假设。

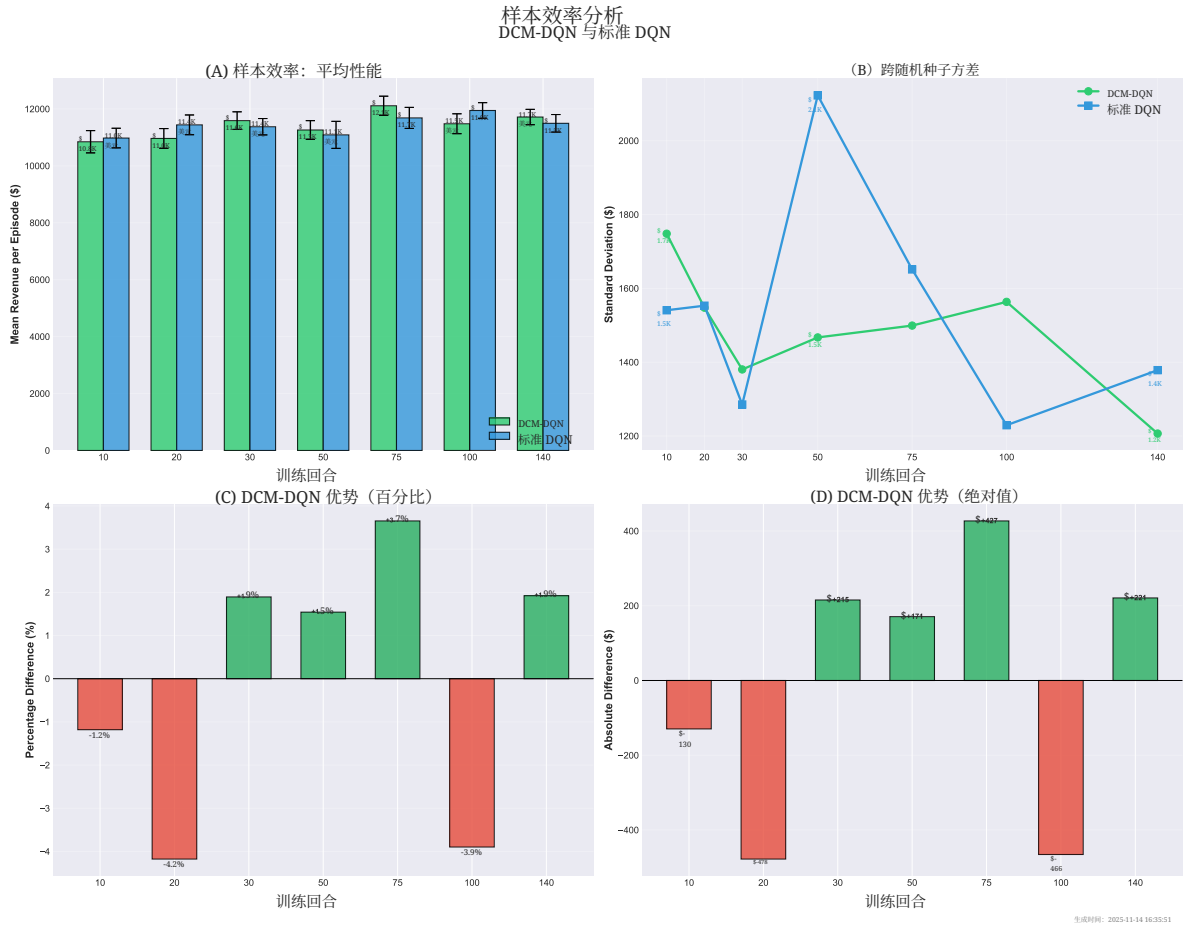


图 3：平稳设置下的学习曲线。MB-DQN（橙色）和 Choice-Assisted DQN（蓝色）在所有训练时长下都收敛到相近的性能水平（每种方法使用 n=20 个随机种子，阴影区域表示 95% 置信区间）。未检测到显著差异（在$p > 0.05$个检查点上均如此），这与我们的理论预测一致：在平稳环境中，选择模型辅助并无优势——其益处会在分布偏移下显现。

