arXiv:2601.00151v1 [cs.LG] 1 Jan 2026

非马尔可夫过程的函数逼近强化学习

Ali Devran Kara *

2026年1月5日

摘要

我们研究非马尔可夫状态与成本过程下采用线性函数逼近的强化学习方法。我们首先考察策略评估方法，并证明在底层非马尔可夫过程满足适当的遍历性条件时，该算法能够收敛。此外，我们证明其极限对应于一个联合算子的不动点；该联合算子由一个正交投影和一个辅助马尔可夫决策过程的贝尔曼算子组成。

对于采用线性函数逼近的Q学习，与马尔可夫情形一样，一般无法保证收敛。不过，我们证明，在基函数根据量化映射选取这一特殊情形下，在类似的遍历性条件下可以证明其收敛。最后，我们将结果应用于部分可观测马尔可夫决策过程，其中有限记忆变量用作状态表示，并为所得学习算法的极限推导出显式误差界。

1 引言

无模型强化学习方法旨在直接从交互数据中计算近似最优的控制策略，或随机控制问题的价值函数，而无需构建动力学模型。尽管这些算法不需要显式了解动力学，但其理论保证依赖于这样一个假设：底层控制问题是马尔可夫决策过程（MDP）。在实践中，这一假设往往过于理想化，只有在模拟环境中才成立。

*作者就职于美国佛罗里达州塔拉哈西的佛罗里达州立大学数学系，电子邮箱：akara@fsu.edu

本文研究当观测到的状态和成本过程是一般随机过程、而非马尔可夫决策过程时的强化学习算法。我们重点关注采用线性函数逼近的方法，并分析其收敛性质，以及在收敛发生时极限的含义。

我们聚焦于线性函数逼近下的两种经典强化学习方法：策略评估和 Q 学习。线性函数逼近是处理高维状态空间最简单的方法之一，也是理论上最易于分析的情形，能够帮助我们理解函数逼近下学习算法的行为。

现有的收敛分析通常假设状态过程是马尔可夫的，且成本仅取决于当前状态和动作。在这些假设下，策略评估和 Q学习分别旨在所选基函数的张成空间内逼近给定策略的价值以及最优状态-动作价值函数。

当马尔可夫假设不成立时，这些迭代的表现并不显而易见。本文探讨的主要问题如下：

- 如果过程不是马尔可夫的，迭代是否会收敛？保证收敛所需的最小假设是什么？
- 如果迭代收敛，其极限代表什么？
- 极限值对我们感兴趣的量的逼近效果如何？特别是，能否得到明确的逼近误差界？

1.1 相关工作

随机控制问题的最优性分析与学习所面临的主要挑战之一是维数灾难。函数逼近方法被广泛用于解决这一问题。特别是，针对完全观测马尔可夫控制问题，采用线性函数逼近的强化学习已得到广泛研究。

[24] 是最早分析全可观测马尔可夫决策过程中的线性函数逼近策略评估的人之一，并证明了 TD(λ) 方法的收敛性。然而，在线性函数逼近下分析最优 Q 值的学习更具挑战性。尤其是，探索策略的不变测度可能不同于贪心策略所诱导的测度，因此算法通常可能无法收敛。 [17] 证明了

在协方差支配条件下，算法能够收敛；该条件关联了贪心策略和探索策略所诱导的特征协方差。这个条件表明，在一般情况下，探索策略不应偏离贪心动作选择太远。

还有一些特殊情形可以保证收敛。首先，在精确表示情形下，如果最优 Q 值位于所选基函数的张成空间中，就可以精确学习。在这种情况下，投影映射与贝尔曼算子的复合等同于贝尔曼算子本身，因此在一致范数下仍是压缩映射[20, 8]。其次，如果基函数是正交归一的（例如在基于离散化的近似中），那么投影映射不仅在 L_2 范数下，而且在一致范数下也是非扩张的，因此无需施加限制性条件即可进行收敛与误差分析 [12]。

对于一般的基函数，Meyn [18]最近证明，尽管投影算子与贝尔曼算子的复合不一定是压缩映射，但当探索策略为 ϵ-贪心策略时，至少存在一个不动点。此外，参数迭代几乎必然有界。

对于全可观测马尔可夫决策过程以外的情形，函数逼近仍相对缺乏研究。 [3] 研究了采用线性函数逼近学习部分可观测马尔可夫决策过程，并假设转移密度和观测密度都可以由基函数精确表示。他们考虑有限记忆变量，并对观测模型施加了严格的可观测性条件，以确保观测分布可逆，并使有限记忆变量的贝尔曼映射能够参数化。该条件保证，任意观测分布都能唯一确定隐状态分布。

已有少数研究探讨了非马尔可夫设置下的Q学习，例如 [5, 4, 10, 15, 23]。在这些近期研究之前，我们注意到，[22] 证明了在将测量值视为状态变量的部分可观测马尔可夫决策过程中Q学习的收敛性；这代表了一类特殊的非马尔可夫动力学。

[10] 分析了基于有限窗口测量值的部分可观测马尔可夫决策过程中的Q学习，并证明在滤波器稳定性条件下可达到近似最优性。类似地， [23]研究了基于历史函数的部分可观测马尔可夫决策过程中的Q学习，并证明其在一般学习率下收敛。

[5] 针对具有有限变量、允许依赖无限过去的复杂环境提出了一种通用强化学习框架，并假设在满足某些正则性条件时转移是平稳的。 [4]通过对无限维可观测历史施加连续性和可测性条件，分析了非马尔可夫环境中的Q学习收敛性，

采用了 [2]开创的基于常微分方程的方法。最后， [15]证明了在非马尔可夫状态过程满足遍历性假设时，表格型Q学习能够收敛，并指出学习得到的价值对应于一个辅助马尔可夫决策过程，因此可以将学习得到的控制策略与最优价值进行比较。

本文将这些结果扩展到遍历性条件下的一般非马尔可夫状态与成本过程的线性函数逼近。我们研究使用线性函数逼近进行的策略评估和Q学习。对于策略评估，我们证明，在遍历性假设下，算法可以收敛。作为一个特例，我们考虑采用有限记忆控制器的部分可观测控制问题。基于 [11, 14] 中提出的有限记忆逼近框架，我们给出学习得到的价值的上界。对于采用线性函数逼近的Q学习，通常无法保证收敛。然而，在离散化的情况下，该算法可简化为对离散化后的非马尔可夫状态过程进行表格型Q学习，因此我们可以应用 [15] 中的结果。此外，对于使用基于离散化的基函数的部分可观测马尔可夫决策过程， [13] 的误差分析在对模型和探索策略限制较少的假设下仍然适用。

1.2 问题表述

我们考虑三个随机过程:

- S_t 是一个取值于 S 的随机过程，表示状态,
- C_t 是一个实值过程，表示成本实现,
- $U_t \sim \gamma(\cdot|S_t)$ 是由某个随机反馈控制函数 γ :S $\to \mathcal{P}$(U) 生成的控制过程。

这里，S $\subset \mathrm{R}^n$ 和 U $\subset \mathrm{R}^m$ 是博雷尔空间，其中 m， $n < \infty$ 为某个有限值。所有过程都定义在滤过概率空间 (Ω $\mathcal{F}$， $\{\mathcal{F}_t\}_{t\geq 0}$， $\mathbb{P}$) 上，并且适应于该滤过。

我们研究应用于这些过程的两种强化学习算法：线性函数逼近下的策略评估（TD(0)）和 Q学习。设 $\{\phi^i(s)\}_{i=1}^d$, ϕ^i : S $\to$R 为一组已知基函数，并记 $\mathbf{\Phi}^\intercal := [\phi^1, \ldots, \phi^d]$。策略评估跟踪参数 $\{\theta_t\}$，其定义如下

$$\theta_{t+1} = \theta_t - \alpha_t \mathbf{\Phi}(S_t) \left[\theta_t^\intercal \mathbf{\Phi}(S_t) - C_t - \beta \theta_t^\intercal \mathbf{\Phi}(S_{t+1})\right], \tag{1}$$

其中，$0 < \beta < 1$ 为折扣因子，α_t 为学习率。

对于 Q学习，基函数扩展到动作空间：$\{\phi^i(s,\ u)\}_{i=1}^d$, $\phi^i : \mathsf{S} \times \mathsf{U} \to \mathsf{R}$，并按如下方式更新参数

$$\theta_{t+1} = \theta_t - \alpha_t \mathbf{\Phi}(S_t, U_t)\Big[\theta_t^\intercal \mathbf{\Phi}(S_t, U_t) - C_t - \beta \min_v \theta_t^\intercal \mathbf{\Phi}(S_{t+1}, v)\Big]. \tag{2}$$

在标准的马尔可夫设置中，状态按照 $S_{t+1} \sim \mathcal{T}(\cdot|S_t,\ U_t)$ 演化，其中转移核为 $\mathcal{T}$；成本仅取决于当前状态和动作：对于某个 $c : \mathsf{S} \times \mathsf{U} \to \mathsf{R}$，有 $C_t = c(S_t,\ U_t)$。对于标准的马尔可夫设置，算法随后旨在近似

$$\sum_{t=0}^{\infty} \beta^t \mathbb{E}_\gamma[c(S_t, U_t)|S_0 = s_0], \quad \inf_\gamma \sum_{t=0}^{\infty} \beta^t \mathbb{E}_\gamma[c(S_t, U_t)|S_0 = s_0, U_0 = u_0],$$

在基函数的张成空间上 $\{\phi^i\}$，其中期望是相对于转移核 $\mathcal{T}$ 和策略 $\gamma(du|s)$ 计算的。上面的第一项表示策略 γ 下累积的无限时域期望折扣成本，我们称之为策略价值 γ。第二项表示在初始状态和动作对为某个 $(s_0,\ u_0)$ 时能够达到的最优价值，也称为 $(s_0,\ u_0)$ 的最优 Q 值或状态-动作价值函数。

在本文中，我们假设过程 S_t、C_t、U_t 不一定遵循标准的马尔可夫设定。我们研究保证迭代（1）和（2）在马尔可夫情形之外收敛的充分条件，并刻画收敛时的极限。我们的主要贡献如下：

- **策略评估（第 2 节）：**我们分析迭代（1）的收敛性，并刻画其极限。

  – 在第 2.1 节中，我们证明了一种用于求解线性方程的随机逼近算法在非马尔可夫噪声下的收敛性。我们通过利用泊松方程分解噪声，扩展了 [1] 的论证，并通过采用适当的遍历性和混合性假设，将这些论证推广到非马尔可夫过程。

  – 在第 2.2 节中，我们构造了一个辅助马尔可夫决策过程，称为 平稳状态 *MDP*，它对应于非马尔可夫状态过程的平稳行为 S_t。

– 在第 2.3 节和 2.4 节中，我们为基函数 $\{\phi^i\}_{i=1}^d$ 定义了正交投影映射，并为平稳状态 MDP 定义了贝尔曼映射。利用随机逼近结果，我们证明了（1）收敛，且其极限与由平稳状态 MDP 的投影映射和贝尔曼映射组成的联合映射的不动点一致。特别地，这意味着，在非马尔可夫过程中进行的迭代会收敛到与在由平稳状态 MDP 生成的马尔可夫过程中进行迭代时相同的极限。

– 在第 2.5 节中，我们分析了在平稳状态 MDP 下，学习得到的价值相对于策略 $\gamma(du|s)$ 的价值的误差。

• **Q学习（第 3 节）：** 我们研究了贪心动作选择下，平稳状态 MDP 的投影贝尔曼算子的行为。与标准 MDP 中的情况一样（这并不令人意外），在非马尔可夫过程中，采用线性函数逼近的 Q学习通常无法收敛，除非满足以下特殊情况之一：(i) 平稳状态 MDP 的成本函数和转移核在所选基函数上完全呈线性；(ii) 折扣后，贪心策略诱导的特征协方差一致地受探索策略诱导的特征协方差支配；或 (iii) 基函数是通过对 $\mathbb{S}$ 和 $\mathbb{U}$ 进行离散化并基于示性函数构造的。

• **部分可观测 MDP（第 4 节）：** 我们将该框架应用于具有有限记忆控制器的部分可观测马尔可夫决策过程。针对有限记忆下的策略评估，我们推导出学习得到的价值的显式误差界，并将误差分解为投影项和有限记忆逼近项；后者与底层系统的滤波器稳定性有关。对于采用有限记忆变量的 Q学习，我们考虑基于离散化的基函数，并为这种情形提供收敛结果和误差分析。

注释1. 在全文中，$K < \infty$ 表示一个一般常数。其值可能在不同步骤中有所不同，但在每一步中，对于给定上下文中的其他变量（例如时间 t 或随机变量）而言，它都是统一的。

2 非马尔可夫过程的策略评估

2.1 遍历性条件下非马尔可夫过程的随机逼近结果

我们定义联合过程 $Z_t := (S_{t+1}, S_t, C_t, U_t)$。我们首先给出主要结果所需的假设。

假设 1. *i.* 对任意有界函数 f，我们有

$$\frac{1}{N}\sum_{t=1}^{N} f(Z_t) \to \int f(z)\pi(dz) \quad a.s.$$

几乎必然地，对某个概率测度 $\pi \in \mathcal{P}(\mathbb{S}^2 \times \mathbb{R} \times \mathbb{U})$成立。

ii. 对于矩阵值函数 $A(Z_t)$ 和向量值函数$b(Z_t)$，定义

$$Y_t^A := \sum_{k=0}^{\infty} \|E[A(Z_{t+k})|\mathcal{F}_t] - A\|, \quad Y_t^b := \sum_{k=0}^{\infty} \|E[b(Z_{t+k})|\mathcal{F}_t] - b\|, \tag{3}$$

其中 $A := \int A(z)\pi(dz)$ 且 $b := \int b(z)\pi(dz)$，并且矩阵采用谱范数。我们假设这些序列在 L_2 中一致有界：$\sup_t \|Y_t^A\|_2 < \infty$ 和 $\sup_t \|Y_t^b\|_2 < \infty$。

iii. $A(Z_t)$ 和 $b(Z_t)$ 是有界函数。

注 2. 若 Z_t 是严格平稳的，则对所有 t，$\|Y_t^A\|_2 = \|Y_0^A\|_2$ 和 $\|Y_t^b\|_2 = \|Y_0^b\|_2$。若没有平稳性，$L_2$ 的有界性仍可扩展到所有 t，如下所示。

在非平稳情况下，使假设1 成立的一个充分条件是强混合系数可求和。对于两个子σ-代数 $\mathcal{A}$，$\mathcal{B} \subset \mathcal{F}$，定义

$$\alpha(\mathcal{A}, \mathcal{B}) := \sup_{A\in\mathcal{A}, B\in\mathcal{B}} |P(A \cap B) - P(A)P(B)|. \tag{4}$$

