arXiv:2603.03523v1 [cs.LG] 3 Mar 2026

面向连续状态强化学习的Q-测度学习：高效实现与收敛

Shengbo Wang[1]

[1]丹尼尔·J· 爱泼斯坦工业与系统工程系，南加州大学 ia

2026年2月

摘要

我们研究无限时域折扣马尔可夫决策过程中连续状态空间下的强化学习，其中数据由马尔可夫行为策略下的单条轨迹在线生成。为避免维护无限维的函数值估计，我们提出一种新颖的*Q-Measure-Learning*，它学习一个支撑于已访问状态-动作对上的有符号经验测度，并通过核积分重构动作价值估计。该方法通过耦合随机逼近，同时估计行为链的平稳分布和 Q-测度，从而实现高效的基于权重的实现，每次迭代的内存开销为 $O(n)$，计算开销为 $O(n)$。在行为链一致遍历的条件下，我们证明诱导的 Q函数几乎必然以上确界范数收敛至核平滑贝尔曼算子的不动点。我们还给出了该极限与最优 Q^* 之间的近似误差界，并将其表示为核带宽的函数。为评估所提算法的性能，我们在双品类库存控制场景中开展了强化学习实验。

1 引言

我们研究工程领域序贯决策中的强化学习（RL），例如库存与收益管理、金融，以及物理系统和机器人系统的基于学习的控制；在这些领域中，系统状态空间最自然地是连续的。即使原始状态是离散的（例如服务系统和制造系统中的队列长度），将其嵌入实向量空间通常也更方便有效，从而为泛化带来有用的局部光滑性概念。受此启发，我们考虑具有连续状态空间 $\mathbb{X} \subset \mathbb{R}^{d_{\mathbb{X}}}$ 以及连续或有限动作空间 $\mathbb{A}$的无限时域折扣马尔可夫决策过程（MDP）的强化学习。我们关注单轨迹场景，其中数据流 $\{(R_t,\ X_t,\ A_t):\ t \geq 1\}$ 由马尔可夫行为策略 π_b在线生成。在标准正则性假设下，最优动作价值函数 Q^* 是贝尔曼最优算子的唯一不动点；该算子在上确界范数下是压缩映射。这一性质是著名的Q学习算法 [32]的理论基础。然而，在连续状态空间场景中，Q^* 是无限维对象，因此若不进行离散化或采用函数近似，表格型Q学习便无法直接应用。因此，要建立高效算法并给出收敛保证并不简单，尤其是在数据由单条轨迹生成时。

同时，我们注意到，在行为策略下进行模拟会诱导出一个马尔可夫链 $\{Z_n:\ n \geq 0\}$在 $\mathbb{Z} = \mathbb{X} \times \mathbb{A}$ 上，并且许多感兴趣的量都可以通过对所访问位置的经验测度进行积分来一致地估计。这启发了一种替代性的设计原则：与其直接在函数空间中逼近 Q^*，不如跟踪一个经验过程——经验Q测度，并通过积分变换恢复对 Q^* 的估计。

具体而言，假设存在一个有限符号测度 ν^* 在 $\mathbb{Z}$ 上，以及一个（假设）已知的 平滑核 K：$\mathbb{Z} \times \mathbb{Z} \to \mathbb{R}_{>0}$，使得

$$Q^*(z) \approx q^*(z) := \int_{\mathbb{Z}} K(z,u)\nu^*(du). \tag{1.1}$$

此外，假设 ν^* 可以由重加权经验过程 $\nu_n = \sum_{k=1}^{n} W_{n\text{来近似，}k}\delta_{Z_{n-1}}$. 那么，近似 Q^* 就归结为设计一种算法，通过迭代更新权重$\{W_{n,k}$来学习 ν^* ： $k = 1$、.....、 $n\}$。

当然， K 的选择并非易事，要实现高效的实现，还需要创新的工程设计。我们发现，很难指定一个完全已知的核 K 和测度 ν^*，使其满足上述表示。不过，在我们的设定中，有一个自然的选择 K，其归一化仅取决于行为链的平稳分布 μ_b Z。由于 μ_b 可以通过其经验测度从轨迹中得到一致估计，我们的算法使用随机逼近联合更新 μ_b 和 ν^* 的估计量，然后通过核积分恢复 Q^* 的估计（1.1）。

这种设计带来两项直接优势。首先，它使我们能够利用遍历马尔可夫链相关经验过程的收敛理论，证明经验Q测度收敛，并由此证明其导出的Q函数估计收敛。其次，它有助于实现节省内存的方案：在第 n次迭代时，只需维护历史位置 $\{Z_0, \ldots, Z_n\}$ 及其对应的权重$\{W_{n,0}, \ldots, W_{n,n}\}$。第 3节提出的高效算法结构还带来了另一项优势。我们设计了更新方式，使得在第 n次迭代时，权重向量 $\{W_{n,0}, \ldots, W_{n,n}\}$ 可在 $O(n)$ 时间内完成更新，从而使整个过程仅需 $O(n)$ 的内存，并在 $O(n^2)$ 的总计算量内完成 n次迭代。

在本文中，我们的贡献组织并呈现如下：

- 我们提出*Q*-测度学习，这是一种在线算法，用于更新符号测度 ν_n 和经验参考测度 μ_n，并通过归一化核积分重构 q_n。
- 在第3节中，我们给出一种高效的基于权重的实现，其每次迭代的计算量与内存开销为 $O(n)$，即在第 n 次迭代时。
- 在第4节中，在行为链的一致遍历性条件下，我们证明几乎必然收敛 $q_n \to q^*$于上确界范数意义下，方法为巴拿赫空间常微分方程方法，其中 q^* 是核平滑贝尔曼算子的唯一不动点。
- 在第5节中，我们量化近似误差 $\|Q^* - q^*\|$，并证明通过调节平滑核的选择，可使该误差任意小。

1.1 相关工作

时序差分方法与Q学习的经典收敛理论在有限马尔可夫决策过程中已较为完备，其中随机逼近与常微分方程论证表明，在步长递减且探索充分的条件下可得到几乎必然收敛 [32, 13, 30, 3]。近期工作为Q学习建立了有限样本保证，既包括可访问生成模型的情形，也包括单轨迹场景 [31, 7, 18]。在连续空间中，一种标准做法是限制在某一函数类内，并分析由此诱导的近似/投影贝尔曼动力学；早期分析表明，收敛可能对近似架构敏感[29], ，而非扩张/平均近似则为稳定性提供了一条原则性途径 [11, 26]。近期工作给出了巴拿赫空间随机逼近（适用于Q学习）的有限次迭代、依赖实例的速率 [21]„ 但所得方法通常需要计算无限维对象。

批量强化学习方法将贝尔曼备份与监督学习相结合，并针对拟合价值迭代及相关方案提供了具有代表性的有限样本分析 [10, 22]。在线性架构中，最小二乘方法提供了另一种批量视角：LSTD 通过正规方程求解投影贝尔曼方程

用于策略评估 [5, 4], , 而扩展到动作价值和策略迭代的方案则包括 LSQL/LSPI 风格的算法 [17, 16]。

核方法对奖励和转移进行平滑，并由此产生平滑的贝尔曼算子。KBRL 通过核平滑构建基于样本的经验 MDP，随后在诱导的代表集 [24, 23] 上求解贝尔曼方程；这种离线、基于模型的方法通常计算开销较大，且难以随转移数据量的增长而有效扩展 [14, 1]。相关的核化方法，如高斯过程强化学习和核化价值近似，也依赖全局核表示，可能带来大量矩阵计算 [9, 28]。

除了“全局学习”方法之外，互补的逐点方法还针对贝尔曼方程开发了多层蒙特卡罗方案，可在指定的状态–动作对处估计 Q^*，而无需构建全局近似 Q^* [2, 19]。这些研究在生成模型设置下提供了多项式样本复杂度保证。

相比之下，我们的方法兼具Q学习每次迭代开销低、易于在线实现的优点，以及核平滑经验MDP方法通常具备的稳定性和收敛保证：它每步只进行一次TD式更新，同时仍能几乎必然地以上确界范数收敛到一个平稳归一化、经核平滑的贝尔曼不动点。

2 记号与假设

设 $\mathbb{X} \subset \mathbb{R}^{d_{\mathbb{X}}}$ 非空且紧致，并设 $\mathbb{A}$ 为 $\mathbb{R}^{d_{\mathbb{A}}}$ 的非空紧子集或有限集（采用 0–1 度量）。虽然我们的收敛结果可推广至紧致波兰空间，但届时第 5 节中的近似分析需要在此类空间上使用具有普适性的核。为使阐述聚焦于强化学习，我们将讨论限制在欧几里得状态空间。令 $\mathcal{X}$和 $\mathcal{A}$ 分别表示 $\mathbb{X}$ 和 $\mathbb{A}$ 上的博雷尔 σ-代数。定义 $(\mathbb{Z}, \mathcal{Z}) := (\mathbb{X} \times \mathbb{A}, \mathcal{X} \times \mathcal{A})$。令 $\mathcal{P}(\mathcal{Z})$ 和 $\mathcal{M}(\mathcal{Z})$ 分别表示 $(\mathbb{Z}, \mathcal{Z})$ 上的概率测度集和有限符号测度集，并令 $C(\mathbb{Z})$ 表示连续函数 $q : \mathbb{Z} \to \mathbb{R}$ 的集合。我们分别为$\mathcal{M}(\mathcal{Z})$ 和 $C(\mathbb{Z})$ 配备全变差范数 $\|\cdot\|_{\mathrm{TV}}$ 和上确界范数 $\|\cdot\|_\infty$。

固定 $\gamma \in (0, 1)$。令 $P(\cdot|x, a)$ 为博雷尔受控转移核。令 $\pi_b(\cdot|x)$ 为马尔可夫行为策略。令 $(\Omega, \mathcal{F}, P)$ 为概率空间，其中 $\mathcal{F}$包含所有支撑受控马尔可夫链 $\{Z_n : n \geq 0\}$ 的 P零测集，使得在 P 下

$$Z_n := (X_n, A_n), \qquad A_n \sim \pi_b(\cdot|X_n), \qquad X_{n+1} \sim P(\cdot|X_n, A_n).$$

令 P_b 表示马尔可夫链 $\{Z_n$ 的转移核：$n \geq 0\}$。

我们还考虑一列独立同分布随机函数 $\{F_n(\omega, \cdot) \in C(\mathbb{Z}) : n \geq 1\}$，定义在概率空间 $(\Omega, \mathcal{F}, P)$ 上，并且与 $\{X_n, A_n : n \geq 0\}$ 相互独立。在整篇论文中，我们考虑滤过

$$\mathcal{F}_n = \sigma(X_0, A_0, \{F_k, X_k, A_k : 1 \leq k \leq n\}).$$

假设1（有界且连续的奖励）。对所有 $n \geq 0$，$\|F_n\| \leq 1$ 几乎处处成立。定义奖励$R_{n+1} = F_{n+1}(X_n, A_n, X_{n+1})$ 和 $r(x, a) = E[F_1(x, a, X_1)]$。则 $r \in C(\mathbb{Z})$。

假设 2（连续动力学）。假设存在一个概率空间 $(\mathbb{W}, \mathcal{W}, \psi)$，其中 $\mathbb{W} \subset \mathbb{R}^{d_{\mathbb{W}}}$ 配备 *Borel* σ-域，并存在一个连续函数 $f : \mathbb{Z} \times \mathbb{W} \to \mathbb{X}$，使得对所有 $B \in \mathcal{X}$ 和 $z \in \mathbb{Z}$,

$$P(B|z) = \int_{\mathbb{W}} \mathbb{1}\{f(z, w) \in B\}\, \psi(dw).$$

在这些假设下，我们考虑最大化无限时域 γ-折扣期望奖励的强化学习问题：

$$\sup_{\pi \in \Pi_{\mathrm{H}}} E^\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{k+1} \middle| X_0 = x \right],$$

其中， Π_{H} 是依赖历史的随机策略类，期望 E^{π} 由策略 π 和转移核 P诱导。

为在这种设定下实现强化学习，我们为Q函数引入贝尔曼方程。在经典随机控制文献中，贝尔曼最优算子 $\mathcal{T}$针对 $Q \in C(\mathbb{Z})$定义如下：

$$\mathcal{T}[Q](z) := r(z) + \gamma \int_{\mathbb{X}} \sup_{a' \in \mathbb{A}} Q(x', a') P(dx'|z).$$

在假设1–2 以及 $\mathbb{A}$的紧致性条件下，算子 $\mathcal{T}$定义良好，并将 $C(\mathbb{Z})$ 映射到其自身——这是最大值定理和费勒性质的标准推论；例如参见 Stachurski [26, 假设 2。1]。由于在我们的设定中，奖励有界于 1，我们定义截断贝尔曼算子$\overline{\mathcal{T}}$： $C(\mathbb{Z}) \to C(\mathbb{Z})$ 如下：

$$\overline{\mathcal{T}}[Q](x,a) := r(x,a) + \gamma \int_{\mathbb{X}} \max_{a' \in \mathbb{A}} \Pi\big(Q(x',a')\big) P(dx'|x,a), \qquad (x,a) \in \mathbb{Z}, \tag{2.1}$$

其中 $\Pi(x) := \max\left\{\min\left\{x, \ \frac{1}{1-\gamma}\right\}, \ -\frac{1}{1-\gamma}\right\}$。

由于 $\mathcal{T}$ 和 $\overline{\mathcal{T}}$ 都是作用于 $C(\mathbb{Z})$和 $\|r\| \leq 1$上的 γ-压缩映射，其唯一不动点 Q^* 满足$\|Q^*\| \leq \frac{1}{1-\gamma}$，因此 $\Pi(Q^*) \equiv Q^*$，故 $\mathcal{T}$和 $\overline{\mathcal{T}}$具有相同的唯一不动点。

此外，众所周知，在假设 1 和 2下，通过相对于 Q^* 采取贪心动作（参见 Stachurski [26]）即可得到一个最优确定性马尔可夫策略。因此，利用行为策略生成的数据进行学习Q^* 是一个核心问题。

2.1 Q-测度

令 $\kappa : \mathbb{Z} \times \mathbb{Z} \to \mathbb{R}_{>0}$ 为一个下界为 $\kappa_\wedge > 0$的连续核。我们将在第 κ 节中具体说明4.1。对于任意概率测度 $\mu \in \mathcal{P}(\mathcal{Z})$，我们定义线性算子 $\mathcal{K}_\mu : C(\mathbb{Z}) \to C(\mathbb{Z})$如下

$$\mathcal{K}_\mu[q](z) := \frac{\int_{\mathbb{Z}} \kappa(z,u) q(u) \mu(du)}{\int_{\mathbb{Z}} \kappa(z,u) \mu(du)}.$$

此外，对于 $\mu \in \mathcal{P}(\mathcal{Z})$，我们定义线性算子 $\Phi_\mu : \mathcal{M}(\mathcal{Z}) \to C(\mathbb{Z})$ 如下

$$\Phi_\mu[\nu](z) := \frac{\int_{\mathbb{Z}} \kappa(z,u) \nu(du)}{\int_{\mathbb{Z}} \kappa(z,u) \mu(du)}$$

对所有 $z \in \mathbb{Z}$均成立。我们注意到，映射 $\nu \to \Phi_\mu[\nu]$ 将作为重构映射（1.1），用于从 ν中估计 Q^*。

引理1。对于定义在 $\mathcal{Z}$ 上的任意概率测度 μ 以及任意 $f, g \in C(\mathbb{Z})$,

$$\|\mathcal{K}_\mu[f] - \mathcal{K}_\mu[g]\| \leq \|f - g\|.$$

因此，平滑截断贝尔曼算子 $\mathcal{T}_\mu := \mathcal{K}_\mu \circ \mathcal{T} : C(\mathbb{Z}) \to C(\mathbb{Z})$ 是关于$\|\cdot\|$的 γ-压缩映射。因此，它存在唯一的不动点 $q_\mu^* \in C(\mathbb{Z})$，使得 $\|q_\mu^*\| \leq \frac{1}{1-\gamma}$。

证明。固定 $z \in \mathcal{Z}$。由于 $\kappa > 0$，我们有

$$|\mathcal{K}_\mu[f](z) - \mathcal{K}_\mu[g](z)| \leq \frac{\int |f(u) - g(u)| \, \kappa(z,u) \mu(du)}{\left|\int \kappa(z,u) \mu(du)\right|} \leq \|f - g\|.$$