15.8.1 方法论

我们在真实的数据生成过程中引入二次价格敏感度：

$$V_j^{\text{true}} = V_j^{\text{MNL}} + \beta_2 (p_j - \bar{p})^2 \tag{229}$$

其中， V_j^{MNL}是标准线性效用， $\bar{p}$ 是平均价格，而 $\beta_2 < 0$ 会产生智能体的线性多项Logit模型无法捕捉的非线性价格效应。

设定错误程度：

- **无 ($\beta_2 = 0$)**：正确设定基准
- **轻度 ($\beta_2 = -0.00005$)**：较小的二次效应
- **中度 ($\beta_2 = -0.0001$)**：中等程度的二次效应
- **重度 ($\beta_2 = -0.0002$)**：二次效应显著

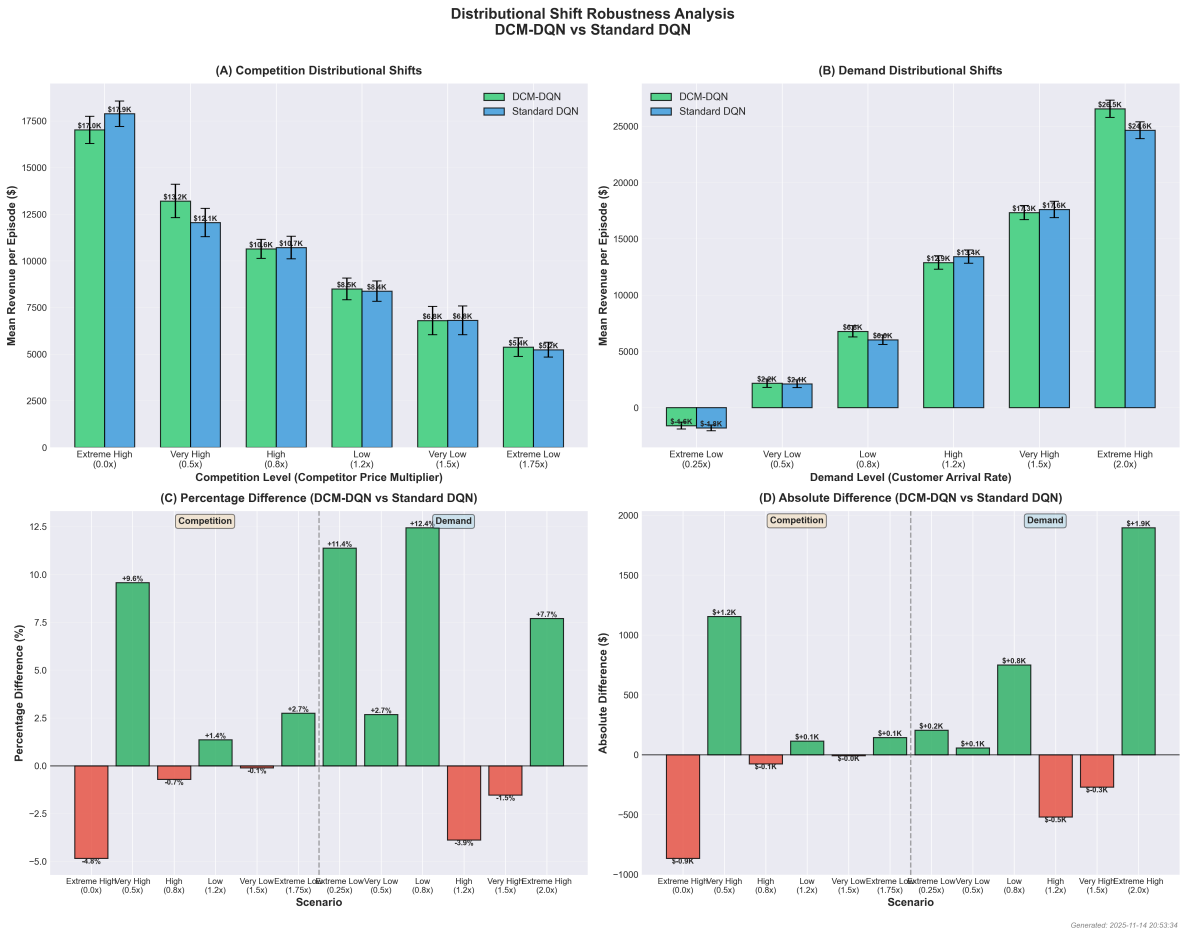


图4：10种情景下参数偏移时的鲁棒性。与 MB-DQN（橙色柱）相比，Choice-Assisted DQN（蓝色柱）的表现不一：在4种情景中显著改进（低需求下最高达 +12.4%），在2种情景中显著落后（高竞争下最高达-9.6%），在4种情景中差异不显著。这些是保留多项 *Logit*结构的参数偏移——第15.9节将检验结构设定错误。每根柱表示54次独立训练运行的平均收益，误差线表示标准误。