令 $\mathcal{F}_j^-$ 表示由 $\{Z_t$ 生成的 $\sigma-$ 代数：$t \le j\}$。类似地，令 $\mathcal{F}_j^+$ 表示由 $\{Z_t$ 生成的 $\sigma-$ 代数：$t \ge j\}$。回顾一下，过程 $\{Z_t\}$ 的强混合系数定义为

$$\alpha(k) := \sup_j \alpha(\mathcal{F}_j^-, \mathcal{F}_{j+k}^+). \tag{5}$$

假设 2. *i.* 随机变量 Y_0^A和 Y_0^b在（3）中定义，满足

$$\|Y_0^A\|_2 < \infty \quad and \quad \|Y_0^b\|_2 < \infty.$$

ii. 混合系数 $\alpha(k)$ 定义于 *(5)* 中，满足

$$\sum_{k=0}^{\infty} \alpha(k)^{1/2} < \infty.$$

引理 1. 假设 2 蕴含假设 1 *(ii)*。也就是说，如果 $\|Y_0^A\|_2 <\infty$, $\|Y_0^b\|_2 < \infty$, 且 $\sum_{k=0}^{\infty} \sqrt{\alpha(k)} < \infty$，那么序列 $\{Y_t^A\}$ 和 $\{Y_t^b\}$对所有 t均一致有界：

$$\sup_t \|Y_t^A\|_2 < \infty, \quad \sup_t \|Y_t^b\|_2 < \infty.$$

证明。 我们仅证明 Y_t^A 的情形。记

$$e_{t+k} := \|E[A(Z_{t+k})|\mathcal{F}_t] - A\|$$

我们从以下直接得到的界开始：

$$\|Y_t^A\|_2 \leq \sum_{k=0}^{\infty} \|e_{t+k}\|_2.$$

下文将用到以下关系：对于一个 $d \times d$ 矩阵 A，有

$$\|A\| \leq \|A\|_F \leq \sqrt{d}\|A\|$$

其中， $\|A\|_F$ 表示弗罗贝尼乌斯范数。对于 e_{t+k}，将矩阵 $A(Z_{t+k})$ 的第 ij 个元素记为 $\tilde{A}^{ij}(Z_{t+k}) - A$ ，我们可以写成

$$\begin{aligned}
\|e_{t+k}\|_2^2 &= E\left[\|E[A(Z_{t+k}) - A|\mathcal{F}_t]\|^2\right] \leq E\left[\sum_{i,j} E[\tilde{A}^{ij}(Z_{t+k})|\mathcal{F}_t]^2\right] \\
&= \sum_{i,j} E\left[E[\tilde{A}^{ij}(Z_{t+k})|\mathcal{F}_t]^2\right] = \sum_{i,j} E\left[\tilde{A}^{ij}(Z_{t+k})E[\tilde{A}^{ij}(Z_{t+k})|\mathcal{F}_t]\right] \\
&= \sum_{i,j} cov\left(\tilde{A}^{ij}(Z_{t+k}), E[\tilde{A}^{ij}(Z_{t+k})|\mathcal{F}_t]\right) + \sum_{i,j} E\left[\tilde{A}^{ij}(Z_{t+k})\right]^2 \\
&= \sum_{i,j} cov\left(\tilde{A}^{ij}(Z_{t+k}), E[\tilde{A}^{ij}(Z_{t+k})|\mathcal{F}_t]\right) + \|E[\tilde{A}(Z_{t+k})]\|_F^2
\end{aligned}$$

这是一个标准结果（例如参见 [19]），对于任意有界的 f

$$Cov(f(Z_{t+k}), E[f(Z_{t+k}|\mathcal{F}_t)]) \leq 4\alpha(k)\|f\|_\infty^2.$$

利用 $A(z)$ 的有界性，我们便可写出：对某个 $K < \infty$，

$$\begin{aligned}
\|Y_t^A\|_2 &\leq \sum_{k=0}^{\infty} \|e_{t+k}\|_2 \leq \sum_{k=0}^{\infty} \sqrt{K\alpha(k)} + \sum_{k=0}^{\infty} \|E[\tilde{A}(Z_{t+k})]\|_F \\
&\leq \sum_{k=0}^{\infty} \sqrt{K\alpha(k)} + \sum_{k=0}^{\infty} \left\| E\left[E[\tilde{A}(Z_k)|\mathcal{F}_0]\right] \right\|_F \\
&\leq \sum_{k=0}^{\infty} \sqrt{K\alpha(k)} + E\left[\sum_{k=0}^{\infty} \left\| E[\tilde{A}(Z_k)|\mathcal{F}_0] \right\|_F\right] \\
&\leq \sum_{k=0}^{\infty} \sqrt{K\alpha(k)} + \sqrt{d}E[Y_0^A] \leq \sum_{k=0}^{\infty} \sqrt{K\alpha(k)} + \sqrt{d}\|Y_0^A\|_2 < \infty.
\end{aligned}$$

□

下面的命题是非马尔可夫过程中策略评估算法收敛的关键结果。主要技术工具引理 2 和 3主要基于 [1]。

具体而言，收敛性证明的主要挑战在于更新中包含的误差项：

$$\delta^\intercal \big[A - A(Z_t)\big]\theta + \delta^\intercal \big[b(Z_t) - b\big].$$

在 [1], 中，这一项针对马尔可夫过程 Z_t 进行了分析，并利用该马尔可夫过程在适当遍历性条件下满足的泊松方程，将其分解为鞅差项和可求和的望远镜项。

对于我们的关键技术工具（引理 2和 3），我们采用类似的策略。具体而言，我们证明，在假设 ***1***下，本例中的非马尔可夫误差项也满足泊松方程，随后可将其分解为鞅差项和可求和的望远镜误差项。尽管整体方法遵循与 [1], 类似的步骤，但推广到非马尔可夫过程并非易事。必须谨慎修订原有分析，例如验证遍历性条件、控制误差项以及处理条件期望。因此，尽管分解思路相似，非马尔可夫情形仍带来了重大的技术挑战，需要采用针对性的方法。

命题 1。 假设 *1* 成立（或者假设 *5*作为假设 *1*的充分条件），且平稳平均矩阵 A为正定矩阵。考虑如下随机逼近迭代：

$$\theta_{t+1} = \theta_t + \alpha_t\big(-A(Z_t)\theta_t + b(Z_t)\big),$$

其中， $A(Z_t)$ 和 $b(Z_t)$ 分别是矩阵值函数和向量值函数。于是， θ_t 几乎必然收敛到极限 θ^*，且该极限满足 $A\theta^* = b$，其中

$$A = E[A(Z)] = \int A(z)\,\pi(dz), \quad b = E[b(Z)] = \int b(z)\,\pi(dz),$$

并且 π 是联合过程 Z_t $=\{S_{t+1}, S_t, C_t, U_t\}$的平稳分布。

证明。 我们首先加上再减去 A 和 b，并注意到 $b = A\theta^*$:

$$\theta_{t+1} = \theta_t + \alpha_t\left(-A(Z_t)\theta_t + A\theta_t + b(Z_t) - b - A\theta_t + A\theta^*\right).$$

定义 $\delta_t := \theta_t - \theta^*$ 和 $M_t := (A - A(Z_t))\theta_t + b(Z_t) - b$，并从两边各减去$\theta^*$，得到

$$\delta_{t+1} = \delta_t + \alpha_t\left(-A\delta_t + M_t\right).$$

将两边平方，得到

$$\begin{aligned}\|\delta_{t+1}\|^2 &= \|\delta_t\|^2 + 2\alpha_t\delta_t^\intercal\left[-A\delta_t + M_t\right] + \alpha_t^2\| - A\delta_t + M_t\|^2 \\ &\le \|\delta_t\|^2 - 2\alpha_t\sigma_{\min}\|\delta_t\|^2 + 2\alpha_t\delta_t^\intercal M_t + \alpha_t^2 2\sigma_{\max}\|\delta_t\|^2 + \alpha_t^2 2\|M_t\|^2\end{aligned} \tag{6}$$

其中， $\sigma_{\min}$ 和 $\sigma_{\max}$ 分别表示 A的最小和最大特征值，并且这里用到了界 $(a+b)^2 \le 2a^2 + 2b^2$。利用假设 $A(Z_t)$ $b(Z_t)$一致有界，我们可得到 $\|M_t\|$的如下上界：

$$\begin{aligned}\|M_t\| &= \|(A - A(Z_t))\theta_t + b(Z_t) - b\| \\ &\le \|(A - A(Z_t))\|\|\theta_t\| + \|b(Z_t) - b\| \\ &\le K(\|\delta_t\| + 1)\end{aligned}$$

对某个 $K < \infty$。于是我们还可以得到，对于某个一般常数 $K < \infty$, $\|M_t\|^2 \le K(\|\delta_t\|^2 + 1)$。利用这一点，我们得到

$$\begin{aligned}\|\delta_{t+1}\|^2 &\le \|\delta_t\|^2 - 2\alpha_t\sigma_{\min}\|\delta_t\|^2 + 2\alpha_t\delta_t^\intercal M_t + \alpha_t^2 K\|\delta_t\|^2 + \alpha_t^2 K \\ &= (1 + K\alpha_t^2)\|\delta_t\|^2 - 2\alpha_t\sigma_{\min}\|\delta_t\|^2 + 2\alpha_t\delta_t^\intercal M_t + \alpha_t^2 K. \end{aligned} \tag{7}$$

我们注意到，Robbins-Siegmund 引理不能直接应用，因为无法保证 $2\alpha_t\delta_t^\intercal M_t$ 非负。尽管如此，我们仍可通过其他论证证明收敛。首先，我们引入以下停止时间：

$$\sigma_n := \inf\{t : \|\delta_t\|^2 > 2^n\}. \tag{8}$$

引理 2。 在假设 *1*下，我们有 $\sum_{t=0}^{k} \mathbb{1}_{\{t+1\leq\sigma_n\}} 2\alpha_t\delta_t^\intercal M_t$ 几乎必然收敛。特别地，$\sum_{t=0}^{k} 2\alpha_t\delta_t^\intercal M_t$ 在事件 $\{\sigma_n = \infty\}$上几乎必然收敛。

证明。 证明见附录A。 □

引理 3。 我们定义停时

$$\sigma(C) := \inf\{t : \|\delta_t\|^2 > C\}.$$

在假设 *1*下，对于任意 $n < \infty$,

$$E\left[\sup_{k>n} \mathbb{1}_{\{k+1\leq\sigma(C)\}} \left(\sum_{t=n}^{k} 2\alpha_t\delta_t^\intercal M_t\right)^2\right] \leq K(1+C^2)\sum_{t=n}^{\infty}\alpha_t^2$$

对于某个常数 $K < \infty$。

证明。 证明见附录B。 □

将 $\mathbb{1}_{\{t+1\leq\sigma_n\}}$ 乘以（7）两边，并注意到 $\mathbb{1}_{\{t+2\leq\sigma_n\}} \leq \hat{} \mathbb{1}_{\{t+1\leq\sigma_n\}}$ 并记 $\mathbb{1}_{\{t+1\leq\sigma_n\}}\delta_t =: \hat{\delta}_t$

$$\begin{aligned}\|\hat{\delta}_{t+1}\|^2 &\leq (1+K\alpha_t^2)\|\hat{\delta}_t\|^2 - 2\alpha_t\sigma_{\min}\|\hat{\delta}_t\|^2 + 2\alpha_t\hat{\delta}_t^\intercal M_t + \alpha_t^2 K \\ &\leq (1+K\alpha_t^2)\|\hat{\delta}_t\|^2 + 2\alpha_t\hat{\delta}_t^\intercal M_t + \alpha_t^2 K \end{aligned} \tag{9}$$

接下来，我们定义

$$X_t := \frac{\|\hat{\delta}_t\|^2}{\prod_{i=1}^{t-1}(1+K\alpha_i^2)}.$$

然后我们观察到

$$X_{t+1} \leq \frac{\|\hat{\delta}_t\|^2}{\prod_{i=1}^{t-1}(1+K\alpha_i^2)} + \frac{2\alpha_t\hat{\delta}_t^\intercal M_t}{\prod_{i=1}^{t}(1+K\alpha_i^2)} + \frac{K\alpha_t^2}{\prod_{i=1}^{t}(1+K\alpha_i^2)}$$

现在引入以下记号:

$$a_t := \frac{2\alpha_t \hat{\delta}_t^\intercal M_t}{\prod_{i=1}^{t}(1+K\alpha_i^2)} + \frac{K\alpha_t^2}{\prod_{i=1}^{t}(1+K\alpha_i^2)}$$

这意味着 $X_{t+1} \leq X_t + a_t$. W 定义为 $U_t := X_t - \sum_{i=1}^{t-1} a_i$. 采用此记号，我们写作

$$\begin{aligned} E[U_{t+1}|\mathcal{F}_t] &= E[X_{t+1}|\mathcal{F}_t] - E[\sum_{i=1}^{t} a_i|\mathcal{F}_t] \\ &= E[X_{t+1}|\mathcal{F}_t] - \sum_{i=1}^{t} a_i \\ &\leq X_t + a_t - \sum_{i=1}^{t} a_i \\ &= X_t - \sum_{i=1}^{t-1} a_i = U_t. \end{aligned} \tag{10}$$

利用引理 3 的证明，我们可以证明

$$\begin{aligned} E[(\sum_{i=1}^{t-1} a_i)^2] &\leq E\left[(\sum_{i=1}^{t-1} 4\alpha_i \hat{\delta}_i^\intercal M_i)^2\right] + \left(\sum_{i=1}^{t-1} 2K\alpha_i^2\right)^2 \\ &\leq K(1+2^{2n})\sum_{i=1}^{\infty}\alpha_i^2 + \left(\sum_{i=1}^{\infty} 2K\alpha_i^2\right)^2 < \infty. \end{aligned}$$

此外，我们有

$$E[X_t^2] \leq E[\|\hat{\delta}_t\|^2] \leq 2^{2n}.$$

综上可知 $\sup_t E[U_t^2] < \infty$. 然后，结合（10），我们可以得出 U_t 是一个 L_2 范数一致有界的超鞅，因此 U_t 几乎必然收敛。此外，利用关于学习率的假设和引理 2，我们还知道 $\sum_{i=1}^{t-1} a_i$ 几乎必然收敛。由此可得 $X_t = U_t + \sum_{i=1}^{t-1} a_i$ 几乎必然收敛。由于根据关于学习率的假设，$\prod_{i=1}^{t-1}(1+K\alpha_i^2)$ 也收敛，因此 $\|\hat{}\delta_t\|^2$ 几乎必然收敛。

回到（9），并重新排列各项，可得

$$2\alpha_t \sigma_{\min}\|\hat{\delta}_t\|^2 \leq \|\hat{\delta}_t\|^2 - \|\hat{\delta}_{t+1}\|^2 + K\alpha_t^2\|\hat{\delta}_t\|^2 + 2\alpha_t\hat{\delta}_t^\intercal M_t + \alpha_t^2 K$$

注意到 $\|\hat{\delta}_t\|^2 \leq 2^{2n}$，将等式两边求和，得到：

$$\sum_{t=0}^{k} 2\alpha_t \sigma_{\min} \|\hat{\delta}_t\|^2 \leq \|\hat{\delta}_0\|^2 - \|\hat{\delta}_{k+1}\|^2 + \sum_{t=0}^{k} \left(K\alpha_t^2 \|\hat{\delta}_t\|^2 + 2\alpha_t \hat{\delta}_t^\intercal M_t + \alpha_t^2 K \right)$$
$$\leq \|\hat{\delta}_0\|^2 + \sum_{t=0}^{k} \alpha_t^2 (1 + 2^{2n}) + \sum_{t=0}^{k} 2\alpha_t \hat{\delta}_t^\intercal M_t.$$

使用引理2 以及关于学习率的条件，右侧的所有项都几乎必然收敛。因此，我们有

$$\sum_{t=0}^{k} 2\alpha_t \sigma_{\min} \|\hat{\delta}_t\|^2 < \infty$$

几乎必然成立，这意味着 $\liminf_k \|\hat{\delta}_t\|^2 \to 0$。由于我们此前已经证明 $\|\hat{\delta}_t\|^2$ 几乎必然收敛，该极限必为 0，即 $\|\hat{\delta}_t\|^2 \to 0$ 几乎必然成立。特别地，在事件 $\{\sigma_n = \infty\}$ 上，$\|\delta_t\|^2 \to 0$ 几乎必然成立。

将 [1, 定理 17] 的论证借助引理 3推广到非马尔可夫过程中，我们可以证明 $P(\{\sigma_n = \infty\}) \to 1$。为完整起见，我们在附录C 中给出了完整证明。

引理 4。 在假设 *1*下，$P(\{\sigma_n = \infty\}) \to 1$。

引理4，随后得出证明。特别地，记 A 为事件 $\|\delta_t\| \to 0$，并记 E_n 为事件 $\{\sigma_n = \infty\}$，则有 $P(E_n \cap A^c) = 0$ 对所有 n均成立。我们可以写成

$$P((\cup_{n=1}^{\infty} E_n) \cap A^c) = P(\cup_{n=1}^{\infty} (E_n \cap A^c)) = \lim_n P(E_n \cap A^c) = 0.$$

因此，再结合 $P(\cup_n E_n) = 1$ 这一事实，我们得到 $P(A^c) = 0$。

□

2.2 平稳 RegimeMDP

回顾联合过程 $Z_t := \{S_{t+1}, S_t, C_t, U_t\}$，其中 S_t 是表示状态过程的随机过程，C_t 是表示成本实现的另一个过程，而 $U_t \sim \gamma(\cdot|S_k)$ 是由某个策略 $\gamma : \mathsf{S} \to \mathcal{P}(\mathsf{U})$ 生成的控制过程。