取上确界$z \in \mathcal{Z}$ 即可得到第一个结论。

由于 $\overline{\mathcal{T}}$在 $\|\cdot\|$中是一个 γ-压缩映射，因此对于所有 q， $q' \in C(\mathbb{Z})$，

$$\begin{aligned}\left\|\overline{\mathcal{T}}_\mu[q] - \overline{\mathcal{T}}_\mu[q']\right\| &= \left\|\mathcal{K}_\mu[\overline{\mathcal{T}}[q]] - \mathcal{K}_\mu[\overline{\mathcal{T}}[q']]\right\| \\ &\le \left\|\overline{\mathcal{T}}[q] - \overline{\mathcal{T}}[q']\right\| \\ &\le \gamma \left\|q - q'\right\|;\end{aligned}$$

即，$\overline{\mathcal{T}}_\mu$ 是 $\|\cdot\|$中的 γ压缩映射。由于（$C(\mathbb{Z})$ $\|\cdot\|$）是巴拿赫空间，巴拿赫不动点定理表明，不动点 q_μ^*存在且唯一。 □

定义 1（ Q-测度。我们将由参考测度 $\mu \in \mathcal{P}(\mathcal{Z})$诱导的 Q-测度定义为

$$\nu_\mu^*(A) := \int_A \overline{\mathcal{T}}[q_\mu^*](z)\mu(dz) \tag{2.2}$$

对所有 $A \in \mathcal{Z}$均成立。注意， $\nu_\mu^* \in \mathcal{M}(\mathcal{Z})$是有限符号测度。

我们构造算法的关键动机在于，我们可以用ν_μ^*表示 q_μ^*

$$q_\mu^*(z) = \mathcal{K}_\mu[\overline{\mathcal{T}}[q_\mu^*]](z) = \frac{\int_{\mathbb{Z}} \kappa(z,u)\nu_\mu^*(du)}{\int_{\mathbb{Z}} \kappa(z,u)\mu(du)} = \Phi_\mu[\nu_\mu^*]. \tag{2.3}$$

因此，如果我们能够学习 ν_μ^*，就可以通过应用 Φ_μ恢复 q_μ^* ；也就是说，用 κ对 ν_μ^* 进行平滑和归一化。

此外，当 $\mathcal{K}_\mu$ 接近恒等算子时， q_μ^* 的$\mathcal{T}_\mu = \mathcal{K}_\mu \circ \mathcal{T}$不动点应接近 Q^* 的 $\mathcal{T}$不动点。我们将在第5节量化这一近似误差。

3 算法与基于权重的高效实现

我们的算法设计利用了由行为策略诱导的马尔可夫链 Z 的遍历性。具体而言，我们作出以下假设。

假设 3（行为链的一致遍历性）。$(Z_n)_{n\ge 0}$ 是一致遍历的：存在 $c < \infty$、$\rho \in (0,1)$，以及唯一的不变概率测度 μ_b，其定义在 $\mathbb{Z}$ 上，并且对所有 $n \ge 0$，

$$\sup_{z\in\mathbb{Z}} \|P_b^n(z,\cdot) - \mu_b(\cdot)\|_{\mathrm{TV}} \le c\rho^n.$$

3.1 Q-测度学习

我们首先对 Q-测度学习算法作出具体且直观的描述。

在算法的整个迭代过程中，我们将维护一个概率测度 μ_n 和一个有限符号测度 ν_n。初始化为 μ_0：$= \delta_{Z_0}$ 和 $\nu_0 := 0$。对于 $n \ge 0$，将q^*-函数的当前迭代近似定义为

$$q_n(z) := \Phi_{\mu_n}[\nu_n](z) = \frac{\int \kappa(z,u)\nu_n(du)}{\int \kappa(z,u)\mu_n(du)}, \tag{3.1}$$

$\forall z \in \mathbb{Z}$。此外，我们选择步长，使其满足 Robbins–Monro 条件。

假设4。我们选择 $\{\alpha_n : n \ge 1\}$ 使得 α_n 单调不减， $\sum_{n=1}^\infty \alpha_n = \infty$，且 $\sum_{n=1}^\infty \alpha_n^2 < \infty$。此外，我们使用 $\beta_n = 1/(n+1)$。

那么，从概念上讲，我们的算法如下进行。

1. 采样（X_{n+1}， R_{n+1}， A_{n+1}）。

算法1： 有限 $\mathbb{A}$的Q-测度学习。

输入： 带宽 $\sigma > 0$，折扣因子 $\gamma \in (0,1)$，步长 $\{\alpha_n, \beta_n : n \geq 1\}$，初始状态 X_0。**初始化：** 轨迹：$= [X_0, A_0]$。设置 $\mathbf{u} \leftarrow [1]$；$\mathbf{W} \leftarrow []$。

对于 $n = 0, 1, 2, \cdots$ **执行**

生成 $[R_{n+1}, X_{n+1}, A_{n+1}]$ 并将其添加到轨迹 Traj 中。

对于 $a \in \mathbb{A}$ **执行**

$$q(a) \leftarrow \frac{\sum_{k=1}^{n} \mathbf{W}_k \kappa((X_{n+1}, a), Z_{k-1})}{\sum_{k=0}^{n} \mathbf{u}_k \kappa((X_{n+1}, a), Z_k)}$$

计算 $Y \leftarrow R_{n+1} + \gamma \max_{a \in \mathbb{A}} \Pi(q(a))$。

重新缩放先前的权重：$\mathbf{u} \leftarrow (1 - \beta_{n+1})\mathbf{u}$ 和 $\mathbf{W} \leftarrow (1 - \alpha_{n+1})\mathbf{W}$。

追加新的权重：$\mathbf{u}$ 追加 β_{n+1}，$\mathbf{W}$ 追加 $\alpha_{n+1} Y$。

2. 计算时序差分目标：

$$Y_{n+1} := R_{n+1} + \gamma \sup_{a \in \mathbb{A}} \Pi\left(q_n(X_{n+1}, a)\right) \in \mathbb{R}. \tag{3.2}$$

3. 通过随机逼近迭代更新 ν_n：

$$\nu_{n+1} := (1 - \alpha_{n+1})\nu_n + \alpha_{n+1} Y_{n+1} \delta_{Z_n}. \tag{3.3}$$

4. 更新经验平稳分布 μ_n：

$$\mu_{n+1} := (1 - \beta_{n+1})\mu_n + \beta_{n+1} \delta_{Z_{n+1}}. \tag{3.4}$$

请注意，在假设 4 和更新式（3.4）下，μ_n 将成为由马尔可夫链 $\{Z_n$ 采样得到的 μ_b 的经验测度：$n \geq 0\}$。

更新规则（3.3）与经典Q学习算法 [32]非常相似。不过，它并不直接更新函数值，而是通过为当前样本分配权重来扩充经验测度，其权重大小与单步贝尔曼目标成正比。此外，该算法经过构造，使得 μ_n，ν_n，q_n，以及 Y_n 均为 $\mathcal{F}_n$-可测的。

3.2 权重表示

本节将说明，前文给出的抽象版Q-测度学习算法可以高效实现。具体而言，我们维护访问过的支撑点 $\{Z_0, \ldots, Z_n\}$ 及权重$\{u_{n,k} : k = 0, \ldots\ldots, n\}$ 和 $\{W_{n,k} : k = 1, \ldots\ldots, n\}$，使得

$$\mu_n = \sum_{k=0}^{n} u_{n,k} \delta_{Z_k}, \qquad \nu_n = \sum_{k=1}^{n} W_{n,k} \delta_{Z_{k-1}}. \tag{3.5}$$

那么，$q_n(X_{n+1}, a)$ 的计算结果就是以下比值

$$q_n(X_{n+1}, a) = \Phi_{\mu_n}[\nu_n](X_{n+1}, a) = \frac{\sum_{k=1}^{n} W_{n,k} \kappa((X_{n+1}, a), Z_{k-1})}{\sum_{k=0}^{n} u_{n,k} \kappa((X_{n+1}, a), Z_k)}.$$

注意，对于每个 $a \in \mathbb{A}$，计算 $q_n(X_{n+1}, a)$ 只需对 $n+1$ 计算 κ 次并计算两个求和。将两个数相乘、求和以及计算 κ 均视为单次操作，则计算 $q_n(X_{n+1}, a)$ 需要 $O(n)$ 次操作。因此，若 $\mathbb{A}$ 有限，取最大值并计算 $Y_{n+1} = R_{n+1} +$

$\gamma \max_{a\in\mathbb{A}} \Pi\big(q_n(X_{n+1}, a)\big)$ 可通过 $O(|\mathbb{A}|n)$ 次操作完成。

另一方面，如果 $\mathbb{A}$ 是连续的，可以针对奖励$q_n(X_{n+1},\ a_j),\ j=1,\ \ldots,\ k$进行 k 次梯度上升。然后，在第n次迭代中，最大化过程将耗费 $O(d_{\mathbb{A}}nk)$ CPU时间。

接下来，为了更新 μ_n 和 ν_n的权重，将（3.5）代入测度递推（3.4）和（3.3），得到如下权重递推：

$$u_{n+1,k} = (1-\beta_{n+1})u_{n,k},\ \text{for } k=0,\ldots,n; \qquad u_{n+1,n+1} = \beta_{n+1}.$$
$$W_{n+1,k} = (1-\alpha_{n+1})W_{n,k},\ \text{for } k=1,\ldots,n; \qquad W_{n+1,n+1} = \alpha_{n+1}Y_{n+1}.$$

这就实现了算法1 中所述的高效实现（我们仅展示有限动作版本）。特别是，我们注意到，将算法1运行至第 m次迭代时，只需维护变量 Traj、$\mathbf{u}$ 和$\mathbf{W}$。因此，内存开销为 $O(m)$。另一方面，在第 n次迭代中，有限动作设置下的操作次数为 $O(|\mathbb{A}|n)$，连续动作设置下为 $O(d_{\mathbb{A}}nk)$，其中最大值通过 k步梯度上升进行近似。因此，总 CPU 时间分别按 $O(|\mathbb{A}|m^2)$和 $O(d_{\mathbb{A}}km^2)$增长。

4 收敛

在本节中，我们证明 Q-测度学习的收敛性。具体而言，我们证明，由 ν_n 重构得到的估计 q_n 几乎必然（a.s.）在上确界范数下收敛到光滑最优Q函数 $q^*_{\mu_b}$。我们还证明，ν_n 在由核 κ 导出的度量下几乎必然收敛到 $\nu^*_{\mu_b}$；我们将先对该度量作出说明。

4.1 归一化高斯核度量

全文中，$|\cdot|$ 表示 $\mathbb{R}^d$ 上的欧几里得范数。本文考虑 $\mathbb{A}$ 的两种情形，它们对应略有不同的核选择。

(i) 若 $\mathbb{A}\subset\mathbb{R}^{d_{\mathbb{A}}}$ 非空且紧致，则 $\mathbb{Z}\subset\mathbb{R}^{d_{\mathbb{X}}+d_{\mathbb{A}}}$。在这种情况下，我们使用 $\kappa_\sigma(z,\ u) = \exp\big(-|z-u|^2/(2\sigma^2)\big)$。

(ii) 若 $\mathbb{A}=\{a_1,\ldots,a_m\}$ 为有限集，对于 $z=(x,\ a_i)$ 和 $u=(y,\ a_j)$，

$$\kappa_\sigma(z,u) := \exp\left(-\frac{|x-y|^2 + \mathbb{1}_{i\neq j}}{2\sigma^2}\right).$$

在上下文清楚时，我们省略下标 σ。

定义 2. 对于前文定义的 $\kappa=\kappa_\sigma$，我们考虑 $\mathcal{M}(\mathcal{Z})$ 上的核均值距离，其定义为 $d(\nu,\ \nu') := \sup_{z\in\mathbb{Z}}\left|\int_{\mathbb{Z}}\kappa(z,\ u)(\nu-\nu')\,(du)\right|$. 由于 κ 未归一化，定义平稳归一化距离

$$D_{\mu_b}(\nu,\nu') := \sup_{z\in\mathbb{Z}}\left|\frac{\int_{\mathbb{Z}}\kappa(z,u)\,(\nu-\nu')\,(du)}{\int_{\mathbb{Z}}\kappa(z,u)\mu_b(du)}\right|.$$

等价地，$D_{\mu_b}(\nu,\ \nu') = \|\Phi_{\mu_b}(\nu-\nu')\|$。

我们首先证明 D_{μ_b} 如上文所定义，是一个有效的度量，可用于量化收敛。命题的证明4.1 推迟至附录A。

命题 4.1 (D_μ 是一个度量)。对于上述情形（i）或（ii）中的任一种，对每个 $\mu\in\mathcal{P}(\mathcal{Z})$，$D_\mu$ 都是一个度量。

注释 1. 下面的收敛结果并不依赖于此处指定的特定高斯型核 κ。我们选择它只是为了确保 D_{μ_b} 是一个度量，并在第 5节得到具体的逼近误差界。特别是，也可以使用在 $\mathbb{Z}\times\mathbb{Z}$ 上远离

证明 $\|q_n - q^*\| \to 0$ 几乎必然地，当 $n \to \infty$时成立。对于 κ的其他选择， D_{μ_b} 定义 2 中的量可能只是一个伪度量（即， $D_{\mu_b}(\nu, \nu') = 0$ 并不一定意味着 $\nu = \nu'$）。此外，由此得到的逼近误差界可能与定理2中的不同。

4.2 几乎必然收敛

我们首先证明马尔可夫链 $\{Z_n$ 的经验分布 $\mu_n \ n \geq 0\}$几乎必然收敛到 D_{μ_b}中的唯一平稳分布 μ_b 。

命题 4.2 (经验过程的收敛)。 在假设 *3* 和 *4*下， $D_{\mu_b}(\mu_n, \mu_b) \to 0$几乎必然收敛，当 $n \to \infty$。

证明。 由于 $\kappa > 0$ 在紧致集合 $\mathbb{Z}$上，因此只需证明 $d(\mu_n, \mu_b) \to 0$ 几乎必然成立。

结合 $\beta_n = 1/(n+1)$，递推关系 $\mu_n = (1-\beta_n)\mu_{n-1} + \beta_n \delta_{Z_n}$给出 $\mu_n = \frac{1}{n+1}\sum_{k=0}^{n} \delta_{Z_k}$—固定 $z \in \mathbb{Z}$并令 $h_z(u) := \kappa(z, u)$，该函数有界且可测。根据一致遍历马尔可夫链的强大数定律 [20, 定理 17.0.1]， $\mu_n[h_z] \to \mu_b[h_z]$ 几乎必然成立。

为了对 z取上确界，我们利用 $\mathbb{Z}$ 是紧致的，且 κ 是连续的，因此关于第一个变量一致连续。所以，对于任意 $\epsilon > 0$，存在 $\delta > 0$，使得只要$|z - z'| \leq \delta$，就有 $\sup_{u\in\mathbb{Z}} |\kappa(z, u) - \kappa(z', u)| \leq \epsilon$。令 $\{z_1, \ldots, z_m\}$ 为 $\mathbb{Z}$ 的一个有限 δ-网。那么，对于任意 $z \in \mathbb{Z}$，取满足 $|z - z_j| \leq \delta$ 的 j，即可得到

$$\left|(\mu_n - \mu_b)[h_z]\right| \leq \left|(\mu_n - \mu_b)[h_{z_j}]\right| + \mu_n[|h_z - h_{z_j}|] + \mu_b[|h_z - h_{z_j}|] \leq \left|(\mu_n - \mu_b)[h_{z_j}]\right| + 2\epsilon.$$

取 $\sup_{z\in\mathbb{Z}}$ 得到

$$d(\mu_n, \mu_b) \leq \max_{1\leq j\leq m} \left|(\mu_n - \mu_b)[h_{z_j}]\right| + 2\epsilon.$$

令 $n \to \infty$，并利用在 z_1，……， z_m 处的几乎必然收敛，可得 $\limsup_{n\to\infty} d(\mu_n, \mu_b) \leq 2\epsilon$ 几乎必然成立。由于 $\epsilon > 0$ 是任意的，因此 $d(\mu_n, \mu_b) \to 0$ 几乎必然成立，进而 $D_{\mu_b}(\mu_n, \mu_b) \to 0$ 几乎必然成立。□

接下来，我们建立本文的主要收敛定理。为简化记号，令 q^* 和 ν^*分别表示 $q^*_{\mu_b}$ 和 $\nu^*_{\mu_b}$。

定理 1 (算法 1的收敛）。在假设 *1–4*下， $\|q_n - q^*\| \to 0$ 和 $D_{\mu_b}(\nu_n, \nu^*) \to 0$几乎必然收敛，当 $n \to \infty$。

为证明 定理 1，我们先考察所关注对象的一些有界性性质 .