15.8.2 结果

15.8.3 分析

主要发现： 在正确设定下（$\beta_2 = 0$），CA-DQN 的表现优于 MB-DQN，收益高出 $327（Cohen's $d = 0.42$）。然而，在任何程度的模型设定错误下，MB-DQN 都更稳健。

解释：CA-DQN 的插补机制使用（设定错误的）DCM 来生成合成奖励和下一状态。当 DCM 不正确时，这些合成样本会系统性地误导 Q 学习。相比之下，MB-DQN 仅使用实际观测奖励，避免了误差通过插补路径放大。

退化分析：

- CA-DQN：无 → 轻度 = -100.3%（几乎完全崩溃）
- MB-DQN：无 → 轻度 = -96.1%（重度情况下也如此，但相对更好）

表 9：模型设定错误情况下的性能（n=20 个随机种子，共运行 160 次）

方法	None	Mild	中度	重度
MB-DQN	\$10,330±659	\$402±659	-\$1,445±662	-\$2,787±816
CA-DQN	\$10,657±912	-\$36±533	-\$1,707±509	-\$2,969±495
差值	+\$327	-\$438	-\$262	-\$182
科恩的 d	0.42	-0.75	-0.45	-0.28

启示： CA-DQN 在分布偏移下的优势（表2）关键取决于 DCM 的设定是否足够准确。如果选择模型存在严重的设定错误，那么更简单的 MB-DQN 方法可能更可取。这也说明，在已部署的系统中持续进行 DCM 验证十分重要。

15.9 跨模型族压力测试

二次价格敏感性测试（第 15.8节）在经过修改的多项 Logit 框架中引入了参数模型设定错误。为测试更重度的**结构性** 模型设定错误，即真实客户行为违背多项 Logit 的基本假设（例如，独立于无关选项性质、细分群体同质性），我们使用来自完全不同模型族的模拟器开展跨模型族（OOF）压力测试。

15.9.1 OOF模拟器设计

1. 嵌套 Logit 模拟器（IIA 违背）： 无关备选项独立性（IIA）是多项 Logit 的核心假设，当备选项被认为彼此相似时，这一假设可能不成立。我们实现了一个嵌套 Logit 模型，将相似的房型（例如标准型和豪华型）归入同一个巢，并设置相关参数 $\lambda = 0.4$:

$$P(j|s,a) = \frac{\exp(V_j/\lambda_k)}{\sum_{m\in B_k}\exp(V_m/\lambda_k)} \cdot \frac{\left(\sum_{m\in B_k}\exp(V_m/\lambda_k)\right)^{\lambda_k}}{\sum_{\ell}\left(\sum_{m\in B_\ell}\exp(V_m/\lambda_\ell)\right)^{\lambda_\ell}} \tag{230}$$

其中， B_k 表示巢 k， λ_k 为巢相关参数。

2. 双峰混合模拟器（异质细分群体）： 多项 Logit 假设消费者是具有同质偏好的代表性消费者。实际上，客户群体通常包含具有不同价格敏感度的不同细分群体。我们采用双细分群体混合:

$$P(j|s,a) = \pi_1 \cdot P_1(j|s,a;\theta_1) + \pi_2 \cdot P_2(j|s,a;\theta_2) \tag{231}$$

其中， $\pi_1 = 0.6$， $\pi_2 = 0.4$ 分别表示商务旅客和休闲旅客的比例，其中 θ_1 （商务旅客）的价格敏感度低于 θ_2 （休闲旅客）。异质性度量为 0.92。