考虑该过程的不变分布 π Z_t $:= \{S_{t+1}, S_t, C_t, U_t\}$ 在假设 1 下。现在我们定义一个马尔可夫决策

过程，以描述平稳情形。成本函数和转移核基于关于平稳测度 $\pi(\cdot)$ 的正则条件分布定义，使得

$$\begin{aligned} c(s,u) &:= E^{\pi}[C|s,u] \quad \forall s,u \in \mathbb{S}\times\mathbb{U} \\ \eta(s_1 \in A|s,u) &:= E^{\pi}\left[\mathbb{1}_{\{S_1\in A\}}|s,u\right] \quad \forall s,u \in \mathbb{S}\times\mathbb{U} \end{aligned} \tag{11}$$

其中，期望是关于 π on$\{S_{t+1},S_t,C_t,U_t\}$ 上的平稳分布而言的。请注意，该 MDP 的成本函数和转移模型取决于平稳分布，进而取决于策略 γ，而该策略会导出特定的平稳测度。为简洁起见，我们在记号中省略这一依赖关系。

我们为该平稳情形 MDP 在策略 γ 下定义如下贝尔曼算子，使得对于 $f \in L_2(\pi,\mathrm{S})$，我们记为

$$T^{\gamma}f(s) := \int_{\mathbb{U}} \left(c(s,u) + \beta \int f(s_1)\eta(ds_1|s,u) \right) \gamma(du|s). \tag{12}$$

类似地，对于 $g \in L_2(\pi,\mathrm{S}\times\mathrm{U})$，我们记为

$$Tg(s,u) := c(s,u) + \beta \int g^{-}(s_1)\eta(ds_1|s,u) \tag{13}$$

其中 $g^{-}(s) :=$ 下确界$_u\, g(s,u)$。

我们将该 MDP 在策略 γ 下的价值函数定义为

$$J_{\beta}^{\pi}(s_0,\gamma) := \sum_{t=0}^{\infty} \beta^t E^{\gamma}[c(\bar{S}_t,\bar{U}_t)|\bar{S}_0 = s_0]$$

其中 $\bar{S}_t$ 表示具有转移核 的马尔可夫过程 $\eta(ds_1|s,u)$，该过程定义于（11）中，其中 $U_t \sim \gamma(\cdot|S_t)$。我们加上横线记号，以区别于原始的非马尔可夫过程 S_t。

2.3 线性函数逼近与投影

我们考虑在通常的内积下，定义于 $s \in \mathrm{S}$ 上、以测度$\pi \in \mathcal{P}(\mathrm{S})$ 为基础的实值函数 L_2 空间。本节中的构造适用于任意测度 $\pi(\cdot)$；不过，π 主要指该过程的平稳测度，尤其是其在 S_t上的边缘测度。

我们引入一组基函数 $\{\phi^i(s)\}_{i=1}^{d}$，其中 $\phi^i(s) : \mathrm{S} \to \mathrm{R}$。我们用 $\mathbf{\Phi}^{\intercal} := [\phi^1,\ldots,\phi^d]$ 表示基函数向量。

假设 3. 在本文余下部分，我们假设对于所有$i = 1$，.....，d，$\|\phi^i\|_\infty \leq 1$ 。

假设 4. 在本文余下部分，我们假设 $\{\phi^i(s)\}$ 在线性意义下独立，且 $L_2(\pi)$ 使得 $E\left[\Phi(S)\Phi^\intercal(S)\right]$ 可逆。

我们将从 $L_2(\pi,\mathbb{S})$ 到 $\mathbf{\Phi}^\intercal := [\phi^1, \ldots, \phi^d]$ 的张成空间的投影映射记为 Π 。特别地，对于某个 $f \in L_2(\pi,\mathbb{S})$，有 $\Pi(f) = \theta_f^\intercal \mathbf{\Phi}$，其中

$$\theta_f = \arg\min_{\theta \in \mathbb{R}^d} \sqrt{\int_{\mathbb{S}} \left|f(s) - \theta^\intercal \mathbf{\Phi}(s)\right|^2 \pi(ds)}. \tag{14}$$

命题 2. 映射 ΠT^γ 在 L_2 范数下是压缩映射，因此存在唯一的不动点。

证明. 对于 $f, g \in L_2(\pi, \mathbb{S})$，我们有

$$\|\Pi T^\gamma(f) - \Pi T^\gamma(g)\|_2 \leq \|T^\gamma(f) - T^\gamma(g)\|_2$$

由于投影是非扩张的。利用詹森不等式，我们进而得到:

$$\begin{aligned}
&\|T^\gamma(f) - T^\gamma(g)\|_2 \\
&\leq \beta \sqrt{\int \left(f(s_1) - g(s_1)\right)^2 \eta(ds_1|s,u)\gamma(du|s)\pi(ds)} \\
&= \beta \sqrt{\int \left(f(s_1) - g(s_1)\right)^2 \pi(ds_1)} = \beta \|f - g\|_2.
\end{aligned}$$

上文利用了这样一个事实：根据构造，$\eta(ds_1|s,u)\gamma(du|s)\pi(ds) = \pi(ds_1, du, ds)$，因为 η 是基于联合过程的平稳分布得到的正则条件分布。此外，平稳分布在连续状态变量 S_t 和 S_{t+1} 上的边缘分布相同，这就说明了最后一步的合理性，从而完成证明。 □

2.4 策略评估算法的收敛

我们考虑以下算法

$$\theta_{t+1} = \theta_t - \alpha_t \mathbf{\Phi}(S_t)\left[\theta_t^\intercal \mathbf{\Phi}(S_t) - C_t - \beta \theta_t^\intercal \mathbf{\Phi}(S_{t+1})\right] \tag{15}$$

其中 α_t 表示学习率，并且我们使用策略 γ下的一条 $\{S_t, U_t, C_t\}_t$ 轨迹。

定理 1. 在假设*1* 和 *4*成立时，若学习率满足$\sum_t \alpha_t = \infty$ 且 $\sum_t \alpha_t^2 < \infty$，则（*15*）中的迭代收敛到某个$\theta^* \in \mathbb{R}^d$。记 $V(s) := \theta^{*\intercal}\mathbf{\Phi}(s)$， $V(s)$ 是联合映射 ΠT^γ的不动点，其中映射 Π和 T^γ分别定义于（*14*）和（*12*）中。

证明。 我们使用命题1 以及

$$A(S_t, S_{t+1}) = -\beta\mathbf{\Phi}(S_t)\mathbf{\Phi}^\intercal(S_{t+1}) + \mathbf{\Phi}(S_t)\mathbf{\Phi}^\intercal(S_t)$$
$$b(S_t, C_t) = \mathbf{\Phi}(S_t)C_t.$$

矩阵 A 和 b 是在联合过程 $(S_t,\ S_{t+1},\ U_t,\ C_t)$的不变测度 π 下定义的。

我们需要证明矩阵 A 是正定的。

引理 5。

$$(\theta - \theta^*)^\intercal E\left[\mathbf{\Phi}(S)\left[C + \beta\theta^\intercal\mathbf{\Phi}(S_1) - \theta^\intercal\mathbf{\Phi}(S)\right]\right] < 0$$

对于任意 $\theta \neq \theta^*$ 其中 θ^* 对应于算子 ΠT^γ的不动点，即 $\theta^{*\intercal}\mathbf{\Phi}(s)$，其中 θ^* 在假设 *4*下是唯一的。

证明。 回顾一下， $\Pi(C+\beta\theta^\intercal\mathbf{\Phi}(S_1))$表示到 $\{\phi^i(s)\}$的张成空间上的投影映射。注意

$$\begin{aligned}\Pi(C + \beta\theta^\intercal\mathbf{\Phi}(S_1)) &= \Pi\left(E\left[C + \beta\theta^\intercal\mathbf{\Phi}(S_1)|S\right]\right)\\ &= \Pi\left(T^\gamma(\theta^\intercal\mathbf{\Phi}(S))\right).\end{aligned}$$

一阶条件意味着 $E\left[\mathbf{\Phi}(S)[C + \beta\theta^\intercal\mathbf{\Phi}(S_1) - \Pi(C + \beta\theta^\intercal\mathbf{\Phi}(S_1))]\right] =0$。然后，加上并减去 $\Pi(C + \beta\theta^\intercal\mathbf{\Phi}(S_1))$：

$$\begin{aligned}&(\theta - \theta^*)^\intercal E[\mathbf{\Phi}(S)[C + \beta\theta^\intercal\mathbf{\Phi}(S_1) - \Pi(C + \beta\theta^\intercal\mathbf{\Phi}(S_1))\\ &\qquad + (\Pi(C + \beta\theta^\intercal\mathbf{\Phi}(S_1)) - \theta^\intercal\mathbf{\Phi}(S))]]\\ &= (\theta - \theta^*)^\intercal E[\mathbf{\Phi}(S)\left(\Pi(C + \beta\theta^\intercal\mathbf{\Phi}(S_1)) - \theta^\intercal\mathbf{\Phi}(S)\right)].\end{aligned}$$

接下来，我们使用等式 $\Pi(C+\beta\theta^\intercal\mathbf{\Phi}(S_1)) = \Pi\left(T^\gamma(\theta^\intercal\mathbf{\Phi}(S))\right)$，并加上再减去 $\theta^{*\intercal}\mathbf{\Phi}(S) = \Pi T^\gamma(\theta^{*\intercal}\mathbf{\Phi}(S))$，以利用复合算子 ΠT^γ (参见命题 2)：

$$\begin{aligned}&(\theta - \theta^*)^\intercal E[\mathbf{\Phi}(S)\left(\Pi(C + \beta\theta^\intercal\mathbf{\Phi}(S_1)) - \theta^\intercal\mathbf{\Phi}(S)\right)]\\ &= (\theta - \theta^*)^\intercal E[\mathbf{\Phi}(S)(\Pi T^\gamma(\theta^\intercal\mathbf{\Phi}(S)) - \theta^{*\intercal}\mathbf{\Phi}(S))]\\ &+ (\theta - \theta^*)^\intercal E[\mathbf{\Phi}(S)(\theta^{*\intercal}\mathbf{\Phi}(S) - \theta^\intercal\mathbf{\Phi}(S))]\\ &\leq \|(\theta - \theta^*)^\intercal\mathbf{\Phi}(S)\|_2\|\Pi T^\gamma(\theta^\intercal\mathbf{\Phi}(S)) - \theta^{*\intercal}\mathbf{\Phi}(S)\|_2\\ &- \|(\theta - \theta^*)^\intercal\mathbf{\Phi}(S)\|_2^2\\ &\leq (\beta - 1)\|(\theta - \theta^*)^\intercal\mathbf{\Phi}(S)\|_2^2 < 0\end{aligned}$$

其中我们使用了柯西-施瓦茨不等式，并且 L_2 范数是相对于不变测度 π而言的。最后一步由 θ^*的唯一性得出。

□

于是我们有

$$\begin{aligned}(-A)(\theta-\theta^*) &= E\left[\beta\mathbf{\Phi}(S)\mathbf{\Phi}^\intercal(S_1)-\mathbf{\Phi}(S)\mathbf{\Phi}^\intercal(S)\right](\theta-\theta^*)\\ &= E\left[\mathbf{\Phi}(S)\left(C+\beta\mathbf{\Phi}^\intercal(S_1)\theta-\mathbf{\Phi}^\intercal(S)\theta\right)\right]\\ &\quad - E\left[\mathbf{\Phi}(S)\left(C+\beta\mathbf{\Phi}^\intercal(S_1)\theta^*-\mathbf{\Phi}^\intercal(S)\theta^*\right)\right]\\ &= E\left[\mathbf{\Phi}(S)\left(C+\beta\mathbf{\Phi}^\intercal(S_1)\theta-\mathbf{\Phi}^\intercal(S)\theta\right)\right]\end{aligned}$$

其中，最后一步由以下事实得出： $\mathbf{\Phi}^\intercal(S)\theta^*$ 是不动点，对应的算子为 ΠT^γ，并且 $\Pi(C+\beta\theta^\intercal\mathbf{\Phi}(S_1)) = \Pi\left(T^\gamma(\theta^\intercal\mathbf{\Phi}(S))\right)$。结合引理 5，可知

$$(\theta-\theta^*)(-A)(\theta-\theta^*) < 0$$

对所有 $\theta \neq \theta^*$均成立，因此利用命题 1 可得 θ_t 收敛到某个 θ' ，且满足 $A\theta' = b$，这意味着

$$E\left[\mathbf{\Phi}(S)\left(C+\beta\mathbf{\Phi}^\intercal(S_1)\theta'-\mathbf{\Phi}^\intercal(S)\theta'\right)\right] = 0$$

那么，正如前面所论证的， θ' 也满足：

$$E\left[\mathbf{\Phi}(S)\left(T^\gamma(\mathbf{\Phi}^\intercal(S)\theta')-\mathbf{\Phi}^\intercal(S)\theta'\right)\right] = 0$$

这进而意味着 $\mathbf{\Phi}^\intercal(S)\theta'$ 是算子 ΠT^γ 的不动点。由于不动点唯一，因此有 $\theta' = \theta^*$ ，证明完毕。

□

2.5 极限值的误差分析

回顾一下

$$J_\beta^\pi(s_0,\gamma) = \sum_{t=0}^{\infty}\beta^t E^\gamma[c(\bar{S}_t,\bar{U}_t)]$$

表示第 2.2 节中定义的平稳状态 MDP 的值，特别地，它是式（ T^γ （12）中给出的贝尔曼算子的不动点。由此可得以下直接的界：

命题 3。在策略 γ 下，联合过程(S_t, C_t, U_t) 的不变测度 π 下，有

$$\|J_\beta^\pi(S,\gamma) - \theta^{*\intercal}\mathbf{\Phi}(S)\|_2 \leq \frac{1}{1-\beta}\|J_\beta^\pi(S,\gamma) - \Pi(J_\beta^\pi(S,\gamma))\|_2.$$

证明。 我们从以下界开始

$$\begin{aligned}\|J_\beta^\pi(S,\gamma) - \theta^{*\intercal}\mathbf{\Phi}(S)\|_2 &\leq \|J_\beta^\pi(S,\gamma) - \Pi T^\gamma(J_\beta^\pi(S,\gamma))\|_2 + \|\Pi T^\gamma(J_\beta^\pi(S,\gamma)) - \theta^{*\intercal}\mathbf{\Phi}(S)\|_2 \\ &\leq \|J_\beta^\pi(S,\gamma) - \Pi(J_\beta^\pi(S,\gamma))\|_2 + \beta\|J_\beta^\pi(S,\gamma) - \theta^{*\intercal}\mathbf{\Phi}(S)\|_2\end{aligned}$$

对于第一项，由于 $J_\beta^\pi(s,\ \gamma)$ 是算子 T^γ（在一致范数下）的不动点，因此有 $\Pi T^\gamma(J_\beta^\pi(S,\ \gamma))\ = \Pi J_\beta^\pi(S,\ \gamma)$。对于第二项，我们利用 $\theta^{*\intercal}\mathbf{\Phi}(S)$ 是 ΠT^γ 的不动点这一事实，而 ΠT^γ 在 L_2 范数下是一个压缩映射。合并各项即得证。 □

该上界与价值函数 $J_\beta^\pi(s,\ \gamma)$ 在平稳测度 π 关于策略 γ 的 L_2 范数下到 $\mathbf{\Phi}$ 的张成空间的投影误差有关。接下来，我们推导近线性价值函数的一致范数差的上界：

假设 5。 我们假设存在某个 θ 和某个常数 $\lambda < \infty$ ，使得

$$\|J_\beta^\pi(s,\gamma) - \hat{\theta}^{\intercal}\mathbf{\Phi}(s)\|_\infty \leq \lambda.$$

命题 4。 在假设 5下，我们有

$$\|J_\beta^\pi(s,\gamma) - \theta^{*\intercal}\mathbf{\Phi}(s)\|_\infty \leq \lambda\left(1 + \frac{2-\beta}{1-\beta}\sqrt{\frac{d}{\sigma_{\min}}}\right)$$

其中， θ^* 是式（15）中通过迭代学习得到的参数。此外，$\sigma_{\min}$ 是矩阵 $E[\mathbf{\Phi}(S)\mathbf{\Phi}^\intercal(S)]$ 在 S 服从不变测度 π时的最小特征值。