引理 2。 对于所有 $n \geq 0$， μ_n 是一个概率测度；$|Y_{n+1}| \leq \frac{1}{1-\gamma}$； $\|\nu_n\|_{\mathrm{TV}} \leq \frac{1}{1-\gamma}$；$\|q_n\| \leq \frac{1}{(1-\gamma)\kappa_-}$。

引理 *2*的证明。 证明如下。首先， μ_{n+1} 是概率测度的凸组合，因此也是概率测度。接下来，根据假设 1， $|R_{n+1}| \leq 1$。此外 $|\Pi(q_n)| \leq \frac{1}{1-\gamma}$，因此 $|Y_{n+1}| \leq 1 + \frac{\gamma}{1-\gamma} = \frac{1}{1-\gamma}$.于是，对于任意有界可测的 f，且 $\|f\| \leq 1$，

$$\begin{aligned} |\nu_{n+1}[f]| &\leq (1-\alpha_{n+1})\,|\nu_n[f]| + \alpha_{n+1}\,|Y_{n+1}|\,|f(Z_n)| \\ &\leq (1-\alpha_{n+1})\,\|\nu_n\|_{\mathrm{TV}} + \alpha_{n+1}\,|Y_{n+1}|\,. \end{aligned}$$

由于 $|Y_{n+1}| \leq \frac{1}{1-\gamma}$，递推关系意味着

$$\|\nu_{n+1}\|_{\mathrm{TV}} \leq (1-\alpha_{n+1})\,\|\nu_n\|_{\mathrm{TV}} + \frac{\alpha_{n+1}}{1-\gamma}.$$

由于 $\nu_0 = 0$，由归纳法可得，对于所有 n，都有 $\|\nu_n\|_{\mathrm{TV}} \leq \frac{1}{1-\gamma}$。

最后，对于任意 $z \in \mathbb{Z}$,

$$\left| \int_{\mathbb{Z}} \kappa(z,u)\nu_n(du) \right| \le \|\nu_n\|_{\mathrm{TV}} \, \|\kappa(z,\cdot)\| \le \|\nu_n\|_{\mathrm{TV}} \cdot$$

同样 $\int_{\mathbb{Z}} \kappa(z,\ u)\mu_n(du) \ge \kappa_\wedge$。因此 $|q_n(z)| \le \frac{\|\nu_n\|_{\mathrm{TV}}}{\kappa_\wedge} \le \frac{1}{1-\gamma()\kappa_\wedge}$. □

定理 *1* 的证明。由于 $d(\mu_n,\ \mu_b) \to 0$，我们可以考虑一个辅助序列 $\bar{q}_n$，即 μ_b-归一化q-函数由 ν_n 导出；即

$$\bar{q}_n(z) := \Phi_{\mu_b}[\nu_n](z) = \frac{\int \kappa(z,u)\nu_n(du)}{\int \kappa(z,u)\mu_b(du)}. \tag{4.1}$$

引理3。 $\|q_n - \bar{q}_n\| \le \frac{D_{\mu_b}(\mu_n,\ \mu_b)}{(1-\gamma)\kappa_\wedge} \to 0$ 几乎必然

引理 *3* 的证明。对于任意 $z \in \mathbb{Z}$,

$$q_n(z) - \bar{q}_n(z) = \int \kappa(z,u)\nu_n(du) \left(\frac{1}{\int_{\mathbb{Z}} \kappa(z,u)\mu_n(du)} - \frac{1}{\int_{\mathbb{Z}} \kappa(z,u)\mu_b(du)} \right).$$

为简便起见，令

$$c_\mu(z) := \int_{\mathbb{Z}} \kappa(z,u)\mu(du). \tag{4.2}$$

根据引理 2，$\left|\int \kappa(z,\ u)\nu_n(du)\right| \le \|\nu_n\|_{\mathrm{TV}} \le \frac{1}{1-\gamma}$。同样 $c_{\mu_n}(z) \ge \kappa_\wedge$ 且 $c_{\mu_b}(z) \ge \kappa_\wedge$，因此

$$\left| \frac{1}{c_{\mu_n}(z)} - \frac{1}{c_{\mu_b}(z)} \right| = \frac{|c_{\mu_n}(z) - c_{\mu_b}(z)|}{c_{\mu_n}(z)c_{\mu_b}(z)} \le \frac{D_{\mu_b}(\mu_n,\mu_b)}{\kappa_\wedge}.$$

因此

$$\|q_n - \bar{q}_n\| \le \frac{D_{\mu_b}(\mu_n,\mu_b)}{(1-\gamma)\kappa_\wedge}.$$

命题 4.2 给出 $D_{\mu_b}(\mu_n,\ \mu_b) \to 0$ 几乎必然成立，因此所述结果成立。 □

因此，为了证明 $\|q_n - q^*\| \to 0$ 几乎必然成立，只需证明 $\|\bar{q}_n - q^*\| \to 0$ 几乎必然收敛。为此，我们分析算法递推。根据递推式（3.3）以及 $\bar{q}_n$ 在（4.1）中的定义，

$$\begin{aligned} \bar{q}_{n+1} &= \Phi_{\mu_b}[\nu_{n+1}] \\ &= \Phi_{\mu_b}[(1-\alpha_{n+1})\nu_n] + \alpha_{n+1}Y_{n+1}\Phi_{\mu_b}[\delta_{Z_n}] \\ &= \bar{q}_n - \alpha_{n+1}\bar{q}_n + \alpha_{n+1}Y_{n+1}\Phi_{\mu_b}[\delta_{Z_n}] \end{aligned} \tag{4.3}$$

为了将 Y_{n+1} 与贝尔曼方程联系起来，我们考虑鞅差

$$D_{n+1} := (Y_{n+1} - E[Y_{n+1}|\mathcal{F}_n])\,\Phi_{\mu_b}[\delta_{Z_n}] \tag{4.4}$$

其中，根据马尔可夫性质

$$E[Y_{n+1}|\mathcal{F}_n] = r(X_n,A_n) + \gamma \int \sup_{a\in\mathbb{A}} \Pi(q_n(x',a))P(dx'|X_n,A_n) = \overline{\mathcal{T}}[q_n](Z_n).$$

因此，令

$$\begin{aligned} G_n(z) &:= \overline{\mathcal{T}}[\bar{q}_n](Z_n)\Phi_{\mu_b}[\delta_{Z_n}](z) - \overline{\mathcal{T}}_{\mu_b}[\bar{q}_n](z) \\ B_n(z) &:= \overline{\mathcal{T}}[q_n](Z_n)\Phi_{\mu_b}[\delta_{Z_n}](z) - \overline{\mathcal{T}}[\bar{q}_n](Z_n)\Phi_{\mu_b}[\delta_{Z_n}](z) \end{aligned} \tag{4.5}$$

对所有 $z \in \mathbb{Z}$，我们可以将 (4.3) 重写为

$$\begin{aligned}\bar{q}_{n+1} &= \bar{q}_n - \alpha_{n+1}\bar{q}_n + \alpha_{n+1}Y_{n+1}\Phi_{\mu_b}[\delta_{Z_n}] \\ &= \bar{q}_n + \alpha_{n+1}\left(\overline{\mathcal{T}}_{\mu_b}[\bar{q}_n] - \bar{q}_n\right) - \alpha_{n+1}\overline{\mathcal{T}}_{\mu_b}[\bar{q}_n] + \alpha_{n+1}D_{n+1} + \alpha_{n+1}\overline{\mathcal{T}}[q_n](Z_n)\Phi_{\mu_b}[\delta_{Z_n}] \\ &= \bar{q}_n + \alpha_{n+1}\left(\overline{\mathcal{T}}_{\mu_b}[\bar{q}_n] - \bar{q}_n\right) + \alpha_{n+1}D_{n+1} + \alpha_{n+1}\left(\overline{\mathcal{T}}[q_n](Z_n)\Phi_{\mu_b}[\delta_{Z_n}] - \overline{\mathcal{T}}_{\mu_b}[\bar{q}_n]\right) \\ &= \bar{q}_n + \alpha_{n+1}\left(\overline{\mathcal{T}}_{\mu_b}[\bar{q}_n] - \bar{q}_n\right) + \alpha_{n+1}(D_{n+1} + G_n + B_n).\end{aligned} \tag{4.6}$$

我们注意到，$\{D_n : n \geq 1\}$、$\{G_n : n \geq 0\}$ 和 $\{B_n : n \geq 0\}$ 分别解释为鞅差噪声、马尔可夫噪声和偏差序列。我们的证明依赖于分别分析这些项，并得到以下结果。

命题 4.3 (常数时间窗口上的一致误差消失)。设

$$\mathcal{N}(n,T) := \left\{ m \geq n : \sum_{k=n}^{m-1} \alpha_{k+1} \leq T \right\}.$$

那么，对于任意固定的 $T \geq 0$,

$$\max_{m \in \mathcal{N}(n,T)} \left\| \sum_{k=n}^{m-1} \alpha_{k+1}(D_{k+1} + G_k + B_k) \right\| \to 0 \quad a.s. \tag{4.7}$$

为简 化定理 1 的证明，我们将命题 4.3 的证明推迟到附录 B .

有了这个误差界，我们接下来使用经典的常微分方程方法（参见 Kushner 和 Yin [15]）证明收敛。需要指出的是，由于我们处理的是巴拿赫空间值对象 $\bar{q}_n \in C(\mathbb{Z})$，因此还需要采用 Kushner 和 Yin [15] 中向量值情形之外的其他技术。

为简化记号，我们定义 $\mathcal{H}[q] := \mathcal{T}_{\mu_b}[q] - q$。注意，由于 $\mathcal{T}_{\mu_b}$ 是 γ-压缩映射，对所有q, $q' \in C(\mathbb{Z})$,

$$\|\mathcal{H}[q] - \mathcal{H}[q']\| \leq \left\|\overline{\mathcal{T}}_{\mu_b}[q] - \overline{\mathcal{T}}_{\mu_b}[q']\right\| + \|q - q'\| \leq (1+\gamma)\|q - q'\|; \tag{4.8}$$

即 $\mathcal{H}$ 是 $(1+\gamma)$-利普希茨的。

采用这一记号，$\bar{q}_n$ 在（4.6）中的递推意味着

$$\bar{q}_m = \bar{q}_n + \sum_{k=n}^{m-1} \alpha_{k+1}\mathcal{H}[q_k] + \sum_{k=n}^{m-1} \alpha_{k+1}(D_{k+1} + G_k + B_k).$$

我们首先陈述所关注的（取值于巴拿赫空间的）常微分方程，以及引理 4 中相关的解的存在唯一性定理。然后，我们在引理 5 中建立该常微分方程不动点的渐近稳定性。

引理 4 （Br´ezis [6], 定理 7.3）。设 $(E, \|\cdot\|)$ 为一个巴拿赫空间，并设 $\mathcal{H} : E \to E$ 是全局利普希茨的；即存在 $\ell < \infty$，使得对所有 q、$q' \in E$、$\|\mathcal{H}[q] - \mathcal{H}[q']\| \leq \ell\|q - q'\|$ 成立。则对每一个 $q_0 \in E$，存在唯一函数 $q : [0,\infty) \to E$，使得 q 连续可微，且对所有 $t \geq 0$

$$q(t) = q_0 + \int_0^t \mathcal{H}(q(s))ds. \tag{4.9}$$

应用引理 4 于 $(E, \|\cdot\|) = (C(\mathbb{Z}), \|\cdot\|)$, 我们将 $q(t; q')$ 定义为以下方程的唯一解

$$q(t;q') := q' + \int_0^t \mathcal{H}[q(s;q')]ds. \tag{4.10}$$

引理 5 (全局指数稳定性 q^*)。 回顾一下， $q^* = q^*_{\mu_b}$ 是 $\mathcal{T}_{\mu_b}$ 在 $C(\mathbb{Z})$中的唯一不动点。对于所有$\overline{\ }t \geq 0$ 和 $q' \in C(\mathbb{Z})$

$$\|q(t, q') - q^*\| \leq e^{-(1-\gamma)t} \|q' - q^*\|.$$

证明 引理 5的证明。 注意，当 $q' = q^*$ 常函数 $t \to q^*$ 满足（4.10）时。因此

$$\begin{aligned}\epsilon(t) &:= q(t; q') - q^* \\ &= (q' - q^*) + \int_0^t (\mathcal{H}[q(s; q')] - \mathcal{H}[q^*])\, ds. \\ &= \epsilon(0) + \int_0^t \left(\overline{\mathcal{T}}_{\mu_b}[q(s; q')] - \overline{\mathcal{T}}_{\mu_b}[q^*] + \epsilon(s)\right) ds.\end{aligned}$$

因此，利用分部积分可得

$$e^t \epsilon(t) = \epsilon(0) + \int_0^t e^s \left(\overline{\mathcal{T}}_{\mu_b}[q(s; q')] - \overline{\mathcal{T}}_{\mu_b}[q^*]\right) ds.$$

取上确界范数，得到

$$e^t \|\epsilon(t)\| \leq \|\epsilon(0)\| + \int_0^t e^s \gamma \|q(s; q') - q^*\|\, ds = \|\epsilon(0)\| + \int_0^t e^s \gamma \|\epsilon(s)\|\, ds$$

其中用到了 $\mathcal{T}_{\mu_b}$ 的压缩性质。 —

因此，令 $x(t) = e^t \|\epsilon(t)\|$，可得

$$x(t) \leq x(0) + \gamma \int_0^t x(s) ds.$$

由 Gr¨onwall 不等式， $x(t) \leq \|x(0)\| e^{\gamma t} = \|\epsilon(0)\| e^{\gamma t}$，因此

$$\|\epsilon(t)\| \leq e^{-t} x(t) \leq e^{-(1-\gamma)t} \|\epsilon(0)\|.$$

这意味着引理 5。 □

将 n 与 m 之间的连续时间定义为

$$\tau_{n,m} := \sum_{k=n}^{m-1} \alpha_{k+1}.$$

因此，对于所有 $k \geq n$，我们有

$$\max \mathcal{N}(n, \tau_{n,k}) = k \quad \text{and hence} \quad \tau_{n,\max \mathcal{N}(n,T)} \leq T. \tag{4.11}$$

我们考虑离散更新与时间为 τ_n、m 时的常微分方程流之间的误差。具体而言，我们

将两者相减，并将误差改写为

$$\begin{aligned}\bar{q}_m - q(\tau_{n,m};\bar{q}_n) = &\underbrace{\sum_{k=n}^{m-1}\alpha_{k+1}\left(\mathcal{H}[\bar{q}_k]-\mathcal{H}[q(\tau_{n,k};\bar{q}_n)]\right)}_{:=\psi_{n,m}}\\ &+\underbrace{\sum_{k=n}^{m-1}\left(\alpha_{k+1}\mathcal{H}[q(\tau_{n,k};\bar{q}_n)]-\int_{\tau_{n,k}}^{\tau_{n,k+1}}\mathcal{H}[q(s;\bar{q}_n)]ds\right)}_{:=\phi_{n,m}}\\ &+\underbrace{\sum_{k=n}^{m-1}\alpha_{k+1}\left(D_{k+1}+G_k+B_k\right)}_{:=\varepsilon_{n,m}}\end{aligned} \tag{4.12}$$

我们分别分析这些项。首先，对于$\psi_{n,\ m}$，有

$$\begin{aligned}\|\psi_{n,m}\| &\le \sum_{k=n}^{m-1}\alpha_{k+1}\left(\left\|\overline{\mathcal{T}}_{\mu_b}[\bar{q}_k]-\overline{\mathcal{T}}_{\mu_b}[q(\tau_{n,k};\bar{q}_n)]\right\|+\|\bar{q}_k-q(\tau_{n,k};\bar{q}_n)\|\right)\\ &\le (1+\gamma)\sum_{k=n}^{m-1}\alpha_{k+1}\|\bar{q}_k-q(\tau_{n,k};\bar{q}_n)\|\end{aligned}$$

这里用到了$\mathcal{T}_{\mu_b}$的压缩性质。

对于$\phi_{n,\ m}$，注意到$\tau_{n,\ k+1}-\tau_{n,\ k}=\alpha_{k+1}$。因此，

$$\begin{aligned}\|\phi_{n,m}\| &\le \sum_{k=n}^{m-1}\int_{\tau_{n,k}}^{\tau_{n,k+1}}\|\mathcal{H}[q(\tau_{n,k};\bar{q}_n)]-\mathcal{H}[q(s;\bar{q}_n)]\|\,ds\\ &\le (1+\gamma)\sum_{k=n}^{m-1}\int_{\tau_{n,k}}^{\tau_{n,k+1}}\|q(\tau_{n,k};\bar{q}_n)-q(s;\bar{q}_n)\|\,ds\\ &\overset{(i)}{=} (1+\gamma)\sum_{k=n}^{m-1}\int_{\tau_{n,k}}^{\tau_{n,k+1}}\left\|\int_{\tau_{n,k}}^{s}\mathcal{H}[q(t;\bar{q}_n)]dt\right\|ds\\ &\le \frac{(1+\gamma)}{2}\sum_{k=n}^{m-1}\sup_{\tau_{n,k}\le s\le\tau_{n,k+1}}\|\mathcal{H}[q(s;\bar{q}_n)]\|\,\alpha_{k+1}^2\\ &\le \sup_{0\le s\le\tau_{n,m}}\|\mathcal{H}[q(s;\bar{q}_n)]\|\sum_{k=n}^{m-1}\alpha_{k+1}^2\end{aligned} \tag{4.13}$$