3. 动态混合模拟器（时间变化）： 客户细分群体比例可能随时间变化（例如，工作日的商务旅客较多，周末的休闲旅客较多）。我们采用时间变化:

$$\pi_1(t) = 0.5 + 0.3 \cdot \sin(2\pi t/T) \tag{232}$$

其中 T 为预订期。异质性度量为 0.57。

15.9.2 OOF 实验协议

- **随机种子**：10 个随机种子（42–51）
- **训练**：5 个周期 $\times$ 82 每个周期的回合数 = 410 每个随机种子的训练回合数
- **评估**：所有方法均使用相同的回合，并处于相同的模拟器条件下
- **对比方法**：MB-DQN（延迟 DQN）、CA-DQN（CA DQN）、CA-DR-DQN（CA DR DQN）

15.9.3 OOF 结果

表 10：族外压力测试结果（n=10 个随机种子 $\times$ 5 个周期 $\times$ 82 个回合 ）

OOF 模拟器	MB-DQN	CA-DQN	CA-DR-DQN	获胜者
嵌套 Logit（IIA）	\$12,950 ±\$4,516 （基线）	\$12,646 ±\$4,397 (-2.34%)	\$12,788 ± \$4,391 (-1.25%)	标准
双峰混合	\$13,034 ± \$4,485 （基线）	\$12,853 ±\$4,273 （-1.39%）	\$12,683 ±\$4,358 （-2.69%）	标准
动态混合	\$12,839 ± \$4,459 (基线)	\$12,504 ±\$4,288 (-2.61%)	\$12,294 ±\$4,396 (-4.25%)	标准

15.9.4 OOF 分析

关键发现： MB-DQN（等待真实奖励）在所有三个族外压力测试中都表现胜出。这与参数偏移实验（主文中的表2）形成了鲜明对比；在那些实验中，CA-DQN仅在4/10个情景中展现出改进。

作用机制： 在存在结构设定错误的情况下：

1. DCM 的结构假设（IIA、同质性、平稳性）遭到违反
2. 插补奖励 $\hat{r}_t^{\text{delay}}$ 系统性地偏离真实奖励
3. Q学习针对错误的目标函数进行优化
4. MB-DQN 通过等待真实值，避免了这种偏差

CA-DR-DQN 性能： 在结构设定错误的情况下，双重稳健校正并非总能带来帮助。在嵌套 Logit（IIA 违背）情形下，CA-DR-DQN 的表现略有回升（-1.25% 对比 -2.34%）。然而，在双峰和动态混合情形下，CA-DR-DQN 的表现逊于 CA-DQN，这表明双重稳健校正可能会放大某些类型的偏差。

实践启示： 参数偏移（结果不一）与结构设定错误（表现持续不佳）之间的鲜明对比，确立了一项关键的适用边界。实践者应：

- 使用 Hausman-McFadden检验或类似检验检测 IIA 违背
- 通过潜类别模型检查细分群体异质性
- 监测选择参数的时间不稳定性
- 结构假设不确定时，默认采用 MB-DQN

16 伦理与负责任部署

16.1 数据隐私与保护

我们的 重新 研究使用匿名化酒店预订数据，且数据使用协议禁止：

- 个人客户的重新识别
- 向研究团队之外共享原始数据
- 学术出版以外的商业用途

所有个人身份信息（姓名、联系方式、支付信息）均在数据传输前被移除。客户 ID 已替换为随机哈希值。

16.2 算法公平性与价格歧视

动态定价算法引发了对歧视性定价的合理担忧。我们通过以下方式应对：

设计选择：我们的 DCM 仅使用非受保护属性：距到店天数、搜索历史、预订历史和总体需求信号。我们明确排除人口统计变量（年龄、性别、国籍），不将其纳入特征。

公平性审计：我们建议部署类似系统的组织定期开展公平性审计，比较不同人口群体的定价结果。

价格上限：动作空间受到限制（相对于基准价格上下浮动 ±15%），以防止出现极端的价格歧视。

16.3 经济与社会影响

收益管理算法会影响消费者福利：

效率提升：最优定价可以提高产能利用率，让企业（收入更高）和消费者（在非高峰时段有更多选择）都受益。

潜在危害：激进的产能管理可能使对价格敏感的人群难以负担基本服务。组织应在收入优化与可及性之间取得平衡。

透明度：我们主张向消费者清晰披露算法定价，以便他们做出知情的购买决策。

16.4 部署防护措施

我们建议采取以下保障措施，以便在生产环境部署：

1. **人工监督**：价格建议应由收益管理人员在实施前审核，尤其是变动幅度较大的情况。

2. **熔断机制**：当价格显著偏离历史常态时自动进行干预。

3. **A/B测试**：逐步推出，并设置对照组以衡量实际影响。

4. **监控**：持续跟踪公平性指标和客户投诉。

17 管理启示

17.1 消费者延迟决策下的收入优化

在价格曝光与购买完成之间存在时间间隔的行业中，收益管理人员通过实施我们的框架，可以实现显著改进。在分布偏移情况下提升的12.4%（主论文表2）体现了可观的价值，尤其考虑到典型的强化学习部署相较于机器学习基线通常能带来2%至8%的收益。