证明。 我们首先加上并减去 $\theta^\intercal\mathbf{\Phi}(s)$:

$$\begin{aligned}&\|J_\beta^\pi(s,\gamma) - \theta^{*\intercal}\mathbf{\Phi}(s)\|_\infty \\ &\leq \|J_\beta^\pi(s,\gamma) - \hat{\theta}^{\intercal}\mathbf{\Phi}(s)\|_\infty + \|\hat{\theta}^{\intercal}\mathbf{\Phi}(s) - \theta^{*\intercal}\mathbf{\Phi}(s)\|_\infty.\end{aligned}$$

根据假设，第一项由 λ界定。我们在 L_2 范数下分析第二项:

$$\begin{aligned}
&\|\hat{\theta}^\intercal \mathbf{\Phi}(S) - \theta^{*\intercal} \mathbf{\Phi}(S)\|_2 \\
&\leq \|\hat{\theta}^\intercal \mathbf{\Phi}(S) - J_\beta^\pi(S,\gamma)\|_2 + \|J_\beta^\pi(S,\gamma) - \theta^{*\intercal} \mathbf{\Phi}(S)\|_2 \\
&\leq \lambda + \frac{1}{1-\beta} \|J_\beta^\pi(S,\gamma) - \Pi(J_\beta^\pi(S,\gamma))\|_2 \\
&\leq \frac{2-\beta}{1-\beta}\lambda.
\end{aligned}$$

对于第二个不等式，我们使用了命题 3。此外，根据假设 5，由于我们是在概率测度下进行分析，因此 $J_\beta^\pi(S,\gamma)$ 与 $\theta^\intercal \mathbf{\Phi}(S)$ 之间的 L_2 距离也有界。对于最后一个不等式，我们利用这样一个事实：由于 $\Pi(J_\beta^\pi(S,\gamma))$ 是 $J_\beta^\pi(S,\gamma)$ 在 π 的 L_2 范数下的投影，因此它与 $J_\beta^\pi(S,\gamma)$ 的距离达到最小值，所以它的误差界必定小于 θ 所达到的误差界 λ 。

另一方面，我们有

$$\begin{aligned}
&\|\hat{\theta}^\intercal \mathbf{\Phi}(S) - \theta^{*\intercal} \mathbf{\Phi}(S)\|_2^2 \\
&= (\theta^* - \hat{\theta}) E[\mathbf{\Phi}(S)\mathbf{\Phi}^\intercal(S)](\theta^* - \hat{\theta}) \geq \|\theta^* - \hat{\theta}\|_2^2 \sigma_{\min}
\end{aligned}$$

其中， $\sigma_{\min}$ 是矩阵 $E[\mathbf{\Phi}(S)\mathbf{\Phi}^\intercal(S)]$ 的最小特征值，且 S 按不变测度 π 分布。请注意，θ 向量的 2 范数是标准的 2 范数，不要将其与函数上 π 下的 L_2 范数混淆。结合目前得到的结果，我们可以写出

$$\|\theta^* - \hat{\theta}\|_2 \leq \frac{2-\beta}{1-\beta} \frac{\lambda}{\sqrt{\sigma_{\min}}}.$$

回到初始项，对于任意 S，我们有

$$\begin{aligned}
&|J_\beta^\pi(s,\gamma) - \theta^{*\intercal} \mathbf{\Phi}(s)| \\
&\leq |J_\beta^\pi(s,\gamma) - \hat{\theta}^\intercal \mathbf{\Phi}(s)| + |\hat{\theta}^\intercal \mathbf{\Phi}(s) - \theta^{*\intercal} \mathbf{\Phi}(s)| \\
&\leq \lambda + \|\theta^* - \hat{\theta}\|_2 \|\mathbf{\Phi}(s)\|_2 \leq \lambda + \frac{2-\beta}{1-\beta} \frac{\lambda\sqrt{d}}{\sqrt{\sigma_{\min}}}
\end{aligned}$$

其中我们使用了对于所有基函数， $\|\Phi^i\|_\infty \leq 1$ 的假设。因此，证明完成。

□

命题4 给出了基于原始过程的平稳分布构造的合成马尔可夫决策过程的学习得到的价值与价值之间的误差界

然而，它并未回答我们真正关心的问题：在状态过程 γ的真实非马尔可夫动力学下，策略 S_t的价值差异。这个问题需要对该过程的混合性质进行更仔细的分析。在本文第4 节中，我们将针对有限记忆策略下的部分可观测马尔可夫决策过程部分回答这一问题；这是非马尔可夫过程的一个特例。

3 关于近似学习最优Q值

本节将重点转向使用线性函数近似来近似学习最优Q值。我们通过以下方式扩展基函数：$\{\phi^i(s,\ u)\}_{i=1}^d$ 其中 $\phi^i(s,\ u) : \mathrm{S}\times\mathrm{U} \to \mathrm{R}$。我们假设对所有 $i = 1,\ \ldots,\ d$，都有 $\|\phi^i\|_\infty \leq 1$。

我们将贪心策略记为 $\gamma_{\theta_t}(s)$，使得 $\min_v \theta_t^\intercal \mathbf{\Phi}(s,\ v) = \theta_t^\intercal \mathbf{\Phi}(s,\ \gamma_{\theta_t}(s))$。考虑以下迭代，

$$\theta_{t+1} = \theta_t - \alpha_t \mathbf{\Phi}(S_t, U_t)\left[\theta_t^\intercal \mathbf{\Phi}(S_t, U_t) - C_t - \beta\theta_t^\intercal \mathbf{\Phi}(S_{t+1}, \gamma_{\theta_t}(S_{t+1}))\right] \quad (16)$$

其中，动作是根据某个时不变探索策略γ 选取的：$\mathrm{S} \to \mathrm{U}$。

式（16）中的最优Q学习迭代分析不同于式（15）中给出的策略评估分析。首先注意，增益矩阵为

$$A(S_t, S_{t+1}, U_t, \theta_t) = -\beta\mathbf{\Phi}(S_t, U_t)\mathbf{\Phi}^\intercal(S_{t+1}, \gamma_{\theta_t}(S_{t+1})) + \mathbf{\Phi}(S_t, U_t)\mathbf{\Phi}^\intercal(S_t, U_t) \quad (17)$$

因此，迭代关于 θ_t并非完全线性。尽管如此， [1] 中的分析在满足某些正则性条件时适用于非线性函数。此外，这种分析也可能扩展到非马尔可夫过程，正如我们在第2节所做的那样。然而，与策略评估方法（见命题 2）不同，联合投影-贝尔曼算子通常并非压缩映射，这主要是由于探索策略与贝尔曼算子中隐含的贪心策略之间存在差异。

备注 3. 请注意，这些方法之间的另一个差异源于遍历性假设。具体而言，假设 *1*中策略评估方法的遍历性条件针对的是增益矩阵 $A(Z_t)$，它与 θ_t 无关。然而，对于 *Q*学习迭代，*Q*学习迭代的增益矩阵（*17*）以非线性方式依赖于参数

因此，必须相应地调整遍历性条件。具体而言，对于任意 $f(Z_t)$ 且满足 $\|f\|_\infty \le 1$，我们定义：

$$\sum_{k=0}^{\infty} \|E[f(Z_{t+k})|\mathcal{F}_t] - \bar{f}\| =: Y_t^f$$

其中 $\bar{f} := \int f(z)\pi(dz)$，其中 π 是联合过程 $Z_t = (S_t, S_{t+1}, C_t, U_t)$ 的平稳分布。该假设经过调整，使得 $\sup_{\{\|f\|\le 1\},\, t} \|Y_t^f\| < \infty$。

回顾针对平稳状态 MDP 定义的贝尔曼算子 (13)

$$Tg(s,u) := c(s,u) + \beta \int \inf_v g(s_1, v)\eta(ds_1|s,u).$$

此外，Π 表示 $L_2(\mathsf{S} \times \mathsf{U}\ \pi)$ 上到 $\{\phi^i(s,u)\}_i$ 的张成空间的正交投影映射。

（16）中的迭代收敛性与联合算子 ΠT 生成的确定性序列的收敛性分析有关。不幸的是，除某些特殊情形外，该映射并非压缩映射：

1）显然，一种情形是成本函数 $c(s,u)$ 和转移模型 $\eta(ds_1|s,u)$ 能够用基函数 $\{\phi^i(s,u)\}$ 完美分解（成本函数 c 使用实参数，核 η 使用符号测度）。这种情形也称为线性马尔可夫决策过程，此时应用贝尔曼算子不会使迭代超出基函数的线性张成空间。因此，联合映射 ΠT 等价于仅应用贝尔曼算子，而贝尔曼算子在一致范数下是压缩映射。2）折扣后，贪心策略所诱导的特征协方差一致地受探索策略所诱导的特征协方差支配。3）当基函数通过对状态空间进行离散化来选择时，投影会将连续状态空间马尔可夫决策过程映射为离散化有限马尔可夫决策过程，因此联合映射保持一致压缩性质。

接下来，我们将更详细地解释情形 (2) 和 (3)。

3.1 贪心策略协方差占优

可以证明，联合映射 ΠT 在 L_2 范数下是压缩映射，但前提是探索策略和贪心策略所诱导的自相关矩阵满足一个略为严格的假设。该假设最初由 [17] 针对马尔可夫决策过程的线性函数逼近 Q学习推导得出。对于非马尔可夫过程，平稳状态 MDP 也需要满足同一假设；该 MDP 对应于非马尔可夫过程在探索策略下的平稳分布。

我们记

$$\Sigma_\gamma := E\left[\mathbf{\Phi}(S,U)\mathbf{\Phi}^\intercal(S,U)\right] \tag{18}$$

其中，$(S,\ U)$服从探索策略 γ 下过程 $(S_t,\ U_t)$ 的不变测度。我们还记 $\gamma_\theta(s) = \arg\min_u \theta^\intercal \mathbf{\Phi}(s,\ u)$ 为参数 θ 的贪心策略。我们定义

$$\Sigma_\theta := E\left[\mathbf{\Phi}(S,\gamma_\theta(S))\mathbf{\Phi}^\intercal(S,\gamma_\theta(S))\right] \tag{19}$$

其中，S 服从 $(S_t,\ U_t)$ 的不变测度。

回顾一下，在（13）中定义的平稳状态 MDP 的贪心动作选择对应的贝尔曼算子，其满足

$$Tg(s,u) := c(s,u) + \beta \int \inf_v f(s_1,v)\eta(ds_1|s,u)$$

还需注意，本节中 Π 表示基函数 $\{\phi^i(s,\ u)\}_{i=1}^d$ 张成空间上的投影映射。

为保证算法的收敛，我们提出以下假设：

假设6。对于所有 $\theta \in \mathbb{R}^d$

$$\beta^2\Sigma_\theta < \Sigma_\gamma.$$

我们注意到，该假设与 [17],中使用的假设相似，并表明对于较大的 β，贪心策略与探索策略彼此接近，而这在实践中可能会有较强的限制。

命题5. 在假设 *6*下，联合算子 ΠT 在 $L_2(\mathbb{S} \times \mathbb{U}\ \pi)$中是一个压缩映射。

证明. 投影映射是非扩张的，因此我们需要证明贝尔曼映射在 L_2 中是一个压缩映射。令 $f(s,u) = \theta_f^\intercal \Phi(s,u)$ 和 $g(s,u) = \theta_g^\intercal \Phi(s,u)$。我们有

$$\|T(f) - T(g)\|_2^2 \leq \beta^2 \int \left(\min_v f(s,v) - \min_v g(s,v)\right)^2 \pi(ds).$$

我们可以证明，$|\min_v f(s,v) - \min_v g(s,v)| \leq \max_\theta |f(s,\gamma_\theta(s)) - g(s,\gamma_\theta(s))|$。将取得最大值的 θ 记为 $\bar{\theta}$:

$$\begin{aligned}
&\beta^2 \int \left(\min_v f(s,v) - \min_v g(s,v)\right)^2 \pi(ds) \\
&\leq \beta^2 (\theta_f - \theta_g)^\intercal \int \mathbf{\Phi}(s,\gamma_{\bar{\theta}}(s)) \mathbf{\Phi}^\intercal(s,\gamma_{\bar{\theta}}(s)) \pi(ds) (\theta_f - \theta_g) \\
&= \beta^2 (\theta_f - \theta_g)^\intercal \Sigma_{\bar{\theta}} (\theta_f - \theta_g) \\
&< (\theta_f - \theta_g)^\intercal \Sigma_\gamma (\theta_f - \theta_g) = \|f - g\|_2^2
\end{aligned}$$

其中，最后一个不等式用到了假设 6 。 □

3.2 离散化下的收敛

到目前为止，在分析中我们一直使用 L_2 范数。我们发现，在贝尔曼算子中，探索策略与贪心策略之间的差异，使最优 Q 值估计的压缩映射分析变得复杂。

本节讨论一种特殊情况：此时投影映射不会扩大函数的上确界范数。因此，在进行压缩映射分析时，可以直接使用一致范数 $\|\cdot\|_\infty$ 。

设 $\{B_i^s\}_{i=1}^{M_s}$ 是 S 的互不相交子集，且 $\cup_{i=1}^{M_s} B_i^s = \mathsf{S}$。同理，设 $\{B_i^u\}_{i=1}^{M_u}$ 是 U 的互不相交子集，且 $\cup_{i=1}^{M_u} B_i^u = \mathsf{U}$。由此得到的离散化会在联合状态-动作变量 $(s, u) \in (\mathsf{S} \times \mathsf{U})$ 上形成矩形离散化。我们将由此得到的联合 $(s, u) \in (\mathsf{S} \times \mathsf{U})$ 变量的离散化区间记为 $\{A_i\}_{i=1}^{(M_s \times M_u)}$。定义以下基函数

$$\phi^i(s,u) = \mathbb{1}_{A_i}(s,u), \text{ for all } i = 1, \ldots, (M_s \times M_u)$$

其中，$\mathbb{1}_{A_i}(s, u)$ 是集合 A_i 的示性函数。注意，投影映射 Π 满足 $\Pi(f)(s, u) = \theta^\intercal \mathbf{\Phi}(s, u)$，其中 $\theta = \Sigma_\gamma^{-1} E_\pi [\mathbf{\Phi}(S, U) f(S, U)]$ 是探索策略下的不变测度 π

策略 γ 其中 Σ_γ 定义见（18）。对于离散化这一特例，基函数 ϕ^i 完全正交归一；对于任意输入 $(s,\ u)$，其中只有一个基函数取值为 1，其余均为 0。因此，我们有$\Sigma_\gamma^{-1}(i,\ i) = \pi^{1}(A_i)$，且非对角元素均为 0。由此可知，对于某个 $(s,\ \overline{u}),\ \in A_i$

$$\begin{aligned}\Pi(f)(s,u) &= \frac{\int_{A_i} f(s',u')\pi(ds',du')}{\pi(A_i)} \\ &= \int_{A_i} f(s',u')\pi_i(ds',du') \le \sup_{s,u\in A_i} |f(s,u)|\end{aligned}$$

其中，$\pi_i(ds,\ du)$ 是在 A_i 上归一化的概率测度。因此，我们有 $\|\Pi(f)\|_\infty \le \|f\|_\infty$，特别是，联合算子 ΠT 在上确界范数下是一个压缩映射。

我们将 $\hat{\mathsf{S}} := \{s^1,\dots,s^{M_s}\}$ 和 $\hat{\mathsf{U}} := \{u^1,\dots,u^{M_u}\}$ 记为。定义映射 $q_s : \mathsf{S} \to \hat{\mathsf{S}}$ 和 $q_u : \mathsf{U} \to \hat{\mathsf{U}}$，使得当 $s \in B_i^s$ 时，$q_s(s) = s^i$，且当 $u \in B_i^u$. 时，$q_u(u) = u^i$

特别地，在离散化条件下，式（16）中的学习算法具有如下形式：对任意 s^i 和 u^j:

$$Q_{t+1}(s^i,u^j) = Q_t(s^i,u^j) - \alpha_t \mathbb{1}_{\{\hat{S}_t=s^i,\hat{U}_t=u^j\}}\left[Q_t(\hat{S}_t,\hat{U}_t) - C_t - \beta V_t(\hat{S}_{t+1})\right]$$