其中（i）由（4.10）得出，最后一个等号则用到了 $\gamma<1$。

根据引理 2 和 $\kappa\le 1$，可得

$$\bar{q}_n(z)=\frac{\int_{\mathbb{Z}}\kappa(z,u)\nu_n(du)}{\int_{\mathbb{Z}}\kappa(z,u)\mu_b(du)}\le\frac{\|\nu_n\|_{\mathrm{TV}}}{\kappa_\wedge}\le\frac{1}{(1-\gamma)\kappa_\wedge}.$$

因此，根据引理 5，我们有

$$\|q(s;\bar{q}_n)-q^*\|\le e^{-(1-\gamma)t}\|\bar{q}_n-q^*\|\le\frac{2}{(1-\gamma)\kappa_\wedge}.$$

从而，由 (4.13)，我们得出

$$\|\phi_{n,m}\| \leq \frac{2}{(1-\gamma)\kappa_\wedge} \sum_{k=n}^{m-1} \alpha_{k+1}^2. \tag{4.14}$$

由于 $\sum_{k=0}^{\infty} \alpha_{k+1}^2 < \infty$，我们有：当 $n \to \infty$,

$$\max_{m \in \mathcal{N}(n,T)} \|\phi_{n,m}\| \to 0 \tag{4.15}$$

处处成立。

因此，回到（4.12），我们有

$$\|\bar{q}_m - q(\tau_{n,m}; \bar{q}_n)\| \leq (1+\gamma) \sum_{k=n}^{m-1} \alpha_{k+1} \|\bar{q}_k - q(\tau_{n,k}; \bar{q}_n)\| + \|\phi_{n,m}\| + \|\varepsilon_{n,m}\|. \tag{4.16}$$

现在，我们将该界加强为在一个常数区间内对所有时刻都成立的统一版本，如 $n \to \infty$。具体而言，我们考察如下定义的统一误差:

$$\Delta_n(T) := \max_{m \in \mathcal{N}(n,T)} \|\bar{q}_m - q(\tau_{n,m}; \bar{q}_n)\|.$$

注意，对于所有 $k \geq n$ 和 $t \in [\tau_{n,k}, \tau_{n,k+1})$

$$\Delta_n(t) = \max_{m \in \mathcal{N}(n,t)} \|\bar{q}_m - q(\tau_{n,m}; \bar{q}_n)\| \geq \|\bar{q}_k - q(\tau_{n,k}; \bar{q}_n)\|.$$

So,

$$\int_{\tau_{n,k}}^{\tau_{n,k+1}} \Delta_n(t)dt \geq \alpha_{k+1} \|\bar{q}_k - q(\tau_{n,k}; \bar{q}_n)\|.$$

结合这些性质与（4.16），我们有

$$\begin{aligned}\Delta_n(T) &\leq (1+\gamma) \sum_{k=n}^{\max \mathcal{N}(n,T)-1} \int_{\tau_{n,k}}^{\tau_{n,k+1}} \Delta_n(t)dt + \max_{m \in \mathcal{N}(n,m)} (\|\phi_{n,m}\| + \|\varepsilon_{n,m}\|) \\ &\leq (1+\gamma) \int_0^T \Delta_n(t)dt + \max_{m \in \mathcal{N}(n,T)} (\|\phi_{n,m}\| + \|\varepsilon_{n,m}\|)\end{aligned}$$

其中最后一个不等式由 $\tau_{n,\,n} = 0$ 以及（4.11）得出。因此，应用 Gr¨onwall 不等式可得

$$\Delta_n(T) \leq e^{(1+\gamma)T} \left(\max_{m \in \mathcal{N}(n,T)} \|\phi_{n,m}\| + \max_{m \in \mathcal{N}(n,T)} \|\varepsilon_{n,m}\| \right). \tag{4.17}$$

因此，回顾 ε_n 的定义，m 并应用命题4.3 和（4.15），我们得出，对于固定的$T > 0$，$\Delta_n(T) \to 0$ 几乎必然地，当 $n \to \infty$。

现在，我们利用 $\Delta_n(T)$ 的几乎必然收敛以及引理 q^* 5 中 的稳定性，得出 $\bar{q}_n \to q^*$几乎必然成立。具体而言，固定 $T > 0$，递归定义 $m_0 := 0$ 并且

$$m_{k+1} := \max \mathcal{N}(m_k, T), \quad k \geq 0.$$

由（4.11），$\tau_{m_k,\,m_{k+1}} \leq T$. 此外，由于 $\alpha_k \to 0$，不难看出 $\tau_{m_k,\,m_{k+1}} \to T$ 因为 $k \to \infty$。特别地，存在足够大的 k_T，使得对于所有 $k \geq k_T$，$\tau_{m_k,\,m_{k+1}} \geq T/2$。

我们首先控制迭代 $\{m_k$ 时的误差： $k \geq 0\}$。令 $E_{k+1}：= \left\|\bar{q}_{m_{k+1}} - q^*\right\|$，则

$$\begin{aligned} E_{k+1} &\leq \left\|\bar{q}_{m_{k+1}} - q(\tau_{m_k,m_{k+1}}; \bar{q}_{m_k})\right\| + \left\|q(\tau_{m_k,m_{k+1}}; \bar{q}_{m_k}) - q^*\right\| \\ &\overset{(i)}{\leq} \Delta_{m_k}(T) + \exp(-(1-\gamma)\tau_{m_k,m_{k+1}}) \left\|\bar{q}_{m_k} - q^*\right\| \\ &= \Delta_{m_k}(T) + \exp(-(1-\gamma)\tau_{m_k,m_{k+1}}) E_k \end{aligned}$$

其中（i）使用了 m_{k+1} 的定义，从而 $m_{k+1} \in \mathcal{N}$（m_k， T），以及引理 5。因此，对于 $k \geq k_T$

$$E_{k+1} \leq e^{-\frac{(1-\gamma)T}{2}} E_k + \Delta_{m_k}(T) \tag{4.18}$$

由于 $\Delta_n(T) \to 0$ 几乎必然成立，对于几乎所有 $\omega \in \Omega$ 以及任意 $\epsilon > 0$，存在 $K_T(\epsilon, \omega) \geq k_T$，使得对所有 $k \geq K_T(\epsilon, \omega)$，

$$\Delta_{m_k}(T, \omega) \leq (1 - e^{-(1-\gamma)T/2})\epsilon.$$

因此，对于这样的 ω 以及所有 $k \geq K_T(\epsilon, \omega)$，

$$E_{k+1}(\omega) \leq e^{-(1-\gamma)T/2} E_k(\omega) + (1 - e^{-(1-\gamma)T/2})\epsilon.$$

不难看出

$$\begin{aligned} E_n &\leq e^{-(1-\gamma)T(n-k)/2} E_k + (1 - e^{-(1-\gamma)T/2}) \sum_{j=0}^{n-k-1} e^{-(1-\gamma)Tj/2}\epsilon \\ &\leq e^{-(1-\gamma)T(n-k)/2} E_k + \epsilon \end{aligned}$$

对所有 $n \geq k \geq K_T(\epsilon, \omega)$ 均成立。由于 ϵ 是任意的，我们得出 $E_n \to 0$ 几乎必然成立。

最后，对于任意 $m \in \mathcal{N}(m_k, T)$，有

$$\begin{aligned} \|\bar{q}_m - q^*\| &\leq \|\bar{q}_m - q(\tau_{m_k,m}, \bar{q}_{m_k})\| + \|q(\tau_{m_k,m}, \bar{q}_{m_k}) - q^*\| \\ &\overset{(i)}{\leq} \Delta_{m_k}(\tau_{m_k,m_{k+1}}) + \|\bar{q}_{m_k} - q^*\| \\ &\overset{(ii)}{\leq} \Delta_{m_k}(T) + E_k \end{aligned}$$

其中（i）对 $m = m_k$，……，m_{k+1} 以及引理 5，而（ii）利用了 $\Delta_{m_k}(\cdot)$ 的单调性以及（4.11）。由于 $\sum_k \alpha_{k+1} = \infty$， $m_k \to \infty$ 当 $k \to \infty$时，我们有

$$\limsup_{k\to\infty} \max_{m \in \mathcal{N}(m_k,T)} \|\bar{q}_m - q^*\| = \limsup_{k\to\infty} [\Delta_{m_k}(T) + E_k] = 0 \quad \text{a.s.}$$

因此，几乎必然成立。

$$0 = \lim_{k\to\infty} \sup_{n \geq k} \max_{m \in \mathcal{N}(m_n,T)} \|\bar{q}_m - q^*\| = \lim_{k\to\infty} \sup_{n \geq m_k} \|\bar{q}_n - q^*\|,$$

这意味着 $\|\bar{q}_n - q^*\| \to 0$，进而由引理 $\|q_n - q^*\| \to 0$ 3 可知，几乎必然成立。

最后，注意到 $\|\bar{q}_n - q^*\| = \|\Phi_{\mu_b}[\nu_n - \nu^*]\| = D_{\mu_b}(\nu_n, \nu^*)$。因此， $D_{\mu_b}(\nu_n, \nu^*) \to 0$ 几乎必然成立 . □

5 用 q^* 逼近 Q^*

我们已证明，随着 $n \to \infty$， q_n 会持续近似 q^* 。现在，我们通过界定最优 Q函数 Q^* （ $\mathcal{T}$的不动点）与平稳平滑不动点 $q^* = q^*_{\mu_b}$ （ $\mathcal{T}_{\mu_b}$的不动点）之间的确定性近似误差，量化平滑带来的偏差。

为简化记号，我们用 d 表示以下度量

$$\mathsf{d}(z,u) := \begin{cases} |z-u|, & \text{case (i): if } \mathbb{A} \subset \mathbb{R}^{d_{\mathbb{A}}} \text{ is compact;} \\ \sqrt{|x-y|^2 + \mathbb{1}_{\{i \neq j\}}}, & \text{case (ii): if } \mathbb{A} = \{a_1, \ldots, a_m\}. \end{cases}$$

在这两种情况下，我们考虑一个参考测度

$$\lambda := \begin{cases} \text{Leb on } \mathbb{Z} \subset \mathbb{R}^{d_{\mathbb{X}}+d_{\mathbb{A}}}, & \text{case (i)}, \\ \text{Leb} \times (\text{counting measure on } \mathbb{A}) \text{ on } \mathbb{Z} = \mathbb{X} \times \mathbb{A}, & \text{case (ii)}, \end{cases}$$

其中 Leb 表示勒贝格测度。此外，我们将 $\mathbb{R}^d$ 中欧几里得球的体积写作

$$V_d(r) := \text{Leb}\big(\{v \in \mathbb{R}^d : |v| \le r\}\big) = \frac{\pi^{d/2} r^d}{\Gamma(d/2+1)}.$$

我们证明以下近似误差界成立。

定理 2（平滑 Q 函数的近似误差）。假设 Q^* 在 d 中是 α-*H¨older* 的，对于某个 $\alpha \in (0,1],$，即存在 $L_Q < \infty$ 使得 $|Q^*(z) - Q^*(u)| \le L_Q \mathsf{d}(z,u)^{\alpha}$ 对所有 z、$u \in \mathbb{Z}$ 成立。于是，对于任意 $\sigma > 0,$

$$\|Q^* - q^*\| \le \frac{L_Q}{1-\gamma}\, \xi_{\mu_b}(\sigma), \tag{5.1}$$

其中

$$\xi_{\mu_b}(\sigma) := \sup_{z \in \mathbb{Z}} \frac{\int_{\mathbb{Z}} \mathsf{d}(z,u)^{\alpha}\, \kappa_\sigma(z,u)\, \mu_b(du)}{\int_{\mathbb{Z}} \kappa_\sigma(z,u)\, \mu_b(du)}. \tag{5.2}$$

此外，假设 $\mu_b \ll \lambda$ 具有密度 $p_b = \frac{d\mu_b}{d\lambda}$，使得 $0 < p_{\underline{b}} \le$ 本质下确界$_\lambda\, p_b \le p_b \le$本质上确界$_\lambda\, p_b < \infty$。同时假设以下一致局部体积条件：存在常数 $0 < v_\wedge \le 1$ 和 $\sigma_\vee > 0$，使得对于所有 $z \in \mathbb{Z}$

$$\begin{aligned} &\textit{case (i):} \quad \text{Leb}\big(\{u \in \mathbb{Z} : |z-u| \le r\}\big) \ge v_\wedge V_{d_{\mathbb{X}}+d_{\mathbb{A}}}(r), \quad \forall r \in (0, \sigma_\vee \sqrt{d_{\mathbb{X}}+d_{\mathbb{A}}}]; \\ &\textit{case (ii):} \quad \text{Leb}\big(\{y \in \mathbb{X} : |x-y| \le r\}\big) \ge v_\wedge V_{d_{\mathbb{X}}}(r), \qquad \forall r \in (0, \sigma_\vee \sqrt{d_{\mathbb{X}}}]. \end{aligned} \tag{5.3}$$

那么，对所有 $\sigma \in (0,\ \sigma_\vee],$

$$\begin{aligned} &\textit{case (i):} \quad \xi_{\mu_b}(\sigma) \le \frac{2}{v_\wedge} \frac{\overline{p}_b}{\underline{p}_b} (d_{\mathbb{X}} + d_{\mathbb{A}})^{\alpha/2}\, \sigma^{\alpha}, \\ &\textit{case (ii):} \quad \xi_{\mu_b}(\sigma) \le \frac{2}{v_\wedge} \frac{\overline{p}_b}{\underline{p}_b} \Big(d_{\mathbb{X}}^{\alpha/2} \sigma^{\alpha} + (m-1)(\text{diam}(\mathbb{X})+1)^{\alpha} e^{-1/(2\sigma^2)} \Big). \end{aligned} \tag{5.4}$$

注 2. 当假设 ¨中的动力学 f 在2上是一致利普希茨连续的，且奖励 $\mathbb{Z}$ 是 Hölder 连续时，最优 Q 函数是 H¨older 连续的；例如参见 Harder 和 Peitz [12]。

注意，根据 $V_d(r)$ 的定义，$v_\wedge \le 1$ 是必要的。此外，我们指出，局部体积条件排除了 $\mathbb{X} \times \mathbb{A}$（情形（i））或 $\mathbb{X}$（情形（ii））在任意点 $z \in \mathbb{Z}$附近呈“空心”的情况。特别地，如果 $\mathbb{Z}/\mathbb{X}$是内部非空的凸体，那么局部体积条件成立。因此，如果 μ_b 具有有界密度，且 $\mathbb{Z}$ 例如是一个满维球或矩形，那么近似误差将随着 $\sigma \downarrow 0$ 以速率 $\lesssim \sigma^{\alpha}$ 趋于 0。

我们还指出，要求局部体积条件在半径 $r \le O(\sqrt{d})$ 范围内成立这一技术要求，源于在 $\mathbb{R}^d$ 中匹配球体积和高斯体积。

定理的证明2。我们首先只假设 Q^* 是 α-H¨在d中的更大者。利用不动点恒等式 $Q^* = \mathcal{T}[Q^*]$

以及 $q^* = q^*_{\mu_b} = \mathcal{K}_{\mu_b}[\mathcal{T}[q^*]]$,

$$Q^* - q^* = \overline{\mathcal{T}}[Q^*] - \mathcal{K}_{\mu_b}[\overline{\mathcal{T}}[q^*]] = \big(Q^* - \mathcal{K}_{\mu_b}[Q^*]\big) + \mathcal{K}_{\mu_b}\big(\overline{\mathcal{T}}[Q^*] - \overline{\mathcal{T}}[q^*]\big).$$

取 $\|\cdot\|$, 利用 $\mathcal{K}_{\mu_b}$ 是非扩张的（引理 1），且 $\mathcal{T}$ 是一个 γ-压缩映射，可得

$$\|Q^* - q^*\| \le \|Q^* - \mathcal{K}_{\mu_b}[Q^*]\| + \gamma \|Q^* - q^*\|,$$

因此 $\|Q^* - q^*\| \le \frac{1}{1-\gamma}\|Q^* - \mathcal{K}_{\mu_b}[Q^*]\|$。

固定 $z \in \mathbb{Z}$ 并定义概率测度

$$w_{z,\sigma}(A) := \frac{\int_A \kappa_\sigma(z,u)\mu_b(du)}{\int_{\mathbb{Z}} \kappa_\sigma(z,v)\mu_b(dv)}.$$