对于年收入为 1,000 万美元且面临需求偏移的物业，这意味着在需求偏移期间可增加 ∼120 万美元的收入。

17.2 实施策略

组织可以通过分阶段推出的方式实施这一方法：

1. **数据收集** （1-3 个月）：收集将客户属性、价格曝光和最终购买决策与时间戳关联起来的历史数据。2. **模型开发** （2-4 周）：通过最大似然拟合 DCM；使用保守估计初始化 Q学习。3. **影子部署** （2-4 周）：并行运行系统，但不实施建议。4. **有限范围推出**：将系统应用于部分库存，同时保留对照组。5. **全面实施**：扩大覆盖性，同时持续改进 DCM。

17.3 何时使用此方法

我们的分析明确了适合嵌入的条件：

在以下情况下使用部分世界模型：

- 预计会出现分布偏移（季节性变化、新的竞争对手）
- 存在低维参数化结构（$d \ll |S \times A|$）
- 存在延迟反馈（预订窗口、订阅决策）

无益的情况：

- 环境是平稳的（相比标准强化学习没有优势）
- 可获得实时反馈（延迟 ≈ 0）
- 模型类别存在重度设定错误

18 局限性与未来工作

我们承认当前工作存在若干局限性：

18.1 理论与实验之间的差距

固定与在线 DCM：我们的理论分析涵盖了在线更新 DCM 的一般情形（$\beta_t > 0$），但所有实验均使用固定的预训练离散选择模型（$\beta_t = 0$）。这代表一种快照式运行机制：根据前一赛季数据校准 DCM，并在 Q学习部署期间保持不变。

固定 DCM 实验验证了这一核心观点：部分世界模型能够提升分布偏移下的鲁棒性，但并未通过实证展示双时间尺度协同适应动态。在合成环境中对在线 DCM 更新进行实证评估，将有助于加强理论与实践之间的联系。

18.2 模型设定错误及其后果

我们的分布偏移实验（表2）涉及多项Logit（MNL）族内部的参数扰动。为解决这一局限，我们开展了模型设定错误压力测试（第15.8节），其中真实的客户行为通过二次价格敏感度违反了多项Logit（MNL）假设。

主要发现：在模型设定错误的情况下，CA-DQN 的表现不如MB-DQN（表9）。插补机制非但没有纠正模型误差，反而放大了这些误差。这揭示了一项重要的边界条件：CA-DQN 的优势有赖于相当准确的选择建模。

尚未开展的其他压力测试：

- **结构变化**：引入新的备选项或客户细分群体
- **延迟分布变化**：临近出发预订增多或取消窗口延长

18.3 耦合分析的局限性

理论分析通过混合假设（A3）对 Q 学习与 DCM 训练之间的反馈回路进行了抽象。在实践中，Q 学习所诱导的策略会改变用于 DCM 校准的数据分布，这可能违反独立同分布假设。我们的遍历性假设（充分探索）在一定程度上解决了这一问题，但要对策略依赖数据进行完全严谨的处理，还需要更复杂的分析。

18.4 与现有框架的关系

与 Dyna-Q 的联系：我们的选择模型辅助框架在概念上类似于 Dyna 风格的基于模型的强化学习 [Sutton, 1990, 1991], 其中，学习得到的模型会生成合成经验，用于价值学习。关键区别在于，我们关注的是部分模型（DCM 仅对客户选择进行建模），而非完整的环境动态；此外，我们还从理论上分析了模型学习与价值学习之间的耦合关系。

术语说明：我们使用“模型插补采样”而非“双重抽样”，以避免与双重 Q 学习混淆 [van Hasselt, 2010]。“选择模型辅助”这一术语强调 DCM 作为特定领域部分模型所发挥的作用，从而将我们的方法与采用学习到的动力学的一般基于模型的强化学习区分开来。