其中 $V_t(s) := \min_v Q_t(s,\ v)$，且 $S_t := q_s(S_t),\ U_t := q_u(U_t)$。

注意，上述算法是在离散化状态和动作过程 $S_t = q_s(S_t)$、$U_t = q_u(U_t)$ 上运行的标准（表格型）Q 学习算法。文献 [15] 研究了该算法在具有随机且依赖于状态的学习率的非马尔可夫过程下的收敛性：

定理2. 对于所有 $s^i \in B_i^s$ 和 $u^j \in B_j^u$ 以及 $\hat{S}_t := q_s(S_t)$、$\hat{U}_t := q_u(U_t)$，考虑

$$Q_{t+1}(s^i,u^j) = Q_t(s^i,u^j) - \alpha_t(s^i,u^j)\left[Q_t(\hat{S}_t,\hat{U}_t) - C_t - \beta V_t(\hat{S}_{t+1})\right].$$

假设对于任意可测有界函数 f，以概率*1*有：

$$\frac{1}{N}\sum_{t=0}^{N-1} f(\hat{S}_{t+1},\hat{S}_t,\hat{U}_t,C_t) \to \int f(\hat{s}_1,\hat{s},\hat{u},c)\pi(d\hat{s}_1,d\hat{s},d\hat{u},dc)$$

对于某个测度 π，使得 $\pi(\hat{\mathbb{S}} \times s^i \times u^j \times \mathbb{R}) > 0$ 对于任意 $(s^i,\ u^j) \in \hat{\mathbb{S}} \times \hat{\mathbb{U}}$。此外，对于学习率，我们假设 $\alpha_t(s^i,\ u^j) = 0$，除非

（$\hat{S}_t$、$\hat{U}_t$）=（s^i、u^j）。此外，

$$\alpha_t(s^i, u^j) = \frac{1}{1 + \sum_{k=0}^{t} 1_{\{\hat{S}_k = s^i, \hat{U}_k = u^j\}}}$$

并且其概率为 1。于是，对于每个 $Q_t(s^i, u^j) \to Q^*(s^i, u^j)$ 几乎必然成立，其中每个 $(s^i, u^j) \in \hat{\mathbb{S}} \times \hat{\mathbb{U}}$ 对 Q^* 都是如此；这里的 *2.2* 节中针对离散化状态和动作构造的平稳状态 *MDP* 的最优 *Q* 值为 Q^*。

4 使用有限记忆对部分可观测马尔可夫决策过程进行函数逼近

4.1 部分可观测马尔可夫决策过程

令 X $\subset \mathbb{R}^m$ 表示一个博雷尔集，它是某个 $m \in$N 的部分可观测马尔可夫决策过程的状态空间。令 Y $\subset \mathbb{R}^n$ 为另一个博雷尔集，表示该模型的观测空间；状态通过观测通道 O 进行观测。观测通道 O 定义为从 X 到 Y 的随机核（正则条件概率），使得对于每个 $x \in$ X，$O(\cdot|x)$ 都是定义在 Y 的 σ-代数 $\mathcal{B}$(Y) 上的概率测度；并且对于每个 $A \in \mathcal{B}$(Y)，$O(A|\cdot)$: X $\to [0,1]$ 都是博雷尔可测函数。U $\in \mathbb{R}^l$ 表示动作空间。可容许策略 γ 是一个控制函数序列 $\{\gamma_t, t \in \mathbb{Z}_+\}$，满足 γ_t 关于由信息变量 $I_t = \{Y_{[0,t]}, U_{[0,t-1]}\}$、 $t \in$N $I_0 = \{Y_0\}$、生成的 σ-代数可测，其中 $U_t = \gamma_t(I_t)$ $t \in \mathbb{Z}_+$ 是取值于 U 的控制动作，且 $Y_{[0,t]} = \{Y_s, 0 \le s \le t\}$、$U_{[0,t-1]} = \{U_s, 0 \le s \le t-1\}$.。我们将所有此类可容许策略构成的集合定义为 Γ。系统的更新规则由以下关系确定：

$$\Pr\big((X_0, Y_0) \in B\big) = \int_B \mu(dx_0) O(dy_0|x_0), \quad B \in \mathcal{B}(\mathbb{X} \times \mathbb{Y}),$$

其中，μ 是初始状态 X_0 的（先验）分布，并且

$$\Pr\Big((X_t, Y_t) \in B \,\Big|\, (X, Y, U)_{[0,t-1]} = (x, y, u)_{[0,t-1]}\Big) = \int_B \mathcal{T}(dx_t|x_{t-1}, u_{t-1}) O(dy_t|x_t),$$

$B \in \mathcal{B}$(X $\times$ Y) $t \in$ N，其中 $\mathcal{T}$是模型的转移核，是从 X $\times$ U 到 X 的随机核。我们将智能体的目标设定为

（决策者）最小化无限时域折扣成本，

$$J_\beta(\mu,\gamma)=E_\mu^\gamma\left[\sum_{t=0}^{\infty}\beta^t c(X_t,U_t)\right] \tag{20}$$

对于某个折扣因子 $\beta\in(0,1)$，在可容许策略 $\gamma\in\Gamma$ 的集合上，其中 $c:$ X $\times$U $\to$ R 是博雷尔可测的阶段成本函数，E_μ^γ 表示在初始状态概率测度 μ 、转移核 $\mathcal{T}$ 以及策略 γ 下的信道 O 所对应的期望。注意， $\mu\in\mathcal{P}$ (X)，其中我们令 $\mathcal{P}$(X) 表示 X 上概率测度的集合。我们将折扣无限时域设置下的最优成本定义为先验分布的函数：

$$J_\beta^*(\mu)=\inf_{\gamma\in\Gamma}J_\beta(\mu,\gamma). \tag{21}$$

对于部分可观测马尔可夫决策过程的分析，一种常见方法是将问题重新表述为一个全可观测马尔可夫决策过程，其中决策者根据可用历史 I_t 跟踪状态 X_t 的后验分布，这也称为信念马尔可夫决策过程。在下文中，我们将采用一种基于有限记忆（窗口）信息变量的、不同但相关的重新表述方法。

4.2 使用有限记忆变量化归为全可观测

以下构造主要取自 [14], 不过，为求完整，我们在此详细介绍该方法。

我们利用N 个阶段之前的预测器和最近的 N 个信息变量（即测量值和动作），构造一个全可观测马尔可夫决策过程的归约。考虑时间 t的以下状态变量：

$$z_t=(\mu_{t-N},h_t) \tag{22}$$

其中，对于 $N\geq 1$

$$\begin{aligned}\mu_{t-N}&=Pr(X_{t-N}\in\cdot|y_{t-N-1},\ldots,y_0,u_{t-N-1},\ldots,u_0),\\ h_t&=\{y_t,\ldots,y_{t-N},u_{t-1},\ldots,u_{t-N}\}\end{aligned}$$

以及 $h_t=y_t$ 对于 $N=0$ ，其中 μ 是 X_0上的先验概率测度。请注意，尽管有限记忆变量 h_t 取决于记忆长度 N，为方便记号起见，我们省略这种依赖关系。

采用此表示的状态空间为 $\mathcal{Z} = \mathcal{P}(\mathbf{X}) \times \mathbf{Y}^{N+1} \times \mathbf{U}^N$，其中我们为 $\mathcal{Z}$ 赋予积拓扑：在 $\mathcal{P}(\mathbf{X})$ 上采用弱收敛拓扑，在 $\mathbf{Y}^{N+1} \times \mathbf{U}^N$ 上采用通常的（坐标）拓扑。

现在我们可以定义阶段成本函数和转移概率。考虑新的成本函数 $\hat{c}: \mathcal{Z} \times \mathbf{U} \to \mathbf{R}$,

$$\hat{c}(z_t, u_t) = \hat{c}(\mu_{t-N}, h_t, u_t) = \int_{\mathbb{X}} c(x_t, u_t) P^{\mu_{t-N}}(dx_t | y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}).$$

此外，我们可以如下定义 $N = 1$ 的转移概率（为简便起见）：对于某个 $A \in \mathcal{B}(\mathcal{Z})$，使得

$$A = B \times \{\hat{y}_{t-N+1}, \hat{u}_t, \ldots, \hat{u}_{t-N+1}\}, \quad B \in \mathcal{B}(\mathcal{P}(\mathbb{X}))$$

我们记作

$$\begin{aligned}
&Pr(z_{t+1} \in A | z_t, \ldots, z_0, u_t, \ldots, u_0) \\
&= Pr(\mu_t \in B, \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \mu_{[t-1,0]}, y_{[t,0]}, u_{[t,0]}) \\
&= \mathbb{1}_{\{y_t, u_t = \hat{y}_t, \hat{u}_t, G(\mu_{t-1}, y_{t-1}, u_{t-1}) \in B\}} P^{\mu_{t-1}}(\hat{y}_{t+1} | y_t, y_{t-1}, u_t, u_{t-1}) \\
&= Pr(\mu_t \in B, \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \mu_{t-1}, y_t, y_{t-1}, u_t, u_{t-1}) \\
&= Pr(z_{t+1} \in A | z_t, u_t) =: \int_A \eta(dz_{t+1} | z_t, u_t)
\end{aligned}$$

其中，映射 G 定义为

$$G(\mu_{t-1}, y_{t-1}, u_{t-1}) = P^{\mu}(X_t \in \cdot | y_{t-1}, \ldots, y_0, u_{t-1}, \ldots, u_0).$$

对于某个可容许策略 γ 和某个初始状态 $z_0 \in \mathcal{Z}$，我们将其诱导的成本记为

$$J_\beta(z_0, \gamma) = \sum_{t=0}^{\infty} \beta^t E^\gamma[\hat{c}(Z_t, U_t)].$$

相应地，我们用 $J^*_\beta(z_0)$ 表示最优价值函数。注意，这种构造不会损失最优性。具体而言，对于固定的 μ_{-N}，假设某个任意策略 γ 从时刻 $-N$ 到 -1 采取行动，则可以证明

$$E\left[J^*_\beta(Z_0)\right] = E\left[J^*_\beta(\mu_{-N}, H_0)\right] = E[J^*_\beta(\mu_0)]$$

其中，左侧的期望是关于 $H_0 = \{Y_0, \ldots, Y_{-N}, U_{-1}, \ldots, U_{-N}\}$ 的，右侧的期望是关于 μ_0 的

$Pr(X_0 \in \cdot | Y_{-1}, \ldots\ldots, Y_{-N}, U_{-1}, \ldots\ldots, U_{-N})$。注意，$J^*_\beta(\mu_0)$ 是式 (21) 中定义的最优价值函数。

因此，我们得到一个完全可观测的马尔可夫决策过程，其成本函数为 $\hat{c}$，转移核为 η，状态空间为 $\mathcal{Z}$。

4.3 有限记忆信念马尔可夫决策过程的近似

上一节构造的有限记忆信念马尔可夫决策过程模型，其状态空间为

$$\mathcal{Z} = \left\{ \pi, y_{[0,N]}, u_{[0,N-1]} : \pi \in \mathcal{P}(\mathbb{X}), y_{[0,N]} \in \mathbb{Y}^{N+1}, u_{[0,N-1]} \in \mathbb{U}^N \right\},$$

其中，第一个坐标概括过去的信息，第二个和最后一个坐标则包含最近 N 个时间步的信息。

对于固定的 $\pi \in \mathcal{P}(\mathbb{X})$，考虑以下集合 $\mathcal{Z}_\pi$

$$\mathcal{Z}_\pi = \left\{ \pi, y_{[0,N]}, u_{[0,N-1]} : y_{[0,N]} \in \mathbb{Y}^{N+1}, u_{[0,N-1]} \in \mathbb{U}^N \right\}$$

使得时刻 t 的状态为 $\hat{z}_t = (\pi, h_t)$。与式（$z_t = (\mu_{t-N}, h_t)$）中定义的状态 22相比，该近似模型使用 π 作为预测器，无论时刻 $t-N$ 的实际预测器是什么。

由于 π 是固定的，我们可以仅将 h_t视为状态。成本函数定义为

$$\hat{c}_\pi(h_t, u_t) = \hat{c}(\pi, h_t, u_t) = \int_{\mathbb{X}} c(x_t, u_t) P^\pi(dx_t | y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}). \tag{23}$$

我们将受控转移模型定义为

$$\eta_\pi(h_{t+1} | h_t, u_t) := \eta\left(\mathcal{P}(\mathbb{X}), h_{t+1} | \pi, h_t, u_t \right). \tag{24}$$

为简单起见，若我们假设 $N = 1$，则转移可改写为某个 $h_{t+1} = (\hat{y}_{t+1}, \hat{y}_t, \hat{u}_t)$ 和 $h_t = (y_t, y_{t-1}, u_{t-1})$

$$\begin{aligned} &\eta_\pi(\hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | y_t, y_{t-1}, u_{t-1}, u_t) = \eta(\mathcal{P}(\mathbb{X}), \hat{y}_{t+1}, \hat{y}_t, \hat{u}_t | \pi, y_t, y_{t-1}, u_{t-1}, u_t) \\ &= \mathbb{1}_{\{y_t = \hat{y}_t, u_t = \hat{u}_t\}} P^\pi(\hat{y}_{t+1} | y_t, y_{t-1}, u_t, u_{t-1}). \end{aligned} \tag{25}$$

我们为该模型定义如下有限记忆策略下的贝尔曼算子γ^N，使得对于任意 f

$$T^N f(h) = \hat{c}_\pi(h^N, \gamma^N(h)) + \beta \int f(h_1)\eta_\pi(dh_{t+1}|h, \gamma^N(h)) \tag{26}$$

我们将近似模型的最优价值函数记为 J_β^N。注意， J_β^N定义在集合 $\mathcal{Z}_\pi$上。不过，只需将其在第一个坐标上定义为关于 $\mathcal{P}(\mathbf{X})$ 的常数，就可以将其扩展到集合 $\mathcal{Z}$ 。

我们还注意到，由于预测器 π 是固定的， J_β^N可以视为定义在有限记忆信息变量 h_t上的函数。

我们定义如下常数:

$$L_t := \sup_{\hat{\gamma}\in\hat{\Gamma}} E_{\mu_0}^{\hat{\gamma}}\left[\|P^{\mu_t}(X_{t+N} \in \cdot|Y_{[t,t+N]}, U_{[t,t+N-1]}) - P^{\pi}(X_{t+N} \in \cdot|Y_{[t,t+N]}, U_{[t,t+N-1]})\|_{TV}\right] \tag{27}$$

它表示在相同观测和控制动作变量 X_{t+N} 条件下， $Y_{[t,t+N]}$当 $U_{[t,t+N-1]}$ 的先验分布分别为X_t 和 μ_t 时， π后验分布之间全变差距离的期望值。该滤波器稳定性项在接下来的误差分析中起着重要作用。可以证明，在某些假设下，当 $L_t \to 0$ 趋于 $N \to 0$时， [11, 14, 16] （在某些情况下以指数速度）趋于零。关于这一分析的更多细节，请参阅。

命题 6. *[14,* 定理 *3.3]* 对于 $z_0 = (\mu_0, h_0)$，若策略 $\hat{}\gamma$ 作用于前 N 步，则有

- 对于有限记忆策略（不一定是最优策略） γ^N

$$E_{\mu_0}^{\hat{\gamma}}\left[\left|J_\beta^N(h_0, \gamma^N) - J_\beta(z_0, \gamma^N)\right|\right] \le \frac{\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t$$

- 对于价值函数之间的差异，有

$$E_{\mu_0}^{\hat{\gamma}}\left[\left|J_\beta^N(h_0) - J_\beta^*(z_0)\right|\right] \le \frac{\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t$$

其中，期望是针对初始有限记忆变量 h_0 的随机实现。

4.4 部分可观测马尔可夫决策过程的有限记忆策略评估

在本节中，我们旨在学习给定有限记忆策略的近似值。具体而言，我们通过设定2，使用第

$$s_t = h_t = \{y_t, \ldots, y_{t-N}, u_{t-1}, \ldots, u_{t-N}\}.$$

特别地，我们还有 $\mathbb{S} = \mathbb{Y}^{N-1} \times \mathbb{U}^N$。我们使用（15）中的相同迭代，使得

$$\theta_{t+1} = \theta_t - \alpha_t \mathbf{\Phi}(S_t)\left[\theta_t^\intercal \mathbf{\Phi}(S_t) - C_t - \beta \theta_t^\intercal \mathbf{\Phi}(S_{t+1})\right] \tag{28}$$