则 $\mathcal{K}_{\mu_b}[Q^*](z) = \int_{\mathbb{Z}} Q^*(u) w_{z,\sigma}(du)$，根据 Hölder 正则性,

$$|Q^*(z) - \mathcal{K}_{\mu_b}[Q^*](z)| \le \int_{\mathbb{Z}} |Q^*(z) - Q^*(u)|\, w_{z,\sigma}(du) \le L_Q \int_{\mathbb{Z}} \mathsf{d}(z,u)^\alpha\, w_{z,\sigma}(du).$$

取上确界$_{z\in\mathbb{Z}}$ 可得 $\|Q^* - \mathcal{K}_{\mu_b}[Q^*]\| \le L_Q\, \xi_{\mu_b}(\sigma)$，从而（5.1）得证。

现在，在附加的密度和局部体积假设下证明（5.4）。对于情形（i），令 $d := d_{\mathbb{X}} + d_{\mathbb{A}}$；对于情形（ii），令 $d := d_{\mathbb{X}}$。

De分母下界：情形 (i)。固定 $\sigma \in (0, \sigma_\vee]$，令 $R := \sigma\sqrt{d}$。对于任意 $z \in \mathbb{Z}$,

$$c_{\mu_b}(z) = \int_{\mathbb{Z}} \kappa_\sigma(z,u)\mu_b(du) = \int_{\mathbb{Z}} e^{-|z-u|^2/(2\sigma^2)}\, p_b(u)\, du.$$

利用 $p_b \ge \underline{p}_b$ 并限制到 $B(z, R) := \{u \in \mathbb{Z} : |z-u| \le R\}$，得到

$$c_{\mu_b}(z) \ge \underline{p}_b \int_{B(z,R)} e^{-|z-u|^2/(2\sigma^2)}\, du. \tag{5.5}$$

定义 $\phi(r) := e^{-r^2/(2\sigma^2)} \in (0,1]$。我们注意到，对于所有 $z \in \mathbb{Z}$,

$$\begin{aligned}\frac{1}{\mathrm{Leb}(B(z,R))}\int_{B(z,R)} \phi(|z-u|)\, du &= E\phi(|z-U|) \\ &= \int_0^1 P\left(\phi(|z-U|) \ge t\right) dt. \\ &= \int_0^1 \frac{\mathrm{Leb}\big(\{u \in B(z,R) : \phi(|z-u|) \ge t\}\big)}{\mathrm{Leb}(B(z,R))} dt.\end{aligned} \tag{5.6}$$

其中 U 是定义在 $B(z, R)$ 上的均匀随机变量。注意，对于 $t \in [0, \phi(R))$,

$$\{u \in \mathbb{Z} : \phi(|z-u|) \ge t\} \supset B(z,R).$$

因此，对于所有 $z \in \mathbb{Z}$

$$\begin{aligned}\mathrm{Leb}\big(\{u \in B(z,R) : \phi(|z-u|) \ge t\}\big) &= \mathrm{Leb}\big(B(z,R) \cap \{u \in \mathbb{Z} : \phi(|z-u|) \ge t\}\big) \\ &= \mathrm{Leb}(B(z,R)) \\ &\ge v_\wedge V_d\,(R)\end{aligned} \tag{5.7}$$

其中，最后一个等式由局部体积条件和 $R \le \sigma_\vee\sqrt{d}$ 得出。

另一方面，对于 $t \in [\phi(R), 1],$，我们有 $\sigma\sqrt{-2\log t} \le R \le \sigma_\wedge\sqrt{d}$，并且

$$\{u \in B(z,R) : \phi(|z-u|) \ge t\} = \left\{u \in \mathbb{Z} : |z-u| \le \sigma\sqrt{-2\log t}\right\}.$$

因此，局部体积条件给出

$$\mathrm{Leb}\big(\{u \in B(z,R) : \phi(|z-u|) \ge t\}\big) \ge v_\wedge V_d\left(\sigma\sqrt{-2\log t}\right). \tag{5.8}$$

结合（5.6）、（5.7）和（5.8），可知对所有 $z \in \mathbb{Z}$

$$\begin{aligned}
\int_{B(z,R)} \phi(|z-u|)\,du &\ge v_\wedge\left(\phi(R)V_d(R) + \int_{\phi(R)}^1 V_d\left(\sigma\sqrt{-2\log t}\right)dt\right)\\
&= v_\wedge\int_0^{\phi(R)} \mathrm{Leb}\big(\{u : |u-z| \le R\}\big)dt + \int_{\phi(R)}^1 \mathrm{Leb}\big(\{u : \phi(|z-u|) \ge t\}\big)dt\\
&\overset{(i)}{=} v_\wedge\int_0^{\phi(R)} \mathrm{Leb}\big(\{u : |u| \le R\}\big)dt + \int_{\phi(R)}^1 \mathrm{Leb}\big(\{u : \phi(|u|) \ge t\}\big)dt\\
&= v_\wedge\int_{|u|\le R} \phi(|u|)du.
\end{aligned}$$

其中（i）利用了勒贝格测度在平移下的不变性，最后一个不等式则由与（5.6）类似的论证得出。

令 $G \sim N(\mathbf{0},\ I_d)$。于是

$$\begin{aligned}
\int_{|u|\le R} \phi(|u|)du &= (2\pi)^{d/2}\sigma^d\,P(|G| \le R/\sigma)\\
&= (2\pi)^{d/2}\sigma^d\,P(|G|^2 \le d)\\
&\ge \tfrac{1}{2}(2\pi)^{d/2}\sigma^d,
\end{aligned}$$

其中，最后一个不等式利用了以下事实：$|G|^2 \sim \chi^2_d$ 的中位数（$|G|^2$）$\le d$。因此，回到（5.5）

$$\inf_{z\in\mathbb{Z}} c_{\mu_b}(z) \ge \tfrac{1}{2}\,\underline{p}_b\,v_\wedge\,(2\pi)^{d/2}\sigma^d. \tag{5.9}$$

分母下界，情形（ii）。固定 $z = (x, a_i) \in \mathbb{X}\times\mathbb{A}$。记 $\mu_b(dy, a_j) = p_b(y, a_j)\,dy$,

$$c_{\mu_b}(x,a_i) = \sum_{j=1}^m \int_{\mathbb{X}} \exp\Big(-\frac{|x-y|^2 + \mathbb{1}_{\{i\ne j\}}}{2\sigma^2}\Big)\,p_b(y,a_j)\,dy \ \ge\ \underline{p}_b\int_{\mathbb{X}} e^{-|x-y|^2/(2\sigma^2)}\,dy.$$

在 $\mathbb{X} \subset \mathbb{R}^{d_\mathbb{X}}$ 上应用与情形（i）相同的论证（并在 $\mathbb{X}$ 上采用欧几里得局部体积条件），可得

$$\inf_{(x,a)\in\mathbb{Z}} c_{\mu_b}(x,a) \ge \tfrac{1}{2}\,\underline{p}_b\,v_\wedge\,(2\pi)^{d_\mathbb{X}/2}\sigma^{d_\mathbb{X}}. \tag{5.10}$$

分子上界，情形（i）。这里 $\mathsf{d}(z,\ u) = |z-u|$ 在 $\mathbb{R}^d$ 上，满足 $d = d_\mathbb{X} + d_\mathbb{A}$。对于任意 $z \in \mathbb{Z}$,

$$\int_{\mathbb{Z}} |z-u|^\alpha\,\kappa_\sigma(z,u)\mu_b(du) \le \overline{p}_b\int_{\mathbb{R}^d} |v|^\alpha \exp\Big(-\frac{|v|^2}{2\sigma^2}\Big)\,dv.$$

令 $G \sim N(0,\ I_d)$。则由高斯矩可得

$$\begin{aligned}\frac{1}{(2\pi)^{d/2}\sigma^d}\int_{\mathbb{R}^d}|v|^\alpha\, e^{-|v|^2/(2\sigma^2)}dv &= \sigma^\alpha\,\mathbb{E}[|G|^\alpha]\\ &\le \sigma^\alpha\,(\mathbb{E}[|G|^2])^{\alpha/2}\\ &= \sigma^\alpha\, d^{\alpha/2},\end{aligned}$$

其中我们使用了詹森不等式。因此，

$$\int_{\mathbb{R}^d}|v|^\alpha\, e^{-|v|^2/(2\sigma^2)}dv \le (2\pi)^{d/2}\sigma^d\sigma^\alpha\, d^{\alpha/2}$$

将其与（5.9）结合可得

$$\xi_{\mu_b}(\sigma) \le d^{\alpha/2}\frac{2\overline{p}_b}{v_\wedge \underline{p}_b}\sigma^\alpha.$$

这说明了（5.4）中的情形（i）。

分子上界，情形（ii）。令 $\mathbb{A} = \{a_1,\dots,a_m\}$, $m := |\mathbb{A}|$, 并记 $z = (x, a_i)$。设$D_{\mathbb{X}} := \mathrm{diam}(\mathbb{X}) := \sup_{x,y\in\mathbb{X}}|x-y| < \infty$。利用 $p_b \le p_b$ 并拆分 $j = i$ 和 $\overline{j} \ne i$,

$$\begin{aligned}&\int_{\mathbb{Z}}\mathsf{d}(z,u)^\alpha\kappa_\sigma(z,u)\mu_b(du)\\ &\le \overline{p}_b\int_{\mathbb{X}}|x-y|^\alpha e^{-|x-y|^2/(2\sigma^2)}dy \;+\; \overline{p}_b\sum_{j\ne i}\int_{\mathbb{X}}(|x-y|^2+1)^{\alpha/2}e^{-(|x-y|^2+1)/(2\sigma^2)}dy\\ &\le \overline{p}_b\int_{\mathbb{R}^d}|v|^\alpha e^{-|v|^2/(2\sigma^2)}dv \;+\; \overline{p}_b(m-1)(D_{\mathbb{X}}^2+1)^{\alpha/2}e^{-1/(2\sigma^2)}\int_{\mathbb{R}^{d_{\mathbb{X}}}}e^{-|v|^2/(2\sigma^2)}dv\\ &= \overline{p}_b(2\pi)^{d_{\mathbb{X}}/2}\sigma^{d_{\mathbb{X}}}\left[\sigma^\alpha\mathbb{E}|G|^\alpha + (m-1)(D_{\mathbb{X}}^2+1)^{\alpha/2}e^{-1/(2\sigma^2)}\right]\\ &\le \overline{p}_b(2\pi)^{d_{\mathbb{X}}/2}\sigma^{d_{\mathbb{X}}}\left[\sigma^\alpha d_{\mathbb{X}}^{\alpha/2} + (m-1)(D_{\mathbb{X}}^2+1)^{\alpha/2}e^{-1/(2\sigma^2)}\right]\end{aligned}$$

其中 $G \sim N(0, I_{d_{\mathbb{X}}})$。

除以上所示的分母下界 $\int_{\mathbb{Z}}\kappa_\sigma(z,u)\mu_b(du)$ 后，对所有$\sigma \in (0, \sigma_\vee]$,

$$\xi_{\mu_b}(\sigma) \le \frac{2}{v_\wedge}\frac{\overline{p}_b}{\underline{p}_b}\left(d_{\mathbb{X}}^{\alpha/2}\sigma^\alpha + (m-1)(D_{\mathbb{X}}^2+1)^{\alpha/2}e^{-1/(2\sigma^2)}\right).$$

这与不等式 $a^2+b^2 \le (|a|+|b|)^2$ 推出了（5.4）中的情形（ii）。 □

6 数值实验

我们在一个具有连续状态空间和有限动作集的双品种缺货损失库存控制问题上测试 Q-测度学习（算法1）。为使行文简洁，部分数值和实现细节将在附录C中介绍。

库存模型

首先，我们定义受控马尔可夫链 $\{(X_t,\ A_t):\ t \ge 0\}$ 在 $\mathbb{Z} = \mathbb{X}\times\mathbb{A}$ 上。状态是现有库存向量

$$X_t = (X_{t,1}, X_{t,2}) \in \mathbb{X} := [0, I_{\max}]^2,$$

其中 $I_{\max} > 0$ 是 th 每种物品的最大库存水平（容量）。动作是订购数量向量 r

$$A_t = (A_{t,1}, A_{t,2}) \in \mathbb{A} := \{0, 1, \ldots, A_{1,\max}\} \times \{0, 1, \ldots, A_{2,\max}\},$$

因此 $\mathbb{A}$ 是有限的。图表使用一个满足 $I_{\max} = 15$ 且 $(A_{1,\max}, A_{2,\max}) = (10, 8)$ 的实例构建。

令 $\{D_t = (D_{t,1}, D_{t,2}) : t \geq 0\}$ 为非负需求向量的独立同分布序列。具体而言，取相关高斯向量的绝对值来引入坐标之间的相关性，即 $D_t = |G_t|$ 其中$G_t \sim N(\mu, \Sigma)$。随后，丢失销售库存动态演化为

$$X_{t+1} = \min\left\{(X_t + A_t - D_t)_+, I_{\max}\right\}, \tag{6.1}$$

其中 $(\cdot)_+$ 表示逐分量正部。

此外，令 h，p，$c \in \mathbb{R}^2_{>0}$ 分别表示持有成本、丢失销售成本和单位订货成本向量，并令 $k > 0$ 表示固定订货成本。单步成本为

$$C_{t+1} = k\,\mathbb{1}\{A_{t,1} + A_{t,2} > 0\} + c^\top A_t + h^\top X_{t+1} + p^\top (D_t - X_t - A_t)_+. \tag{6.2}$$

由于 D 是无界的，为大致匹配假设 1，我们对成本进行归一化以获得有界奖励；也就是说，$R_{t+1} = -C_{t+1}/C_{\max}$ 其中 $C_{\max}$ 以高概率满足 $C_t \leq C_{\max}$。

在这种设定下，动态规划原理成立，其奖励函数为 $r(x, a) = E[R_1 | X_0 = x, A_0 = a]$，最优 Q^* 是下式的唯一解：$\mathcal{T}[Q^*] = Q^*$，其中 $\mathcal{T}$由式（2.1）给出，其受控转移核为 $P(A|x, a) = E[\mathbb{1}\{X_1 \in A\} | X_0 = x, A_0 = a]$ 对所有 $A \in \mathcal{X}$均成立。

算法设置与基准

我们使用完全探索型行为策略下单条轨迹生成的数据，实现基于权重的算法1 。具体而言，我们在有限动作集上采用均匀行为策略；也就是说，$\pi_b(a|x) = 1/|\mathbb{A}|$ 对所有 $a \in \mathbb{A}$ 以及 $x \in \mathbb{X}$均成立。在这一探索策略下，我们通过实证验证状态空间得到了充分探索，这支持定理2 中的密度假设得到满足（参见附录3中的图C）。

尽管 $\mathbb{A}$ 是有限的，我们仍将其视为 $\mathbb{R}^2$ 的一个紧子集，并实现第4.1节中的情形(i)核。对于论文中展示的图，我们设 $\sigma = 1$。我们指出，在额外实验中，对于 $\sigma \in [0.5, 2]$，我们也观察到了类似的收敛和策略误差表现。相反，若将 σ 选得过小（例如 $\sigma = 0.1$），则会对 q_n平滑不足，导致估计明显更粗糙，收敛也更慢。此外，我们按照假设4选择步长，其中 $\beta_n = 1/(n+1)$，且 $\alpha_n \sim 1/n$。

由于需求相关且存在正的固定订货成本，我们并不知道其闭式解。因此，我们使用动态规划在量化状态空间上计算最优Q^* 的近似 Q_{DP} ，并通过蒙特卡罗估计需求分布。

收敛性与策略行为

图 1 展示了训练期间计算的两项收敛性诊断指标：由 q_n导出的贪心策略的估计折扣回报，以及一条跟踪固定评估网格上状态-动作对误差的均方根误差（RMSE）曲线。需要注意的是，这两幅图均由 单条训练轨迹生成。图中显示，随着已完成迭代次数 n的增加，回报上升而均方根误差下降。这与第 4 节中的收敛机制一致：随着耦合跟踪器（μ_n，ν_n）趋于稳定，诱导函数 $q_n = \Phi_{\mu_n}[\nu_n]$ 逐渐接近其极限不动点 q^*，相应的贪心策略也趋于稳定。

我们还注意到，两张图都显示与最优回报以及均方根误差为 0 之间存在持续差距。这与我们的理论一致：平滑参数 $\sigma > 0$ 会引入严格为正的近似误差，因此 $\|q^* - Q^*\| > 0$。