对于定义在 $\mathbb{S} = \mathbb{Y}^{N-1} \times \mathbb{U}^N$ 上的给定基函数 $\{\phi^i\}_i$。

推论 1 （定理 1的推论）。 设假设 *4* 和假设 *1* 对$Z_t = (S_t, S_{t+1}, c(X_t, U_t)\ U_t)$ 成立，其中 S_t 是有限记忆策略 γ^N 下的有限记忆变量 *28*。那么，（ *28* ）中的迭代收敛到某个θ^*。

遍历性 在这一部分，我们研究有限记忆过程 $\{h_t\}$ 的长期行为。注意，该过程不是马尔可夫链。不过，在有限记忆策略 γ 下，联合过程 (h_t, x_t, u_t) 是一个马尔可夫链。例如，对于 $N = 2$ 以及某个 $B_1, B_2 \in \mathcal{B}(\mathbb{Y})\ B_3,$ $B_4 \in \mathcal{B}(\mathbb{U})\ B_5 \in \mathcal{B}(\mathbb{X})$，记作 $I_{t+1} = \{(y, x, u)_{t+1}, \ldots, (y, x, u)_0\}$

$$\begin{aligned} &Pr(Y_{t+2} \in B_1, Y_{t+1} \in B_2, U_{t+1} \in B_3, X_{t+2} \in B_5, U_{t+2} \in B_4 | I_{t+1}) \\ &= \int_{x_{t+1} \in B_5} \int_{x_{t+2} \in \mathbb{X}} \int_{y_{t+2} \in B_2} \int_{u_{t+2} \in B_4} \mathbb{1}_{(y_{t+1} \in B_2, u_{t+1} \in B_3)} \\ &\quad \gamma(du_{t+2} | y_{t+2}, y_{t+1}, u_{t+1}) O(dy_{t+2} | x_{t+2}) \mathcal{T}(dx_{t+2} | x_{t+1}, u_{t+1}) \end{aligned}$$

这表明联合过程是一个马尔可夫链。我们注意到，该马尔可夫过程的几何遍历性是在有限记忆策略 1 下满足假设 γ^N 的充分条件。

然而，一般而言，仅凭转移核 $\mathcal{T}(\cdot|x,\ u)$ 的性质无法保证满足这一条件。这是因为有限记忆变量 h_t 包含过去的控制动作，因此控制策略对过去控制动作的依赖使遍历性分析变得复杂。例如，对于类型为$u_t \sim \gamma($ $\cdot|u_{t-1})$的策略，动作过程以及由此产生的有限记忆过程的遍历性显然取决于随机策略 $\gamma(\cdot|u_{t-1})$。

注意，如果有限记忆策略 γ 和转移核 $\mathcal{T}$满足小化条件，那么增广过程就是指数遍历的，因此满足假设1。

假设 7。存在非平凡测度 $\lambda_x(\cdot)$ 和 $\lambda_u(\cdot)$，使得

$$\mathcal{T}(dx_1|x,u) \geq \lambda_x(dx_1)$$
$$\gamma(du|h) \geq \lambda_u(du)$$

对于所有 $(x,u) \in \mathbb{X} \times \mathbb{U}$ 以及所有 $h \in \mathbb{Y}^N \times \mathbb{U}^{\mathbb{N}-\mathbb{1}}$。

引理 6。假设 7 蕴含联合过程1 的假设 (h_t, x_t, u_t)。特别地，在假设7下，增广马尔可夫链 (h_t, x_t, u_t) 在有限记忆策略下是指数遍历的。

证明。我们给出关于 $N=2$ 的证明：考虑链 (h_t, x_t, u_t) 从某个起始点 $(y_1, y_0, u_0, x_1, u_1)$ 出发时的两步转移：

$$Pr(dy_3, dy_2, du_2, dx_3, du_3|y_1, y_0, u_0, x_1, u_1)$$
$$= \int_{x_2 \in \mathbb{X}} \gamma(du_3|y_3, y_2, u_2) O(dy_3|x_3) \mathcal{T}(dx_3|x_2, u_2) \gamma(du_2|y_2, y_1, u_1) O(dy_2|x_2) \mathcal{T}(dx_2|x_1, u_1)$$
$$\geq \int_{x_2} \gamma(du_3|y_3, y_2, u_2) O(dy_3|x_3) \mathcal{T}(dx_3|x_2, u_2) \lambda_u(du_2) O(dy_2|x_2) \lambda_x(dx_2)$$
$$=: \lambda_h(du_3, dy_3, dy_2, du_2, dx_3)$$

非平凡测度 $\lambda_h(\cdot)$ 与起始点无关，因此可以证明 (h_t, x_t, u_t) 是指数遍历的（例如参见 [7, 引理 3.3]）。

□

备注 4。对于任何不满足假设 γ 7 的有限记忆策略7，总可以构造一个满足该假设的扰动版本。特别地，令 γ' 为满足小集策略条件的任意策略。那么，扰动后的策略 $\hat{\gamma}(du|h) = (1-\epsilon)\gamma(du|h) + \epsilon\gamma'(du|h)$ 按构造满足假设7。

学习得到的价值的误差界 在上一节中，我们观察到，使用迭代（15），可以学习算子的固定不动点ΠT^N其中，Π是投影映射，且 T^N定义见（26）。接下来，我们将学习得到的价值函数 $\theta^{*\intercal}\mathbf{\Phi}(h)$ 与算子的固定不动点 T^N 进行比较。我们注意到，算子的固定不动点 T^N是有限记忆策略 γ^N 在第4.3节构造的近似模型中的价值函数，我们将其记为 $J_\beta^N(h, \gamma^N)$。然而，这并不是原始部分可观测环境中有限记忆策略的价值。

接下来的结果给出了在原始环境中，学习得到的价值函数相对于有限记忆策略真实价值的误差上界。

ˆ**假设 8.** 我们假设存在某个 θ 和某个常数 $\lambda < \infty$，使得

$$\|J_\beta^N(h, \gamma^N) - \hat{\theta}^\intercal \mathbf{\Phi}(h)\|_\infty \leq \lambda.$$

定理 3。 假设假设 8成立。假设不可观测状态在时间 $-N$ 根据某个 $\mu_{-N} \in \mathcal{P}(\mathbb{X})$开始，并且有限记忆策略 γ^N 从时间 $t = 0$开始执行。我们用 h_0表示从时间 $t = -N$到 $t = 0$的有限记忆变量。对于 (μ_{-N}, h_0)，若策略 ˆγ 作用于前 N 步，则有

$$\begin{aligned} &E_{\mu_{-N}}^{\hat{\gamma}} \left[\left| J_\beta(z_0, \gamma^N) - \theta^{*\intercal} \mathbf{\Phi}(h_0) \right| \right] \\ &\leq \frac{\|c\|_\infty}{(1-\beta)} \sum_{t=0}^{\infty} \beta^t L_t + \lambda \left(1 + \frac{2-\beta}{1-\beta} \sqrt{\frac{d}{\sigma_{\min}}} \right) \end{aligned}$$

其中，期望是关于初始有限记忆变量的随机实现 h_0而言的。此外，$\sigma_{\min}$ 是矩阵 $E[\mathbf{\Phi}(H)\mathbf{\Phi}^\intercal(H)]$ 的最小特征值，其中 H 服从不变测度π。

证明。 证明是命题6 和命题 4的应用。

$$\begin{aligned} &E_{\mu_{-N}}^{\hat{\gamma}} \left[\left| J_\beta(z_0, \gamma^N) - \theta^{*\intercal} \mathbf{\Phi}(h_0) \right| \right] \\ &\qquad \leq E_{\mu_{-N}}^{\hat{\gamma}} \left[\left| J_\beta(z_0, \gamma^N) - J_\beta^N(h, \gamma^N) \right| \right] + E_{\mu_{-N}}^{\hat{\gamma}} \left[\left| J_\beta^N(h, \gamma^N) - \theta^{*\intercal} \mathbf{\Phi}(h_0) \right| \right] \end{aligned}$$

第一项由命题 6 给出界，第二项由命题4给出界。

□

4.5 部分可观测马尔可夫决策过程离散化下的收敛与近似最优性

如第 3节所述，除少数特殊情况外，Q 学习算法通常无法保证收敛。正如第 3.2节所述，基于离散化的基函数就是这类特殊情况之一。

本节将介绍一种针对部分可观测马尔可夫决策过程有限记忆变量的离散化方法，并基于 [13] 展示由此得到的算法的收敛性和近似最优性。

对于弱 Feller 信念马尔可夫决策过程（[6, 9]），[21, 定理 3.16]证明了有限动作策略的近似最优性。若 U 是紧集，则可以构造有限个动作集合，且近似误差可任意小。因此，在下文中我们假设动作空间是有限的

设 $\{B_i\}_{i=1}^{M}$ 是 $\mathbf{Y}$ 的互不相交子集，使得 $\cup_{i=1}^{M} B_i = \mathbf{Y}$。由此得到的离散化进而意味着对有限记忆变量和动作变量（h，u）$\in$（$\mathbf{Y} \times \mathbf{U}$）^{N}的离散化。我们将联合（$h$，$u$）$\in$（$\mathbf{Y} \times \mathbf{U}$）^{N}变量对应的离散化分箱记为 $\{A_i\}_{i=1}^{(M\times|\mathbb{U}|)^N}$。我们定义以下基函数

$$\phi^i(h,u) = \mathbb{1}_{A_i}(h,u), \text{ for all } i = 1,\ldots,(M\times|\mathbb{U}|)^N$$

其中，$\mathbb{1}_{A_i}(h,u)$ 是集合 A_i 的示性函数。

与第3.2节类似，我们令 $q(h)$表示量化映射，该映射使用本节中的构造将连续取值的有限记忆变量映射为其离散化版本。

因此，对于每个 h^i、$i \in\{1,\ldots,M^N\}$以及每个 u^j, $j \in \{1,\ldots,|\mathbf{U}|^N\}$，我们考虑以下迭代

$$Q_{t+1}(h^i,u^j) = Q_t(h^i,u^j) - \alpha_t(h^i,u^j)\left[Q_t(q(H_t),U_t) - c(X_t,U_t) - \beta V_t(q(H_{t+1}))\right] \tag{29}$$

其中，我们将 $V_t(h) = \min_v Q_t(h,v)$ 记为：

以下内容根据本文的结果改编自 [13]:

假设 9。

1. *If* $(q(H_t),U_t) = (h^i,u^j)$

$$\alpha_t(h^i,u^j) = \frac{1}{1+\sum_{k=0}^{t} 1_{\{q(H_k)=h^i,U_k=u^j\}}}$$

否则 α_t（h,u）$=0$。

2. 在任何平稳的 {无记忆或有限记忆探索}策略（记作 γ）下，真实状态过程 $\{X_t\}_t$ 都是正 *Harris* 常返的，因而存在唯一的不变测度 π。

3. 在探索阶段，每个 (h^i,u^j) 对都会被访问无限多次。

4. $\mathbb{Y} \subset \mathbb{R}^n$ 是紧的。

5. $O(dy|x) = g(x,y)\lambda(dy)$，且 $g(y,x)$ 关于 y 满足利普希茨条件，使得$|g(x,y) - g(x,y')| \le \alpha_{\mathbb{Y}}\|y-y'\|$ 对每个 $y,y' \in \mathbb{Y}$ 和某个 $\alpha_{\mathbb{Y}} < \infty$ 的 $x \in \mathbb{X}$ 均成立。

6. 阶段成本函数 $c(x,u)$ 有界，满足 $\sup_{x,u} c(x,u) = \|c\|_\infty < \infty$。

定理 4. • 在探索策略满足假设 *9* 的条件下，（*29*）中的迭代收敛到某个 $Q^*(h,u)$。

• 考虑学习策略 γ^N，其满足 $\gamma^N(h) =$使.....最小$_u Q^*(h,u)$。我们假设不可观测状态在时刻 $-N$ 根据某个 $\mu_{-N} \in \mathcal{P}(\mathbb{X})$开始，并且学习得到的有限记忆策略 γ^N 从时刻 $t = 0$开始执行。我们将从时刻 $t = -N$ 到 $t = 0$的有限记忆变量记为 h_0。对于$z_0 = (\mu_{-N}, h_0)$，若策略 $\hat{}\gamma$在前 N步执行，则有

$$E^{\hat{\gamma}}_{\mu_{-N}}\left[\left|J_\beta(z_0,\gamma^N) - J^*_\beta(z_0)\right|\right] \le \frac{2\|c\|_\infty}{(1-\beta)}\sum_{t=0}^{\infty}\beta^t\hat{L}_t + \frac{\beta}{(1-\beta)^2}\|c\|_\infty\alpha_{\mathbb{Y}}L_{\mathbb{Y}}$$

其中，期望是关于初始有限记忆变量 h_0 的随机实现的，其中

$$L_{\mathbb{Y}} := \max_i \sup_{y,y'\in B_i}\|y-y'\|,$$

$$\hat{L}_t := \sup_{\hat{\gamma}\in\hat{\Gamma}} E^{\hat{\gamma}}_{\mu}\left[\|P^{\pi_t^-}(X_{t+N}\in\cdot|\hat{Y}_{[t,t+N]},U_{[t,t+N-1]}) - P^{\pi^*}(X_{t+N}\in\cdot|\hat{Y}_{[t,t+N]},U_{[t,t+N-1]})\|_{TV}\right]$$

$\hat{}$使得，关于离散化观测的滤波器稳定性项为 L_t，且 $\alpha_{\mathbb{Y}}$ 为信道 O的密度函数g 的 *Lipschitz* 常数。

引理2的证明

证明。 我们记

$$e_t(\theta) := \delta^\intercal \left[A - A(Z_t)\right] \theta + \delta^\intercal \left(b(Z_t) - b\right).$$

此外，根据假设1，我们还定义

$$\phi_t(\theta) := \sum_{k=0}^{\infty} E\left[e_{t+k}(\theta)|\mathcal{F}_t\right]$$

注意，在假设 $A(z)$ 和 $b(z)$ 一致有界的条件下，我们有

$$\left|E[\mathbb{1}_{\{t+1\leq\sigma_n\}} e_{t+k}(\theta_t)|\mathcal{F}_t]\right| \leq K(2^n+1)\left(\|A - E[A(Z_{t+k})|\mathcal{F}_t]\| + \|b - E[b(Z_{t+k})|\mathcal{F}_t]\|\right).$$

根据假设1，我们知道 $\phi_t(\theta) \in L_2$。此外，我们有以下 L_2 关于 $\phi_t(\theta)$ 的界：

$$\begin{aligned}
\|\phi_t(\theta_t)\mathbb{1}_{\{t+1\leq\sigma_n\}}\|_2 &= \left\|\sum_{k=0}^{\infty} E[\mathbb{1}_{\{t+1\leq\sigma_n\}} e_{t+k}(\theta_t)|\mathcal{F}_t]\right\|_2 \\
&\leq K(2^n+1)\left\|\sum_{k=0}^{\infty}\left(\|A - E[A(Z_{t+k})|\mathcal{F}_t]\| + \|b - E[b(Z_{t+k})|\mathcal{F}_t]\|\right)\right\|_2 \\
&\leq K(2^n+1)\|Y_t^A + Y_t^b\|_2 < \infty \text{ uniformly for all } t. \qquad (30)
\end{aligned}$$

其中 Y_t^b和 Y_t^A定义见 (3)，最后一步由假设1 (ii) 得出。

我们记

$$\begin{aligned}
E[e_t(\theta)|F_t] &= \phi_t(\theta) - E[\phi_{t+1}(\theta)|F_t] \\
&= \phi_{t+1}(\theta) - E[\phi_{t+1}(\theta)|F_t] + \left(\phi_t(\theta) - \phi_{t+1}(\theta)\right).
\end{aligned}$$

我们用 $\tau^k := (k+1)$ 表示 $\wedge(\sigma_n - 1)$。不失一般性地，我们假设

$\sigma_n > 2$，并记作：

$$
\begin{aligned}
&\sum_{t=1}^{k+1} \mathbb{1}_{\{t+1\le\sigma_n\}} 2\alpha_t \delta_t^\intercal M_t = \sum_{t=1}^{\tau^k} 2\alpha_t \delta_t^\intercal M_t = \sum_{t=1}^{\tau^k} 2\alpha_t E[e_t(\theta_t)|F_t] \\
&= \sum_{t=1}^{\tau^k} 2\alpha_t \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right) + \sum_{t=1}^{\tau^k} 2\alpha_t \left(\phi_t(\theta_t) - \phi_{t+1}(\theta_t)\right) \\
&= \sum_{t=1}^{\tau^k} 2\alpha_t \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right) \\
&\quad + \sum_{t=1}^{\tau^k} 2\alpha_t \phi_t(\theta_t) - \sum_{t=0}^{\tau^k-1} 2\alpha_{t+1}\phi_{t+1}(\theta_t) \\
&\quad + \sum_{t=0}^{\tau^k-1} 2\alpha_{t+1}\phi_{t+1}(\theta_t) - \sum_{t=1}^{\tau^k} 2\alpha_t \phi_{t+1}(\theta_t) \\
&= \sum_{t=1}^{\tau^k} 2\alpha_t \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right) \\
&\quad + \sum_{t=1}^{\tau^k} 2\alpha_t \left(\phi_t(\theta_t) - \phi_t(\theta_{t-1})\right) \\
&\quad + \sum_{t=1}^{\tau^k-1} 2(\alpha_{t+1} - \alpha_t)\phi_{t+1}(\theta_t) \\
&\quad + 2\alpha_1\phi_1(\theta_0) - 2\alpha_{\tau^k}\phi_{\tau^k+1}(\theta_{\tau^k})
\end{aligned}
$$

我们分别分析这些项：

首项： 我们首先研究该项：$\sum_{t=1}^{\tau^k} 2\alpha_t \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right)$。我们首先注意到，$\sum_{t=1}^{k+1} 2\alpha_t \mathbb{1}_{\{t+1\le\sigma_n\}} \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right)$ 是一个鞅。此外，对于这个鞅的增量，我们有

$$
\begin{aligned}
&\sum_t 4\alpha_t^2 E\left[\mathbb{1}_{\{t+1\le\sigma_n\}} \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right)^2\right] \\
&\le \sum_t 16\alpha_t^2 E[\mathbb{1}_{\{t+1\le\sigma_n\}}\phi_{t+1}^2(\theta_t)] = \sum_t 16\alpha_t^2 \|\mathbb{1}_{\{t+1\le\sigma_n\}}\phi_{t+1}(\theta_t)\|_2^2 \\
&\le \sum_t 16K\alpha_t^2(2^{2n}+1)\|Y_{t+1}^A + Y_{t+1}^b\|_2^2 < \infty.
\end{aligned}
$$

对于某个一般常数 $K < \infty$，这里用到了$\|\mathbb{1}_{\{t+1\leq\sigma_n\}}\phi_{t+1}(\theta_t)\|_2 \leq K(2^n+1)\|Y_{t+1}^A + Y_{t+1}^b\|_2$ 对于某个 $K < \infty$ 按照（30）中相同的步骤。此外，对于最后一步，我们用到了在假设 1下 $\sup_t \|Y_{t+1}^A + Y_{t+1}^b\|_2 < \infty$。因此，我们得到一个增量方差可求和的鞅，从而$\sum_{t=1}^{k+1} 2\alpha_t \mathbb{1}_{\{t+1\leq\sigma_n\}}(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]$ 几乎必然收敛。

第二项： 现在我们关注项 $\sum_{t=1}^{\tau^k} 2\alpha_t\left(\phi_t(\theta_t) - \phi_t(\theta_{t-1})\right)$。等价地，我们可以研究

$$\sum_{t=1}^{k+1} \mathbb{1}_{\{t+1\leq\sigma_n\}} 2\alpha_t\left(\phi_t(\theta_t) - \phi_t(\theta_{t-1})\right)$$

利用以下事实： $\phi_t \in L_2$ 根据假设1

$$\phi_t(\theta_t) - \phi_t(\theta_{t-1}) = \sum_{k=0}^{\infty} E\left[e_{t+k}(\theta_t) - e_{t+k}(\theta_{t-1})|F_t\right]$$

$$= \sum_{k=0}^{\infty} E\left[(\delta_t - \delta_{t-1})^\intercal[A - A(Z_{t+k})]\theta_{t-1} + \delta_t^\intercal[A - A(Z_{t+k})](\theta_t - \theta_{t-1})|F_t\right]$$

注意，在事件 $t \leq \sigma_n$ 上，利用 A、$A(Z_t)$ b、$b(Z_t)$ 的有界性，我们有

$$\delta_t - \delta_{t-1} = \alpha_{t-1}(-A\delta_t + M_t) \leq \alpha_{t-1}K(1 + 2^{\frac{n}{2}})$$
$$\theta_t - \theta_{t-1} = \alpha_{t-1}(-A(Z_t)\theta_t + b(Z_t)) \leq \alpha_{t-1}K(1 + 2^{\frac{n}{2}}).$$

利用这些界，并按照 (30) 中完全相同的步骤，再根据假设1，可得存在某个一般常数 $K < \infty$ 使得

$$\|\mathbb{1}_{\{t+1\leq\sigma_n\}}\left(\phi_t(\theta_t) - \phi_t(\theta_{t-1})\right)\|_2 \leq \alpha_{t-1}K(1 + 2^n)\|Y_t^A + Y_t^b\|_2.$$

因此，我们写作

$$\lim_{k\to\infty} E\left[\sum_{t=1}^{k+1}\left|\mathbb{1}_{\{t+1\leq\sigma_n\}} 2\alpha_t\left(\phi_t(\theta_t) - \phi_t(\theta_{t-1})\right)\right|\right]$$
$$= E\left[\sum_{t=1}^{\infty}\left|\mathbb{1}_{\{t+1\leq\sigma_n\}} 2\alpha_t\left(\phi_t(\theta_t) - \phi_t(\theta_{t-1})\right)\right|\right]$$
$$= \sum_{t=1}^{\infty} 2\alpha_t E\left[\left|\mathbb{1}_{\{t+1\leq\sigma_n\}}\left(\phi_t(\theta_t) - \phi_t(\theta_{t-1})\right)\right|\right]$$
$$\leq \sum_{t=1}^{\infty} 2\alpha_t\alpha_{t-1}K(1 + 2^n)\|Y_t^A + Y_t^b\|_2 < \infty$$

最后一步用到了 $\|Y_t^A + Y_t^b\|_2$ 在 t 上的一致有界性。因此，我们可以得出 $\sum_{t=1}^{k+1} \left|\mathbb{1}_{\{t+1\leq\sigma_n\}} 2\alpha_t(\phi_t(\theta_t) - \phi_t(\theta_{t-1}))\right| < \infty$ 几乎必然成立，因而 $\sum_{t=1}^{k+1} \mathbb{1}_{\{t+1\leq\sigma_n\}} 2\alpha_t(\phi_t(\theta_t) - \phi_t(\theta_{t-1}))$几乎必然收敛。

第三项： 现在我们研究项 $\sum_{t=1}^{\tau^k-1} (\alpha_{t+1} - \alpha_t)\phi_{t+1}(\theta_t)$。

$$
\begin{aligned}
&E\left[\lim_{k\to\infty} \sum_{t=1}^{\tau^k-1} |(\alpha_{t+1} - \alpha_t)\phi_{t+1}(\theta_t)|\right] \\
&= E\left[\lim_{k\to\infty} \sum_{t=1}^{k} \mathbb{1}_{\{t+1\leq\sigma_n-1\}} |(\alpha_{t+1} - \alpha_t)\phi_{t+1}(\theta_t)|\right] \\
&= E\left[\sum_{t=1}^{\infty} \mathbb{1}_{\{t+1\leq\sigma_n-1\}} |(\alpha_{t+1} - \alpha_t)\phi_{t+1}(\theta_t)|\right] \\
&\leq \sum_{t=1}^{\infty} (\alpha_t - \alpha_{t+1}) E\left[\mathbb{1}_{\{t+1\leq\sigma_n\}} |\phi_{t+1}(\theta_t)|\right] \\
&\leq \sum_{t=1}^{\infty} (\alpha_t - \alpha_{t+1}) \|\mathbb{1}_{\{t+1\leq\sigma_n\}} \phi_{t+1}(\theta_t)\|_2 \\
&\leq K(2^n + 1) \sum_{t=1}^{\infty} (\alpha_t - \alpha_{t+1}) = K(2^n + 1)\alpha_1
\end{aligned}
$$

因此，当 $k \to \infty$ 时， $\sum_{t=1}^{\tau^k-1} (\alpha_{t+1} - \alpha_t)\phi_{t+1}(\theta_t)$ 几乎必然收敛。

最后一项： 最后，对于 $2\alpha_1\phi_1(\theta_0) - 2\alpha_{\tau^k}\phi_{\tau^k+1}(\theta_{\tau^k})$，我们有

$$
2\alpha_{\tau^k}\phi_{\tau^k+1}(\theta_{\tau^k}) \to \begin{cases} \alpha_{\sigma_n-1}\phi_{\sigma_n-1}(\theta_{\sigma_n-1}) & \text{if } \sigma_n < \infty \\ \lim_{k\to\infty} \alpha_{k+1}\phi_{k+2}(\theta_{k+1}) = 0, & \text{if } \sigma_n = \infty \end{cases}
$$

对于最后一部分，使用与之前类似的论证，我们可以证明

$$
E[\sum_{k=0}^{\infty} \mathbb{1}_{\{k+1\leq\sigma_n\}} (\alpha_k\phi_{k+1}(\theta_k))^2] < \infty
$$

这进而意味着，在 $\{\sigma_n = \infty\}$ 上， $\sum_{k=0}^{\infty}(\alpha_k\phi_{k+1}(\theta_k))^2 < \infty$ 几乎必然成立，并且 $\alpha_k\phi_{k+1}(\theta_k) \to 0$ 几乎必然成立。

最后一步： 到目前为止，我们已经证明 $\sum_{t=1}^{k+1} \mathbb{1}_{\{t+1\leq\sigma_n\}} 2\alpha_t\delta_t^\intercal M_t$ 几乎必然收敛。因此，这立即意味着 $\sum_{t=1}^{k+1} 2\alpha_t\delta_t^\intercal M_t$在事件 $\sigma_n = \infty$上几乎必然收敛，因为对于所有 t， $\mathbb{1}_{\{t+1\leq\sigma_n\}} = 1$ 在 $\sigma_n = \infty$ 上。

□

B 引理证明3

对于任意 $k > n$，我们有：

$$\mathbb{1}_{\{k+1\leq\sigma(C)\}}\left(\sum_{t=n}^{k}2\alpha_t\delta_t^\intercal M_t\right)^2 \leq \left(\sum_{t=n}^{k}\mathbb{1}_{\{t+1\leq\sigma(C)\}}2\alpha_t\delta_t^\intercal M_t\right)^2.$$

此外，将 τ^k 记为：$= k \wedge (\sigma(C) - 1)$。我们有

$$E\left[\sup_{k>n}\left(\sum_{t=n}^{k}\mathbb{1}_{\{t+1\leq\sigma(C)\}}2\alpha_t\delta_t^\intercal M_t\right)^2\right] = E\left[\sup_{k>n}\mathbb{1}_{\{\sigma(C)>n\}}\left(\sum_{t=n}^{\tau^k}2\alpha_t\delta_t^\intercal M_t\right)^2\right]$$

接着，我们写作

$$\begin{aligned}
&\sum_{t=n}^{\tau^k}2\alpha_t\delta_t^\intercal M_t = \sum_{t=n}^{\tau^k}2\alpha_t E[e_t(\theta_t)|F_t] \\
&= \sum_{t=n}^{\tau^k}2\alpha_t\left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right) + \sum_{t=n}^{\tau^k}2\alpha_t\left(\phi_t(\theta_t) - \phi_{t+1}(\theta_t)\right) \\
&= \sum_{t=n}^{\tau^k}2\alpha_t\left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right) \\
&\quad + \sum_{t=n}^{\tau^k}2\alpha_t\phi_t(\theta_t) - \sum_{t=n-1}^{\tau^k-1}2\alpha_{t+1}\phi_{t+1}(\theta_t) \\
&\quad + \sum_{t=n-1}^{\tau^k-1}2\alpha_{t+1}\phi_{t+1}(\theta_t) - \sum_{t=n}^{\tau^k}2\alpha_t\phi_{t+1}(\theta_t) \\
&= \sum_{t=n}^{\tau^k}2\alpha_t\left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right) \\
&\quad + \sum_{t=n}^{\tau^k}2\alpha_t\left(\phi_t(\theta_t) - \phi_t(\theta_{t-1})\right) \\
&\quad + \left(\sum_{t=n}^{\tau^k-1}2(\alpha_{t+1}-\alpha_t)\phi_{t+1}(\theta_t)\right)\mathbb{1}_{\{\sigma(C)>n+1\}} \\
&\quad + 2\alpha_n\phi_n(\theta_{n-1}) - 2\alpha_{\tau^k}\phi_{\tau^k+1}(\theta_{\tau^k})
\end{aligned}$$

我们分别分析这些项：

第一项： 对于第一项，我们首先回顾，对于任意 $k > n$

$$Z_n^k := \sum_{t=n}^{k} \mathbb{1}_{\{t+1 \le \sigma(C)\}} 2\alpha_t \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right)$$

是一个鞅序列。按照引理2 中的相同步骤，我们有

$$\begin{aligned}
E[(Z_n^k)^2] &= E\left[\left(\sum_{t=n}^{k} \mathbb{1}_{\{t+1 \le \sigma(C)\}} 2\alpha_t \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right)\right)^2\right] \\
&= \sum_{t=n}^{k} 4\alpha_t^2 E\left[\mathbb{1}_{\{t+1 \le \sigma(C)\}} \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right)^2\right] \\
&= \sum_{t=n}^{k} 16K\alpha_t^2 (C^2+1) \|Y_t^A + Y_t^b\|_2^2 \\
&\le 16K(C^2+1)\left(\sup_t \|Y_t^A + Y_t^b\|_2^2\right) \sum_{t=n}^{k} \alpha_t^2 = K'(C^2+1) \sum_{t=n}^{k} \alpha_t^2
\end{aligned}$$

对于某个 K， $K' < \infty$。因此，利用 Doob 最大不等式和单调收敛定理，我们可以写出

$$\begin{aligned}
E[\sup_{n<k} |Z_n^k|^2] &= \lim_{N\to\infty} E\left[\sup_{n<k<N} |Z_n^k|^2\right] \\
&\le 4 \sup_{n<k<N} E[|Z_n^k|^2] \le 4K'(C^2+1) \sum_{t=n}^{\infty} \alpha_t^2.
\end{aligned}$$

于是我们可以写成

$$\begin{aligned}
&E\left[\sup_{k>n} \mathbb{1}_{\{\sigma(C)>n\}} \left(\sum_{t=n}^{\tau^k} 2\alpha_t \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right)\right)^2\right] \\
&= E\left[\sup_{k>n} \left(\sum_{t=n}^{k} \mathbb{1}_{\{t+1 \le \sigma_n\}} 2\alpha_t \left(\phi_{t+1}(\theta_t) - E[\phi_{t+1}(\theta_t)|F_t]\right)\right)^2\right] \\
&= E\left[\sup_{k>n} |Z_n^k|^2\right] \le 4K'(C^2+1) \sum_{t=n}^{\infty} \alpha_t^2.
\end{aligned}$$

第二项：我们按照引理 2中的相同步骤，并写成

$$
\begin{aligned}
& E\left[\sup_{k>n} \mathbb{1}_{\{\sigma(C)>n\}}\left(\sum_{t=n}^{\tau^k} 2\alpha_t\left(\phi_t(\theta_t)-\phi_t(\theta_{t-1})\right)\right)^2\right] \\
&= E\left[\sup_{k>n}\left(\sum_{t=n}^{k} \mathbb{1}_{\{t+1\le\sigma(C)\}} 2\alpha_t\left(\phi_t(\theta_t)-\phi_t(\theta_{t-1})\right)\right)^2\right] \\
&\le E\left[\sup_{k>n}\left(\sum_{t=n}^{k} 4\alpha_t^2\right)\left(\sum_{t=n}^{k} \mathbb{1}_{\{t+1\le\sigma(C)\}}\left(\phi_t(\theta_t)-\phi_t(\theta_{t-1})\right)^2\right)\right] \\
&\le \sum_{t=n}^{\infty} 4\alpha_t^2 \sum_{t=n}^{\infty} E\left[\mathbb{1}_{\{t+1\le\sigma(C)\}}(\phi_t(\theta_t)-\phi_t(\theta_{t-1}))^2\right] \\
&= \sum_{t=n}^{\infty} 4\alpha_t^2 \sum_{t=n}^{\infty} \|\mathbb{1}_{\{t+1\le\sigma(C)\}}\left(\phi_t(\theta_t)-\phi_t(\theta_{t-1})\right)\|_2^2 \\
&\le \sum_{t=n}^{\infty} 4\alpha_t^2 \sum_{t=n}^{\infty} \alpha_{t-1}^2(1+C^2)\|Y_t^A+Y_t^b\|_2^2 \\
&\le K(1+C^2)\sum_{t=n}^{\infty}\alpha_t^2\sum_{t=n}^{\infty}\alpha_{t-1}^2
\end{aligned}
$$