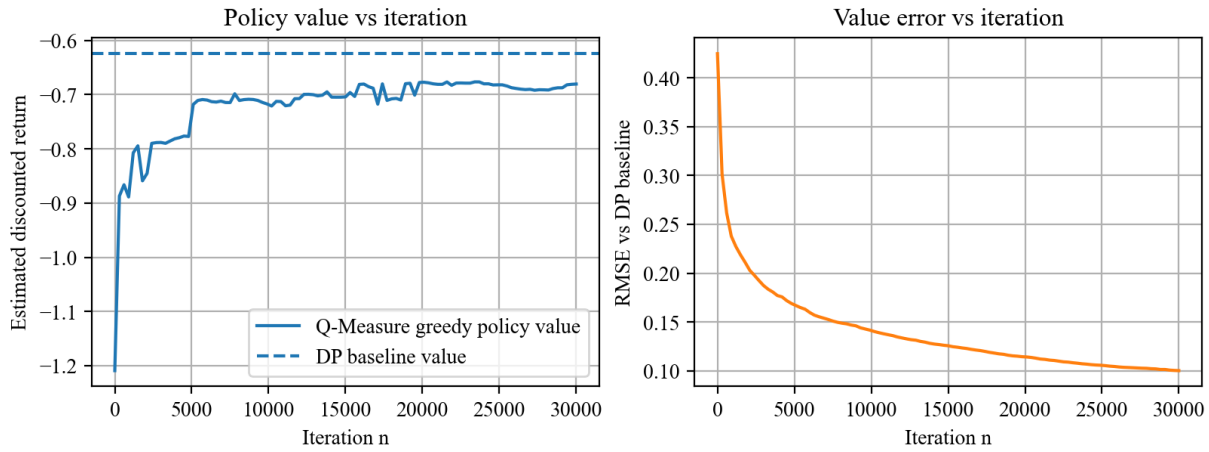

图1：左：由 q_n 导出的贪心策略的估计折扣回报。右： Q^* 估计误差，曲线呈现出与收敛相符的稳定/下降趋势。

图 2 展示了在库存状态空间上学得的贪心策略，并将其与相同的离线基准策略进行比较。这两种策略呈现出相似的结构，这正是损失销售系统的典型特征：库存较低时，策略会下达正订单；库存足够高时，策略则倾向于不下订单。由此产生的 $\mathbb{X}$ 订货区域与非订货区域划分高度一致，这支持了学得策略在定性上的正确性。

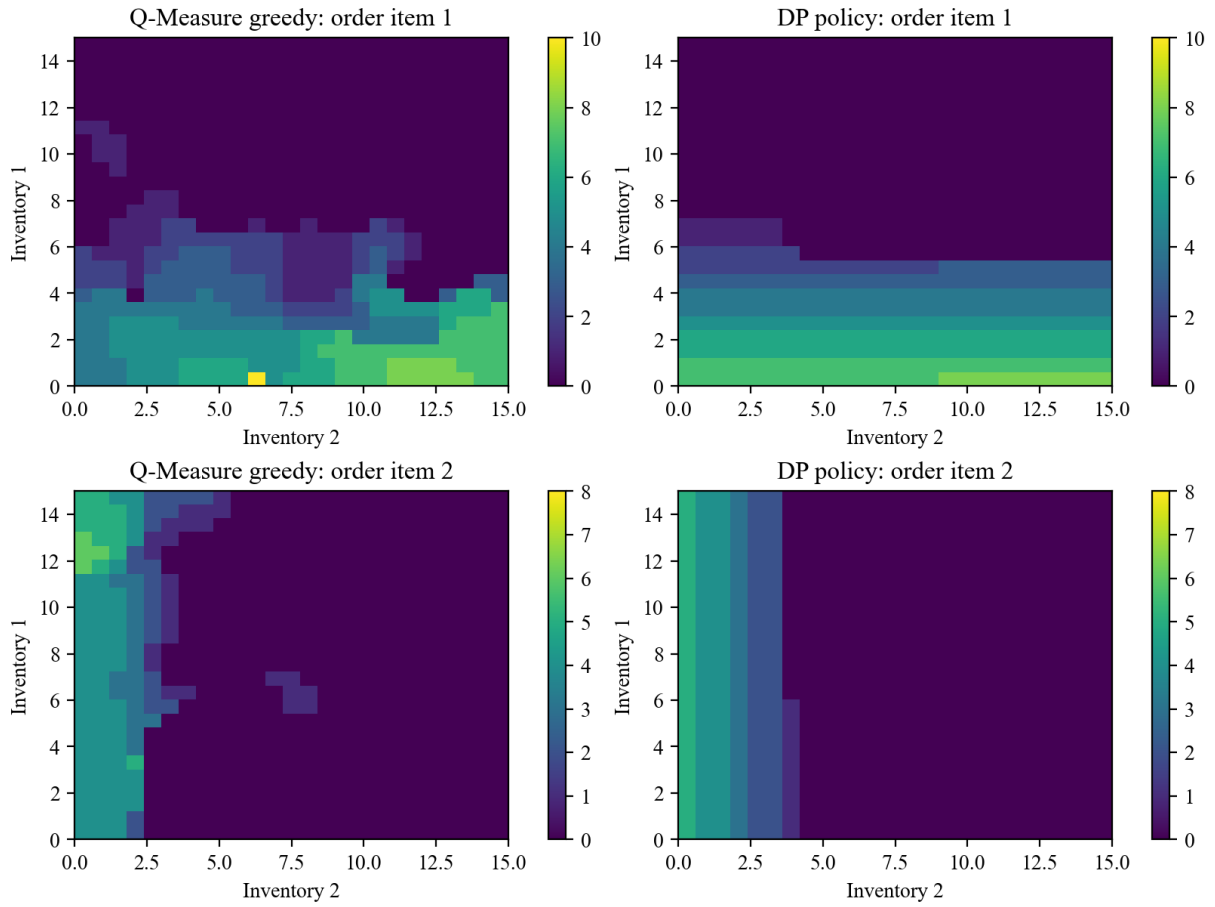

图 2：由 q_n 产生的贪心策略与网格上的动态规划策略对比。

参考文献

[1] Barreto, A. M., Precup, D., and Pineau, J.（2016年）。实用的基于核的强化学习。机器学习研究杂志，17(67):1–70。3[2] Beck, C., Jentzen, A., Kleinberg, K., and Kruse, T.（2025年）。用于离散时间、高维随机最优控制问题贝尔曼方程的多项式运行时间非线性蒙特卡洛方法。应用数学与优化，91(1):26。3[3] Borkar, V. S. and Meyn, S. P.（2000年）。用于随机逼近和强化学习收敛性的常微分方程方法。*SIAM* 控制与优化杂志，38(2):447–469。2[4] Boyan, J. A.（2002年）。技术更新：最小二乘时序差分学习。机器学习，49(2):233–246。3[5] Bradtke, S. J. and Barto, A. G.（1996年）。用于时序差分学习的线性最小二乘算法。机器学习，22(1):33–57。3[6] Br´ezis, H.（2011年）。泛函分析、索伯列夫空间与偏微分方程。施普林格。10[7] Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., and Maguluri, S. T.（2022年）。非线性随机逼近的有限样本分析及其在强化学习中的应用。自动化学报，146:110623。2

[8] Doob, J. L.（1953年）。随机过程。John Wiley & Sons。25[9] Engel, Y., Mannor, S., and Meir, R.（2005年）。利用高斯过程进行强化学习。载于第22届国际机器学习会议论文集，第201–208页。3[10] Ernst, D., Geurts, P., and Wehenkel, L.（2005年）。基于树的批量强化学习。机器学习研究杂志，6。2[11] Gordon, G. J.（1995年）。动态规划中的稳定函数逼近。载于1995年机器学习论文集，第261–268页。爱思唯尔。2[12] Harder, H. and Peitz, S.（2024年）。强化学习与最优控制中价值函数的连续性和平滑性。载于2024年IEEE第63届决策与控制会议（CDC），第1935–1940页。IEEE。15[13] Jaakkola, T., Jordan, M., and Singh, S.（1993年）。随机迭代动态规划算法的收敛性。神经信息处理系统进展，6。2[14] Jong, N. and Stone, P.（2006年）。用于强化学习的基于核的模型。载于ICML核机器与强化学习研讨会。3[15] Kushner, H. J. and Yin, G. G.（2003年）。随机逼近与递归算法及其应用。施普林格。10[16] Lagoudakis, M. G. and Parr, R.（2003年）。最小二乘策略迭代。机器学习研究杂志，4(12月):1107–1149。3[17] Lagoudakis, M. G., Parr, R., and Littman, M. L.（2002年）。控制中的强化学习最小二乘方法。载于希腊人工智能会议论文集，第249–260页。施普林格。3[18] Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y.（2024年）。Q学习是否达到极小极大最优？紧致的样本复杂度分析。运筹学，72(1):222–236。2[19] Meunier, M., Reisinger, C., and Zhang, Y.（2025年）。通过多层蒙特卡罗实现熵正则化马尔可夫决策过程的高效学习。3[20] Meyn, S. P. and Tweedie, R. L.（2012年）。马尔可夫链与随机稳定性。施普林格科学与商业媒体。8[21] Mou, W., Khamaru, K., Wainwright, M. J., Bartlett, P. L., and Jordan, M. I.（2022年）。巴拿赫空间中最优的方差缩减随机逼近。arXiv预印本 arXiv:2201.08518。2[22] Munos, R. and Szepesv´ari, C.（2008年）。拟合价值迭代的有限时间界。机器学习研究杂志，9(5)。2[23] Ormoneit, D. and Glynn, P.（2002年）。平均成本问题中的基于核的强化学习。IEEE自动控制汇刊，47(10):1624–1636。3[24] Ormoneit, D. and Sen, ´S.（2002年）。基于核的强化学习。机器学习，49(2):161–178。3[25] Rudin, W.（1974年）。实分析与复分析。麦格劳-希尔科学、工程与数学。24[26] Stachurski, J.（2008年）。通过非扩张逼近实现连续状态动态规划。《计算经济学》，31(2):141–160。2, 4

[27] Steinwart, I.（2001年）。论核对支持向量机一致性的影响。机器学习研究杂志，2（11月）：67–93。24[28] Taylor, G.和Parr, R.（2009年）。用于强化学习的核化价值函数近似。载于第*26*届国际机器学习年会论文集，第1017–1024页。3[29] Tsitsiklis, J.和Van Roy, B.（1997年）。带函数近似的时序差分学习分析。*IEEE*自动控制汇刊，42(5):674–690。2[30] Tsitsiklis, J. N.（1994年）。异步随机逼近与Q学习。机器学习，16(3):185–202。2[31] Wainwright, M. J.（2019年）。带锥压缩算子的随机逼近：Q学习的尖锐L无穷范数界。*arXiv*预印本 *arXiv:1905.06265*。2[32] Watkins, C. J.和Dayan, P.（1992年）。Q学习。机器学习，8(3):279–292。1，2，6

附录

命题的证明 4.1

我们将使用以下来自泛函分析的辅助结果。

引理 6 (Rudin [25], 定理 6.19)。 设 $\mathbb{Z}$ 为紧豪斯多夫空间，且 $\mu \in \mathcal{M}(\mathcal{Z})$。若对所有 $f \in C(\mathbb{Z})$都有 $\int_{\mathbb{Z}} f d\mu = 0$，则 $\mu = 0$。

No请注意，这是 Riesz 表示定理关于 $C(\mathbb{Z})$ 的唯一性陈述。

引理 7 (Steinwart [27], 例 1)。 固定 $\sigma > 0$。对于每个非空紧致 $K \subset \mathbb{R}^d$，高斯核 $\kappa(x, y) = \exp(-|x-y|^2/(2\sigma^2))$ 在 K 上具有普适性；也就是说，对于每个 $f \in C(K)$ 和 $\epsilon > 0$，存在一个由 κ 诱导的函数 g，使得 $\|f - g\| \leq \epsilon$。

有了这些结果，我们证明命题 4.1。

命题 *4.1*的证明。 首先，我们注意到，容易证明 D_μ 是非负、对称的，并且满足三角不等式。我们验证 $D_\mu(\nu, \nu') = 0$ 意味着 $\nu = \nu'$。假设 $D_\mu(\nu, \nu') = 0$并令 $m := \nu - \nu'$。

接下来，我们验证 κ 在 $\mathbb{Z}$上的全称性。在情形（i）中，$\mathbb{Z} \subset \mathbb{R}^{d_{\mathbb{X}}+d_{\mathbb{Z}}}$ 是紧致的，因此引理 7 适用，其中 $d = d_{\mathbb{X}} + d_{\mathbb{Z}}$。在情形（ii）中，定义可逆映射 $\psi\colon \mathbb{X} \times \mathbb{A} \to \mathbb{R}^{d_{\mathbb{X}}+m}$ 如下：$\psi(x, a_i) = (x, e_i/\sqrt{2})$, 其中 (e_i) 是 $\mathbb{R}^m$ 的标准基。于是 $\psi(\mathcal{Z})$ 是紧致的，且对所有 $z, u \in \mathcal{Z}$ 和

$$\kappa(z, u) = \exp\left(-\frac{|\psi(z) - \psi(u)|^2}{2\sigma^2}\right)$$

因此 κ_σ 是紧致集 $\psi(\mathbb{Z}) \subset \mathbb{R}^{d+m}$上的高斯核，它在 $\psi($ $\mathbb{Z})$.

定义 $L\colon C(\mathbb{Z}) \to \mathbb{R}$ 由 $L[f] := \int_{\mathbb{Z}} f dm$给出；于是 $|L[f]| \leq \|f\|_\infty \|m\|_{\mathrm{TV}}$，因此 L是连续的。

为论证 $m = 0$，我们采用反证法，并假设 $\|m\|_{\mathrm{TV}} > 0$。任取 $f \in C(\mathbb{Z})$ 和 $\eta > 0$。我们证明 $L[f] < \eta$。由全称性，存在一个希尔伯特空间 $(H, \langle\cdot,\cdot\rangle_H)$、一个特征映射$\Psi\colon \mathbb{Z} \to H$，满足 $\kappa(x, y) = \langle\Psi(x)\ \Psi(y)\rangle_H$，以及一个由此诱导的函数 $g(\cdot) = \langle w, \Psi(\cdot)\rangle_H$，使得

$$\|f - g\| < \frac{\eta}{2\|m\|_{\mathrm{TV}}}.$$

令 $H_0 := \overline{\text{张成}\{\Psi(z) : z \in \mathbb{Z}\}}$，并将 $H \ni w = w_0 + w_\perp$ 写成满足 $w_0 \in H_0$ 和 $w_\perp \in H_0^\perp$的形式。于是 $g(\cdot) = \langle w_0, \Psi(\cdot)\rangle_H$。由于 $w_0 \in H_0$，我们可以选择

$$w_n = \sum_{i=1}^{n} \alpha_i \Psi(z_i) \in \operatorname{span}\{\Psi(z) : z \in \mathbb{Z}\}$$

使得 $\|w_0 - w_n\|_H < \eta/(2M\|m\|_{\mathrm{TV}})$，其中 $M := \sup_{z\in\mathbb{Z}} \|\Psi(z)\|_H = \sup_z \sqrt{\kappa(z, z)} = 1$。则函数 $h(\cdot) := \langle w_n, \Psi(\cdot)\rangle_H = \sum_{i=1}^n \alpha_i \kappa(z_i, \cdot)$ 满足

$$\|g - h\| \leq M\|w_0 - w_n\|_H < \eta/(2\|m\|_{\mathrm{TV}}).$$

由于 κ 在紧致的 $\mathbb{Z}$ 上为正，$D_\mu(\nu, \nu') = 0$ 意味着，对所有 $z \in \mathbb{Z}$，$\int_{\mathbb{Z}} \kappa(z, u) m(du) = 0$。因此，对所有 i，$L[\kappa(z_i, \cdot)] = 0$，并且我们有 $L[h] = 0$。所以

$$|L[f]| \leq |L[f-g]| + |L[g-h]| + |L[h]| \leq \|m\|_{\mathrm{TV}} (\|f - g\| + \|g - h\|) < \eta.$$

由于 $\eta > 0$ 是任意的， $L[f] = 0$ 对所有 $f \in C(\mathbb{Z})$ 。但根据引理 6，只有在 $m = 0$的情况下才会出现这种情况，这与我们关于 $\|m\|_{\mathrm{TV}} > 0$的假设矛盾。因此必有 $m = 0$；即 $\nu = \nu'$。 □

B 命题 4.3 的证明

我们首先介绍鞅和马尔可夫链的两个经典结果。

引理 8 (Doob [8], 第七章，定理 4.1)。 设$(M_n, \mathcal{F}_n)_{n\geq 0}$ 是一个实值鞅，满足上确界$_{n\geq 0}\, E[M_n^2] < \infty$。则 M_n 几乎必然收敛，且当 $n \to \infty$ 趋于 L^2 时，在 L^2 中收敛于一个平方可积随机变量M_∞。特别地，收敛的实数列是柯西列，因此当 $n \to \infty$ 时，上确界$_{m\geq n}\, |M_m - M_n| \to 0$ 几乎必然成立。