对于某个常数 $K<\infty$。

第三项我们使用柯西-施瓦茨定理以及

$\mathbb{1}_{\{t+1\le\sigma(C)-1\}} \le \mathbb{1}_{\{t+1\le\sigma(C)\}}$ 来写出

$$
\begin{aligned}
&E\left[\sup_{k>n}\mathbb{1}_{\{\sigma(C)>n+1\}}\left(\sum_{t=n}^{\tau^k-1}2(\alpha_{t+1}-\alpha_t)\phi_{t+1}(\theta_t)\right)^2\right]\\
&\le E\left[\sup_{k>n}\left(\sum_{t=n}^{k-1}\mathbb{1}_{\{t+1\le\sigma(C)-1\}}2(\alpha_{t+1}-\alpha_t)\phi_{t+1}(\theta_t)\right)^2\right]\\
&\le E\left[\sup_{k>n}\left(\sum_{t=n}^{k-1}4(\alpha_t-\alpha_{t+1})\right)\left(\sum_{t=n}^{k-1}\mathbb{1}_{\{t+1\le\sigma(C)\}}(\alpha_t-\alpha_{t+1})\ \phi_{t+1}(\theta_t)^2\right)\right]\\
&\le \sum_{t=n}^{\infty}4(\alpha_t-\alpha_{t+1})\sum_{t=n}^{\infty}(\alpha_t-\alpha_{t+1})\left\|\mathbb{1}_{\{t+1\le\sigma(C)\}}\phi_{t+1}(\theta_t)\right\|_2^2\\
&\le K(C^2+1)\sum_{t=n}^{\infty}(\alpha_t-\alpha_{t+1})\sum_{t=n}^{\infty}(\alpha_t-\alpha_{t+1})\\
&\le K(C^2+1)\alpha_n^2
\end{aligned}
$$

其中我们用到了 $\left\|\mathbb{1}_{\{t+1\le\sigma(C)\}}\phi_{t+1}(\theta_t)\right\|_2^2 \le (C^2+1)K\|Y_t^A+Y_t^b\|_2^2$ 这一事实，并且根据假设，$\sup_t\|Y_t^A+Y_t^b\|_2^2<\infty$。

最后一项：

$$
\begin{aligned}
&E\left[\sup_{k<n}\mathbb{1}_{\{\sigma(C)>n\}}\left(2\alpha_n\phi_n(\theta_{n-1})-2\alpha_{\tau^k}\phi_{\tau^k+1}(\theta_{\tau^k})\right)^2\right]\\
&\le 4\alpha_n^2E\left[\mathbb{1}_{\{n+1\le\sigma(C)\}}\phi_n(\theta_{n-1})^2\right]+E\left[\sup_{k<n}\mathbb{1}_{\{\sigma(C)>n\}}\sum_{t=n}^{\tau^k}(\alpha_t\phi_{t+1}(\theta_t))^2\right]\\
&\le 4\alpha_n^2\|\mathbb{1}_{\{n\le\sigma(C)\}}\phi_n(\theta_{n-1})\|_2^2+E\left[\sup_{k<n}\sum_{t=n}^{k}\mathbb{1}_{\{t+1\le\sigma(C)\}}\alpha_t^2\phi_{t+1}(\theta_t)^2\right]\\
&\le K\alpha_n^2(1+C^2)+\sum_{t=n}^{\infty}\alpha_t^2\|\mathbb{1}_{\{t+1\le\sigma(C)\}}\phi_{t+1}(\theta_t)\|_2^2\\
&\le K\alpha_n^2(1+C^2)+K(1+C^2)\sum_{t=n}^{\infty}\alpha_t^2
\end{aligned}
$$

C 引理4 的证明

证明。 我们引入以下停止时（σ_n 已在（8）中介绍）：

$$\sigma_n := \inf\{t : \|\delta_t\|^2 > 2^n\}$$
$$\tau_n := 1 + \sup\{t < \sigma_{n+1} : \|\delta_t\|^2 \le 2^n\}.$$

利用 $\|M_t\|$ 上的界，使得 $\|M_t\| \le K(\|\delta_t\| + 1)$，其中某个 $K < \infty$，我们可以写成

$$\|\delta_{t+1}\|^2 - \|\delta_t\|^2 \le \alpha_t K(1 + \|\delta_t\|^2) + K\alpha_t^2(1 + \|\delta_t\|^2)$$

其中某个一般常数为 $K < \infty$。如果我们定义集合

$$C_n := \{\forall t \ge n : \|\delta_{t+1}\|^2 - \|\delta_t\|^2 \le \frac{1}{2}(\|\delta_t\|^2 + 1)\}$$

那么存在某个 $r < \infty$，使得 $P(C_n) = 1$ 对所有 $r \ge n$ 成立。

对于 C_r，我们有 $\|\delta_{t+1}\|^2 + 1 \le \frac{3}{2}(\|\delta_t\|^2 + 1)$ 对所有 $t \ge r$ 成立，因此可得对所有 $t \ge r$，$\|\delta_t\|^2 + 1 \le \frac{3}{2}^{t-r}(\|\delta_r\|^2 + 1)$。考虑

$$\{\sigma_n \le n\} = \{\sup_{r \le t \le n} \|\delta_t\|^2 \ge 2^n\} \cup \cup_{t=1}^{r-1}\{\|\delta_t\|^2 \ge 2^n\}.$$

注意，利用 $A(Z_t)$ 和 $b(Z_t)$ 上的界，对于每个固定的 $t < r$，$\lim_{n\to\infty} P(\|\delta_t\|^2 \ge 2^n) = 0$。于是我们有

$$\lim_{n\to\infty} P(C_r \cap (\sigma_n \le n)) = \lim_{n\to\infty} P(C_r \cap (\sup_{r \le t \le n} \|\delta_t\|^2 \ge 2^n))$$
$$\le \lim_{n\to\infty} P(\|\delta_r\|^2 + 1 \ge 2^n \frac{3}{2}^{r-n}) = 0.$$

由于，$P(C_r) = 1$，因此我们有

$$\lim_{n\to\infty} P(n < \sigma_n) = 1.$$

现在我们定义

$$B_n := C_n \cap (n < \sigma_n) \tag{31}$$

使得 $P(B_n) \to 1$。注意，在 $\{\sigma_{n+1} < \infty\}$ 上，$\|\delta_{\sigma_{n+1}}\|^2 \ge 2^{n+1}$。此外，在 B_n 上，$\tau_n \ge \sigma_n > n$，并且我们有 $\|\delta_{(\tau_n - 1)}\|^2 \le 2^n$。接着我们写作，

$$\|\delta_{\tau_n}\|^2 \le \frac{3}{2}\|\delta_{(\tau_n-1)}\|^2 + \frac{1}{2} \le \frac{3}{2}2^n + \frac{1}{2}.$$

由此可得，在 $B_n \cap(\sigma_{n+1} < \infty)$ 上

$$\|\delta_{\sigma_{n+1}}\|^2 - \|\delta_{\tau_n}\|^2 \geq 2^{n+1} - \frac{3}{2}2^n - \frac{1}{2} \geq \frac{2^n}{4} \tag{32}$$

对所有 $n \geq 2$。

现在我们关注上界。利用式（7），在$B_n \cap (\sigma_{n+1} < \infty)$上，我们有

$$\begin{aligned}
\|\delta_{\sigma_{n+1}}\|^2 - \|\delta_{\tau_n}\|^2 &\leq \sum_{t=\tau_n}^{\sigma_{n+1}-1} K\alpha_t^2\|\delta_t\|^2 + \sum_{t=\tau_n}^{\sigma_{n+1}-1} 2\alpha_t\delta_t^\intercal M_t \\
&\leq 2^{n+1}\sum_{t=n}^{\infty} K\alpha_t^2 + \sup_{n<t<\sigma_{n+1}} \left|\sum_{k=t}^{\sigma_{n+1}-1} 2\alpha_k\delta_k^\intercal M_k\right| \\
&\leq 2^{n+1}\sum_{t=n}^{\infty} K\alpha_t^2 + 2\sup_{n<t\leq\sigma_{n+1}} \left|\sum_{k=n}^{t-1} 2\alpha_k\delta_k^\intercal M_k\right| \\
&\leq 2^{n+1}\sum_{t=n}^{\infty} K\alpha_t^2 + 2\sup_{n<t} \mathbb{1}_{\{t\leq\sigma_{n+1}\}} \left|\sum_{k=n}^{t-1} 2\alpha_k\delta_k^\intercal M_k\right|
\end{aligned}$$

根据引理 3，我们有

$$\begin{aligned}
&E\left[\sup_{t>n} \mathbb{1}_{\{t\leq\sigma_{n+1}\}} \left(\sum_{k=n}^{t-1} 2\alpha_k\delta_k^\intercal M_k\right)^2\right] \\
&= E\left[\sup_{t>n-1} \mathbb{1}_{\{t+1\leq\sigma_{n+1}\}} \left(\sum_{k=n}^{t} 2\alpha_k\delta_k^\intercal M_k\right)^2\right] \\
&\leq E\left[\sup_{t>n} \mathbb{1}_{\{t+1\leq\sigma_{n+1}\}} \left(\sum_{k=n}^{t} 2\alpha_k\delta_k^\intercal M_k\right)^2\right] + E\left[\mathbb{1}_{\{n+1\leq\sigma_{n+1}\}} \left(2\alpha_n\delta_n^\intercal M_n\right)^2\right] \\
&\leq K(1+2^{2n+2})\sum_{k=n}^{\infty}\alpha_k^2 + K(1+2^{2n+2})\alpha_n^2 \leq K(1+2^{2n+2})\sum_{k=n}^{\infty}\alpha_k^2
\end{aligned}$$

其中我们使用了一个通用的 $K < \infty$，它可能在不同步骤中发生变化。由此可得

$$E\left[\mathbb{1}_{B_n\cap(\sigma_{n+1}<\infty)}(\|\delta_{\sigma_{n+1}}\|^2 - \|\delta_{\tau_n}\|^2)^2\right] \leq K2^{2n}\left(\sum_{t=n}^{\infty}\alpha_t^2\right)^2 + K2^{2n}\sum_{t=n}^{\infty}\alpha_t^2$$

对于某个常数 $K < \infty$。结合这个界与（32），我们可以写成

$$K\left(\sum_{t=n}^{\infty}\alpha_t^2\right)^2 + K\sum_{t=n}^{\infty}\alpha_t^2 \geq 2^{-2n}E\left[\mathbb{1}_{B_n\cap(\sigma_{n+1}<\infty)}(\|\delta_{\sigma_{n+1}}\|^2 - \|\delta_{\tau_n}\|^2)^2\right]$$
$$\geq \frac{1}{16}P(B_n \cap (\sigma_{n+1} < \infty)).$$

注意到 $P(B_n) \to 1$ (见 (31))，并且 $\sum_{t=n}^{\infty}\alpha_t^2 \to 0$，因此我们得出 $P(\sigma_n < \infty) \to 0$. □

参考文献

[1] Albert Benveniste、Michel M´etivier 和 Pierre Priouret。自适应算法与随机逼近，第22卷。Springer Science & Business Media，2012年。[2] V. S. Borkar 和 S. P. Meyn。用于随机逼近与强化学习收敛的常微分方程方法。《*SIAM* 控制与优化杂志》，38(2):447–469，2000年。[3] Qi Cai、Zhuoran Yang 和 Zhaoran Wang。部分观测下的强化学习：具有可证明样本效率的线性函数逼近。载于国际机器学习会议，第2485–2522页。PMLR，2022年。[4] S. Chandak、V.S. Borkar 和 P. Dodhia。非马尔可夫环境中的强化学习。《系统与控制快报》，185:105751，2024年。[5] S. Dong、B. van Roy 和 Z. Zhou。简单智能体，复杂环境：利用智能体状态实现高效强化学习。《机器学习研究杂志》，23(1):11627–11680，2022年。[6] E.A. Feinberg、P.O. Kasyanov 和 N.V. Zadioanchuk。转移概率弱连续的平均成本马尔可夫决策过程。《数学运筹研究》，37(4):591–607，2012年11月。[7] O. Hernandez-Lerma。自适应马尔可夫控制过程，第79卷。Springer Science & Business Media，2012年。[8] Chi Jin、Zhuoran Yang、Zhaoran Wang 和 Michael I Jordan。具有线性函数逼近的可证明高效强化学习。《运筹学数学》，48(3):1496–1521，2023年。

[9] A. D. Kara、N. Saldi 和 S. Y¨uksel。非线性滤波器的弱费勒性质。《系统与控制快报》，134:104–512，2019年。[10] A. D. Kara 和 S. Yuksel。部分可观测马尔可夫决策过程有限记忆Q学习的收敛性，以及滤波器稳定性下学习策略的近似最优性。*arXiv*预印本 *arXiv:2103.12158*，2021年。[11] A. D. Kara 和 S. Yuksel。部分可观测马尔可夫决策过程中的有限记忆反馈策略的近似最优性。《机器学习研究杂志》，23(1):1–46，2022年。[12] Ali Kara、Naci Saldi 和 Serdar Y¨uksel。一般空间上的马尔可夫决策过程Q学习：弱连续性下通过量化实现收敛与近似最优性。《机器学习研究杂志》，24(199):1–34，2023年。[13] Ali D. Kara、Erhan Bayraktar 和 Serdar Y¨uksel。连续空间部分可观测马尔可夫决策过程的近似最优近似与有限记忆策略。《系统科学与复杂性杂志》，38:238–270，2025年。[14] Ali Devran Kara 和 Serdar Y¨uksel。部分可观测马尔可夫决策过程有限记忆Q学习的收敛性，以及滤波器稳定性下学习策略的近似最优性。《运筹学数学》，48(4):2066–2093，2023年。[15] Ali Devran Kara 和 Serdar Yuksel。一般信息结构和非马尔可夫环境下随机控制的Q学习。《机器学习研究汇刊》，2024年。特色认证。[16] C. McDonald 和 S. Y¨uksel。通过多布鲁申系数实现指数滤波器稳定性。《概率电子通信》，25，2020年。[17] F. C. Melo、S. P. Meyn 和 I. M. Ribeiro。函数逼近强化学习分析。载于第*25*届国际机器学习会议论文集，第664–671页，2008年。[18] Sean Meyn。强化学习中的投影贝尔曼方程。*IEEE* 自动控制汇刊，2024年。[19] Emmanuel Rio。强混合过程的协方差不等式。载于《*IHP* 概率年刊 ´与统计学，第29卷，第587–597页，1993年。

[20] Andrzej Ruszczy´nski 和 Shangzhe Yang。用于非凸非光滑条件随机优化的函数模型方法。*SIAM* 优化期刊，34(3):3064–3087，2024年。[21] N. Saldi、T. Linder 和 S. Y¨uksel。离散时间随机控制中的有限逼近：量化模型与渐近最优性。施普林格，Cham，2018年。[22] S. P. Singh、T. Jaakkola 和 M. I. Jordan。部分可观测马尔可夫决策过程中的无状态估计学习。机器学习会议论文集 *1994*，第284–292页，1994年。[23] Amit Sinha、Matthieu Geist 和 Aditya Mahajan。用于部分可观测马尔可夫决策过程的周期性基于智能体状态的Q学习。神经信息处理系统进展，37:62123–62159，2024年。[24] J. N. Tsitsiklis 和 B. Van Roy。函数逼近下时序差分学习的分析。*IEEE* 自动控制汇刊，42(5):674–690，1997年。