引理 9 (泊松方程)。 假设假设 3。设 $f \in C(\mathbb{Z})$ 并令 $f_c : = f - \mu_b[f]$。定义

$$v_f(z) := \sum_{t=0}^{\infty} P_b^t[f_c](z), \qquad z \in \mathbb{Z}. \tag{B.1}$$

那么，级数 (B.1) 在 $\mathbb{Z}$ 上绝对且一致收敛， v_f 有界可测，并且它满足泊松方程 $v_f - P_b[v_f] = f_c$ 此外，

$$\|v_f\| \leq \frac{2c}{1-\rho} \|f\| . \tag{B.2}$$

引理的证明 9。 根据假设 3，对于任意有界的 h 以及任意 $z \in \mathbb{Z}$,

$$\left|P_b^t[h](z) - \mu_b[h]\right| \leq \|h\| \left\|P_b^t(z,\cdot) - \mu_b(\cdot)\right\|_{\mathrm{TV}} \leq c\rho^t \|h\| .$$

将此应用于 $h = f_c$并利用 $\mu_b[f_c] = 0$得到 $\|P_b^t[f_c]\| \leq c\rho^t \|f_c\| \leq 2c\rho^t \|f\|$。因此 $\sum_{t\geq 0} \|P_b^t f_c\| < \infty$，这意味着 (B.1) 的绝对一致收敛，并得到 (B.2)。随后，绝对一致收敛使得可以逐项应用 P_b，从而得到 $P_b[v_f] = \sum_{t\geq 0} P_b^{t+1}[f_c] = \sum_{t\geq 1} P_b^t[f_c]$，因此$v_f - P_b[v_f] = f_c$。

□

命题的证明 4.3。 定义如下：对于 $m \geq n$,

$$S_{n,m}^D := \sum_{k=n}^{m-1} \alpha_{k+1} D_{k+1}, \qquad S_{n,m}^G := \sum_{k=n}^{m-1} \alpha_{k+1} G_k, \qquad S_{n,m}^B := \sum_{k=n}^{m-1} \alpha_{k+1} B_k.$$

我们将证明这些和式中每一个的几乎必然收敛。

偏差项 $S_{n,\,m}^B$**：** 根据 $\mathcal{N}(n,\ T)$ 的定义，对于每个 $m \in \mathcal{N}(n,\ T)$，都有 $\sum_{k=n}^{m-1} \alpha_{k+1} \leq T$，于是 ce

$$\max_{m\in\mathcal{N}(n,T)} \left\|S_{n,m}^B\right\| \leq \sum_{k=n}^{\max \mathcal{N}(n,T)-1} \alpha_{k+1} \|B_k\| \leq T \sup_{k\geq n} \|B_k\| . \tag{B.3}$$

对于每个 k，根据 B_k 在（4.5）中的定义、界 $\|\Phi_{\mu_b}[\delta_{Z_k}]\| \leq \kappa_\wedge^{-1}$以及 γ对 $\mathcal{T}$的压缩映射性质，

$$\|B_k\| \leq \frac{1}{\kappa_\wedge} \left\|\overline{\mathcal{T}}[q_k] - \overline{\mathcal{T}}[\bar{q}_k]\right\| \leq \frac{\gamma}{\kappa_\wedge} \|q_k - \bar{q}_k\| .$$

引理 3 可得 $\|q_k - \bar{q}_k\| \to 0$ 几乎处处成立，因此 $\sup_{k\geq n} \|B_k\| \to 0$ 几乎处处成立。因此，由（B.3），可得

$$\max_{m\in\mathcal{N}(n,T)} \left\|S_{n,m}^B\right\| \to 0 \tag{B.4}$$

几乎必然地，当 $n \to \infty$。

一个 δ-网论证： 我们控制 $S^D_{m,\ n}$ 和 $S^G_{m,\ n}$ 的上确界误差的策略，始于观察到这些误差关于 z是一致利普希茨的。因此，对于任意 $\delta>0$，我们都可以构造一个有限的 δ-网，只需跟踪网内这些点处的误差。

具体而言，由于 $\mathbb{Z}$是紧致的，对于所有 z，z'，$u\in\mathbb{Z}$，由中值定理可得

$$|\kappa(z,u)-\kappa(z',u)|\le\frac{\operatorname{diam}(\mathbb{Z})}{\sigma^2}|z-z'|.$$

Here 当 $\mathbb{A}$ 为有限集时，我们滥用记号，将 $|(x,\ a_i)-(x',\ a_j)|=|x-x'|+\mathbb{1}_{i\neq j}$表示为 .

设 $L_\kappa:=\operatorname{diam}(\mathbb{Z})/\sigma^2$。回忆一下，在（4.2）中，我们定义了 $c_{\mu_b}(z)=\int_{\mathbb{Z}}\kappa(z,\ u)\ \mu_b(du)$。因此，$z\to c_{\mu_b}(z)$是$L_\kappa$-利普希茨的。因此，对每个 u，z，$z'\in\mathbb{Z}$,

$$\begin{aligned}|\Phi_{\mu_b}[\delta_u](z)-\Phi_{\mu_b}[\delta_u](z')|&=\left|\frac{\kappa(z,u)}{c_{\mu_b}(z)}-\frac{\kappa(z',u)}{c_{\mu_b}(z')}\right|\\&\le\left|\frac{\kappa(z,u)-\kappa(z',u)}{c_{\mu_b}(z)}\right|+\left|\frac{\kappa(z',u)(c_{\mu_b}(z)-c_{\mu_b}(z'))}{c_{\mu_b}(z)c_{\mu_b}(z')}\right|\\&\le L_\kappa\left(\frac{1}{\kappa_\wedge}+\frac{1}{\kappa_\wedge^2}\right)|z-z'|;\end{aligned}$$

即 $\Phi_{\mu_b}[\delta_u]$ 是利普希茨连续的，常数为 $L_\Phi:=L_\kappa(\frac{1}{\kappa_\wedge}+\frac{1}{\kappa_\wedge^2})$。此外，$\|\Phi_{\mu_b}[\delta_u]\|\le\frac{1}{\kappa_\wedge}$.

此外，根据引理 2，$|Y_{k+1}|\le\frac{1}{1-\gamma}$因此 $|E[Y_{k+1}|\mathcal{F}_k]|\le\frac{1}{1-\gamma}$。于是 $D_{k+1}=\big(Y_{k+1}-E[Y_{k+1}|\mathcal{F}_k]\big)\Phi_{\mu_b}[\delta_{Z_k}]$满足

$$\|D_{k+1}\|\le\frac{2}{(1-\gamma)\kappa_\wedge},\qquad|D_{k+1}(z)-D_{k+1}(z')|\le\frac{2L_\Phi}{1-\gamma}|z-z'|.\tag{B.5}$$

此外，在 $\|\mathcal{T}_{\mu_b}[\bar{q}_k]\|\le\frac{1}{1-\gamma}$与$\|\mathcal{I}[\bar{q}_k]\|\le\frac{1}{1-\gamma}$下，马尔可夫项 $G_k=\mathcal{T}[\bar{q}_k](Z_k)\Phi_{\mu_b}[\delta_{Z_k}]-\mathcal{T}_{\mu_b}[\bar{q}_k]$满足一致利普希茨界

$$\begin{aligned}|G_k(z)-G_k(z')|&\le\frac{L_\Phi}{1-\gamma}|z-z'|+\mathcal{K}_{\mu_b}[\overline{\mathcal{T}}[\bar{q}_k]](z)-\mathcal{K}_{\mu_b}[\overline{\mathcal{T}}[\bar{q}_k]](z')\\&\le\frac{L_\Phi}{1-\gamma}|z-z'|+\frac{\int_{\mathbb{Z}}|c_{\mu_b}(z')\kappa(z,u)-c_{\mu_b}(z)\kappa(z',u)|\,\overline{\mathcal{T}}[\bar{q}_k](u)\mu_b(du)}{c_{\mu_b}(z)c_{\mu_b}(z')}\\&\le\frac{2L_\Phi}{1-\gamma}|z-z'|.\end{aligned}\tag{B.6}$$

固定 $\epsilon>0$。设 $L_0:=\frac{2}{1-\gamma}L_\Phi$, 且 $\delta:=\frac{\epsilon}{L_0T}$. 由于 $\mathbb{Z}$是紧致的，存在有限 δ-网 $\{z_1,\dots,z_l\}\subset\mathbb{Z}$，使得对任意 $z\in\mathbb{Z}$，存在满足 $|z-z_j|\le\delta$的 $j\le N$。于是，对任意L-利普希茨函数 $h:\mathbb{Z}\to\mathbb{R}$，

$$\|h\|=\sup_{z\in\mathbb{Z}}|h(z)|\le\max_{1\le j\le l}\left|h(z^{(j)})\right|+L\delta.\tag{B.7}$$

如果 $m\in\mathcal{N}(n,\ T)$，那么 $\sum_{k=n}^{m-1}\alpha_{k+1}\le T$，因此由 (B.5)–(B.6) 可知，函数 $S^D_{n,\ m}$ 和 $S^G_{n,\ m}$ 都是 L_0T-利普希茨的。因此，对所有 n,

$$\begin{aligned}\max_{m\in\mathcal{N}(n,T)}\left\|S^D_{n,m}\right\|&\le\max_{1\le j\le l}\max_{m\in\mathcal{N}(n,T)}\left|S^D_{n,m}(z_j)\right|+\epsilon,\\\max_{m\in\mathcal{N}(n,T)}\left\|S^G_{n,m}\right\|&\le\max_{1\le j\le l}\max_{m\in\mathcal{N}(n,T)}\left|S^G_{n,m}(z_j)\right|+\epsilon.\end{aligned}\tag{B.8}$$

因此，只需控制有限多个点 z_1，... ，z_l 处的误差过程。

鞅项 D_{k+1}。固定 $j\in\{1,\dots,l\}$ 并考虑标量过程 $\{S^D_{0,m}(z_j):\ m\ge0\}$。由于$E[D_{k+1}(z_j)\ |\mathcal{F}_k]=0$ 根据 D_{k+1}的定义，$\{S^D_{0,m}(z_j):\ m\ge0\}$ 是一个 $\mathcal{F}_m$- 鞅。此外，利用（B.5）

以及鞅差的正交性,

$$E\big[(S^D_{0,m}(z_j))^2\big] = \sum_{k=0}^{m-1} \alpha_{k+1}^2 E\big[D_{k+1}(z_j)^2\big]$$
$$\leq \frac{4}{(1-\gamma)^2\kappa_\wedge^2} \sum_{k=0}^{\infty} \alpha_{k+1}^2 < \infty.$$

因此，引理 8 成立，我们由此得出

$$\max_{m\in\mathcal{N}(n,T)} \big|S^D_{n,m}(z_j)\big| \leq \sup_{m\geq n} \big|S^D_{0,m}(z_j) - S^D_{0,n}(z_j)\big| \to 0$$

几乎必然收敛，当 $n\to\infty$。

对有限集 $j = 1,\ldots,l$ 取最大值仍能保持几乎必然收敛。因此，(B.8)表明，对于任意 $\epsilon > 0$，存在 Ω_ϵ 使得 Ω_ϵ^c 是 P-零测集，且对所有 $\omega\in\Omega_\epsilon$

$$\limsup_{n\to\infty} \max_{m\in\mathcal{N}(n,T)} \big\|S^D_{n,m}\big\| \leq \epsilon.$$

令 $\Omega_* = \bigcap_{k=1}^\infty \Omega^{1/k}$。显然， $P(\Omega_* = 1)$，且对所有 $\omega\in\Omega_*$

$$\limsup_{n\to\infty} \max_{m\in\mathcal{N}(n,T)} \big\|S^D_{n,m}\big\| = 0 \tag{B.9}$$

这表明几乎必然收敛。

通过泊松方程得到马尔可夫项 G_k。定义 $\forall u\in\mathbb{Z}$,

$$f_{j,k}(u) := \overline{\mathcal{T}}[\bar{q}_k](u)\Phi_{\mu_b}[\delta_u](z_j).$$

注意到

$$\mu_b[f_{j,k}] = \int_{\mathbb{Z}} \overline{\mathcal{T}}[\bar{q}_k](u)\frac{\kappa(u,z_j)}{c_{\mu_b}(z_j)}\mu_b(du) = \overline{\mathcal{T}}_{\mu_b}[\bar{q}_k](z_j).$$

那么，根据 G_k 的定义,

$$G_k(z_j) = f_{j,k}(Z_k) - \mu_b[f_{j,k}]. \tag{B.10}$$

此外，利用 $\big\|\mathcal{T}[\bar{q}_k]\big\| \leq \frac{1}{1-\gamma}$ 和 $\|\Phi_{\mu_b}[\delta_u]\| \leq \kappa_\wedge^{-1}$，可得

$$\|f_{j,k}\| \leq \frac{1}{(1-\gamma)\kappa_\wedge}. \tag{B.11}$$

Let $v_{j,\ k}$： $\mathbb{Z}\to\mathbb{R}$ 为与 f_j 相关的泊松方程的解k。根据引理 9 ,

$$v_{j,k} = \sum_{t=0}^{\infty} P_b^t[f_{j,k} - \mu_b[f_{j,k}]] \tag{B.12}$$

其中该级数绝对收敛。由于 $f_{j,k}$ 是 $\mathcal{F}_k$-可测的，因此 $v_{j,k}$ 也 $\mathcal{F}_k$-可测。又由引理 9,

$$v_{j,k} - P_b v_{j,k} = f_{j,k} - \mu_b[f_{j,k}],$$

and

$$\begin{aligned}\|v_{j,k}\| &\le \frac{2c}{1-\rho}\|f_{j,k}\| \\ &\le \frac{2c}{1-\rho}\cdot\frac{1}{(1-\gamma)\kappa_\wedge} \\ &=: b_v.\end{aligned} \tag{B.13}$$

因此，由（B.10）以及 Z_k的马尔可夫性质，

$$\begin{aligned}G_k(z_j) &= v_{j,k}(Z_k) - P_b[v_{j,k}](Z_k) \\ &= v_{j,k}(Z_k) - E\big[v_{j,k}(Z_{k+1})|\mathcal{F}_k\big].\end{aligned}$$

定义鞅差 $\eta_{j,\ k+1} := v_{j,\ k}(Z_{k+1}) - E\left[v_{j,k}(Z_{k+1})|\mathcal{F}_k\right]$. 那么 $E[\eta_{j,k+1}|\mathcal{F}_k] = 0$, $|\eta_{j,\ k+1}| \le 2\|v_{j,\ k}\| \le 2b_v$, 且

$$G_k(z_j) = v_{j,k}(Z_k) - v_{j,k}(Z_{k+1}) + \eta_{j,k+1}. \tag{B.14}$$

因此，对于 $m \ge n$,

$$\begin{aligned}S^G_{n,m}(z_j) &= \sum_{k=n}^{m-1}\alpha_{k+1}G_k(z_j) \\ &= \underbrace{\sum_{k=n}^{m-1}\alpha_{k+1}\eta_{j,k+1}}_{=:M_{j,n,m}} + \underbrace{\sum_{k=n}^{m-1}\alpha_{k+1}\big(v_{j,k}(Z_k) - v_{j,k}(Z_{k+1})\big)}_{=:T_{j,n,m}}\end{aligned} \tag{B.15}$$

其中 $M_{j,\ n,\ m}$ 是鞅部分，而 $T_{j,\ n,\ m}$ 是望远镜求和部分。

鞅部分：注意，$M_{j,\ 0,\ m}$ 是一个 $\mathcal{F}_m$-鞅，且

$$E[M^2_{j,0,m}] = \sum_{k=0}^{m-1}\alpha^2_{k+1}E[\eta^2_{j,k+1}] \le 4b_v^2\sum_{k=0}^{\infty}\alpha^2_{k+1} < \infty.$$

因此，根据引理8

$$\max_{m\in\mathcal{N}(n,T)}|M_{j,n,m}| \le \sup_{m\ge n}|M_{j,0,m} - M_{j,0,n}| \to 0 \quad \text{a.s.}$$

望远镜部分：对于所有 $m \ge n$,

$$\begin{aligned}T_{j,n,m} &= \alpha_{n+1}v_{j,n}(Z_n) - \alpha_m v_{j,m-1}(Z_m) + \sum_{k=n+1}^{m-1}(\alpha_{k+1}-\alpha_k)v_{j,k}(Z_k) \\ &\quad + \sum_{k=n+1}^{m-1}\alpha_k(v_{j,k}(Z_k) - v_{j,k-1}(Z_k)).\end{aligned} \tag{B.16}$$

由于 $|v_{j,\ k}(Z_k)| \le \|v_{j,\ k}\| \le b_v$ 且 α_k 单调不增，因此（B.16）中的前三项之和有上界 $3b_v\alpha_{n+1}$。

对于最后一项求和，我们使用引理 $v_{j,\ k}$ 中的表示形式，见 9 和（B.12）

$$\begin{aligned}\|v_{j,k} - v_{j,k-1}\| &= \left\|\sum_{t=0}^{\infty}P_b^t[f_{j,k} - f_{j,k-1} - \mu_b[f_{j,k} - f_{j,k-1}]]\right\| \\ &\le \frac{4c}{1-\rho}\|f_{j,k} - f_{j,k-1}\|.\end{aligned}$$

此外，

$$\begin{aligned}\|f_{j,k} - f_{j,k-1}\| &\le \sup_{u\in\mathbb{Z}} |\Phi_{\mu_b}[\delta_u](z_j)| \left\|\overline{\mathcal{T}}[\bar{q}_k] - \overline{\mathcal{T}}[\bar{q}_{k-1}]\right\| \\ &\le \frac{\gamma}{\kappa_\wedge} \|\bar{q}_k - \bar{q}_{k-1}\|.\end{aligned}$$

由（4.3），$\bar{q}_k - \bar{q}_{k-1} = \alpha_k\big(Y_k\Phi_{\mu_b}[\delta_{Z_{k-1}}] - \bar{q}_{k-1}\big)$。因此，根据引理 2，

$$\|\bar{q}_k - \bar{q}_{k-1}\| \le \alpha_k\left(\frac{1}{(1-\gamma)\kappa_\wedge} + \frac{1}{(1-\gamma)\kappa_\wedge}\right) = \frac{2\alpha_k}{(1-\gamma)\kappa_\wedge}.$$

结合最后三个显示式可得

$$\|v_{j,k} - v_{j,k-1}\| \le \frac{8c\gamma}{(1-\rho)(1-\gamma)\kappa_\wedge^2}\,\alpha_k. \tag{B.17}$$

因此，对于所有 $m \ge n$，

$$|T_{j,n,m}| \le 3b_v\alpha_{n+1} + \frac{8c\gamma}{(1-\rho)(1-\gamma)\kappa_\wedge^2}\sum_{k=n+1}^{\infty}\alpha_k^2 \to 0.$$

当 $n\to\infty$ 时。特别地，处处有 $\max_{m\in\mathcal{N}(n,T)}|T_{j,\,n,\,m}| \to 0$。将其与鞅部分结合，并回顾（B.15），可得 $\max_{m\in\mathcal{N}(n,T)}\left|S^G_{n,\,m}(z_j)\right| \to 0$ 几乎必然成立。

采用（B.9）中的相同策略，我们将这一几乎必然收敛 $\max_{m\in\mathcal{N}(n,T)}\left|S^G_{n,\,m}(z_j)\right| \to 0$ 转换为

$$\limsup_{n\to\infty}\max_{m\in\mathcal{N}(n,T)}\left\|S^G_{n,m}\right\| = 0 \quad \text{w.p.1.} \tag{B.18}$$

步骤 5：得出命题。 结合（B.4）、（B.9）和（B.18），我们得到 in

$$\begin{aligned}&\max_{m\in\mathcal{N}(n,T)}\left\|\sum_{k=n}^{m-1}\alpha_{k+1}(D_{k+1}+G_k+B_k)\right\| \\ &\le \max_{m\in\mathcal{N}(n,T)}\left\|S^D_{n,m}\right\| + \max_{m\in\mathcal{N}(n,T)}\left\|S^G_{n,m}\right\| + \max_{m\in\mathcal{N}(n,T)}\left\|S^B_{n,m}\right\| \to 0\end{aligned}$$

几乎必然成立，当 $n\to\infty$，从而证明了（4.7）。 □

C 数值实验细节与其他测试

C.1 参数规格与可视化细节

我们选择折扣因子 $\gamma = 0.7$、带宽 $\sigma = 1$ 和初始状态 $X_0 = (0,0)$。

回顾一下，c、h、p、k 分别表示订货成本、持有成本、缺货损失成本和固定订货成本。对于第 6 节中介绍的实验，我们采用

$$c = (0.3, 0.35), \quad h = (0.05, 0.04), \quad p = (0.8, 0.9), \quad \text{and} \quad k = 0.2.$$

需求由 $D = |G|$ 给出，其中 $G \sim N(\mu,\Sigma)$，其中

$$\mu = (5,4), \quad \text{and} \quad \Sigma = \begin{bmatrix} 3 & -0.21 \\ -0.21 & 1 \end{bmatrix}.$$

不难看出，在这种设定下，需求呈负相关。

Q^*的基准近似记为 Q_{DP}，通过在离散化状态空间上进行动态规划计算得到。我们将区间 $[0, I_{\max} = 15]$ 均匀划分为 25 个子区间，从而得到 625 个正方形，每个正方形表示一个状态。动作是 $\mathbb{A} = \{0, 1, \ldots, A_{1,\max} = 10\} \times \{0, 1, \ldots, A_{2,\max} = 8\}$，因此 $|\mathbb{A}| = 99$。动态规划过程采用值迭代，在这些 $625 \times 99 = 61875$ 点上计算 Q_{DP}，然后将与某个状态对应的正方形内的值设为常数。

对于图 1，策略价值通过蒙特卡罗方法计算，即对256个价值估计取平均值；每个价值估计都通过模拟当前策略诱导的马尔可夫链构造

$$\pi_n(x) := \arg\max_{(a_1,a_2)\in\mathbb{A}} q_n(x,(a_1,a_2)) = \arg\max_{(a_1,a_2)\in\mathbb{A}} \Phi_{\mu_n}[\nu_n](x,(a_1,a_2)) \tag{C.1}$$

经过 $T = 200$ 步；即

$$\bar{V} = \frac{1}{256}\sum_{j=1}^{256} V_j; \qquad V_j := \sum_{t=0}^{T} \gamma^t R_{j,t+1}.$$

对于图 2，我们按照（ $n = 30000$）中的方法计算并绘制末次迭代策略（在 $n = 30000$），并将其与由 Q_{DP}诱导的贪心策略进行比较。图中分别展示了两种策略在（ a_1， a_2）平面上的订货量 x_1和 x_2。

第6节的其他诊断图

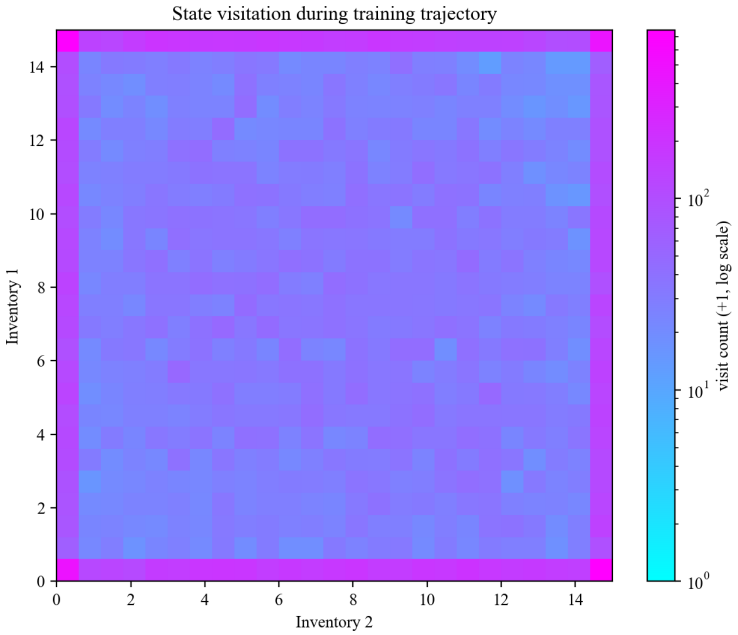

图3：行为策略下的访问次数：整个状态空间均得到了充分覆盖。

图 3 展示了在第 6节所述数值实验设置下库存状态的访问情况。我们观察到，均匀随机化的行为策略探索了整个状态空间。

图4 比较了两个具有代表性的离散动作 $q_n(\cdot, a)$对应的Q函数切片与基准 Q_{DP}， 分别为 $a = (0,2)$和 $a = (0,5)$。学习得到的曲面在定性上与基准吻合，但存在明显的全局价值偏移，约为 $+0.1$，其中 q_n 高估了 Q^*。这表明，图*1*中的均方根误差曲线所示约为 0.1的残差误差主要由这一全局偏移造成。

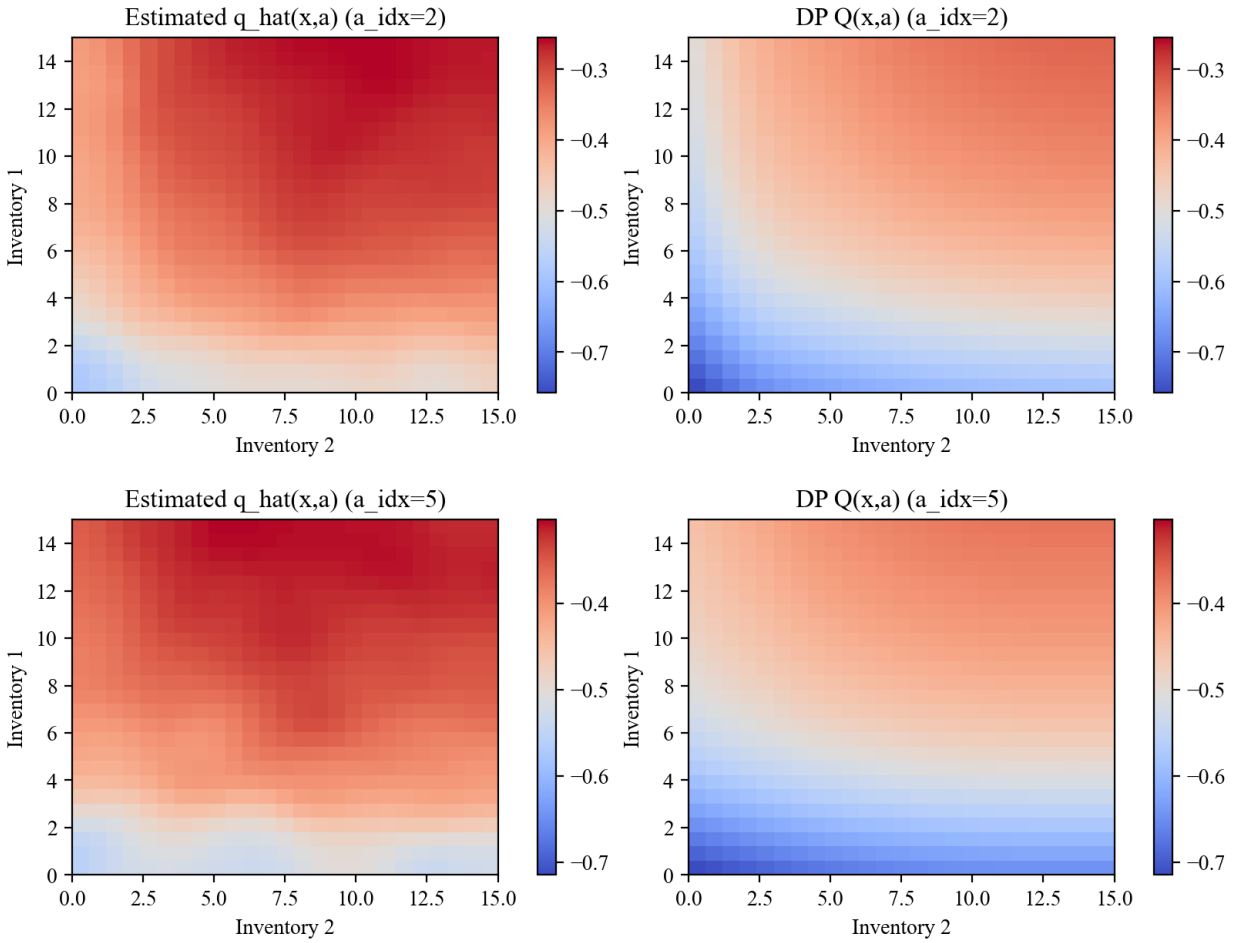

图 4：Q-测度估计 $q_n(\cdot,\ a)$ 与动态规划基准 $a = (0,\ 2)$ 和 $(0,\ 5$ $)$.

C.3 部分覆盖下的学习结果

另一方面，我们还在探索策略无法探索整个状态空间的情形下实现了该算法；见图5。在这种情况下，需求 $D = |G'|$，其中 $G' \sim N((8,7)\ \Sigma)$，与之前相同的 Σ 。因此，需求高于早先的实验设置。我们观察到，策略只探索状态空间的左下角。相应地， q_n 对Q^* 的估计在行为策略访问的区域仍然准确，但在行为策略无法到达的区域则变得不可靠。

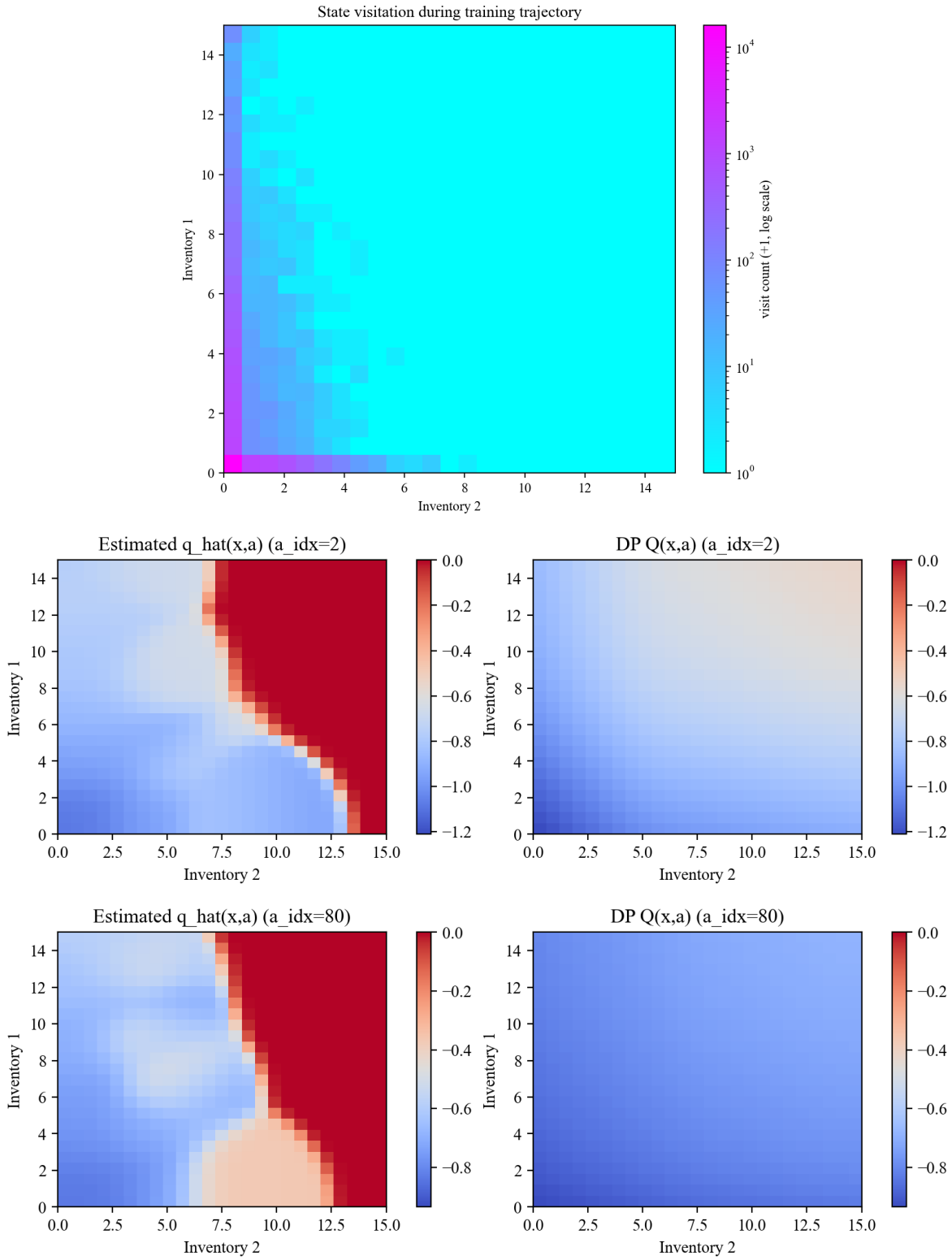

图5：库存状态的部分访问：行为策略仅探索状态空间的左下角。 Q^* 的Q测度估计在左下区域是准确的，但在右上区域存在偏差。