arXiv:2603.04226v1 [math.OC] 4 Mar 2026

# 具有有限可加估值的马尔可夫决策过程中的最优策略*

János Flesch† Arkadi Predtetchinski‡ William Sudderth§
Xavier Venel¶

2026年3月5日

**摘要**

我们研究无限期马尔可夫决策过程（马尔可夫决策过程），其中决策者通过聚合无限的预期阶段奖励流来评估其每一策略。我们方法的关键特征在于，聚合是借助阶段集合上给定的弥散电荷（一种弥散的有限可加概率测度）来完成的。Neyman [2023]的结果表明，在此设定下，对于每一个具有有限状态空间和行动空间的马尔可夫决策过程，只要该弥散电荷满足货币时间价值原理，决策者便拥有一个纯最优策略。他的结果提出了这样一个问题：在不对聚合电荷作额外假设的情况下，最优策略是否存在。我们以一个反例给出否定回答。在精心构造的聚合电荷下，该马尔可夫决策过程根本不存在最优策略，无论是纯策略还是随机化策略。

**关键词：马尔可夫决策过程、有限可加测度、最优策略。**

## 1 引言

马尔可夫决策过程（MDP）是最常用于描述和研究动态决策的模型之一。关于马尔可夫决策过程已有非常丰富的文献，并且它们已应用于各种情境（例如 Puterman [1990]）。

*我们感谢 Fabio Maccheroni 提供了非常有益的讨论。

†定量经济学系，马斯特里赫特大学，邮政信箱 616，6200 MD，荷兰。电子邮箱：jflesch@maastrichtuniversity.nl

‡微观经济学与公共经济学系，马斯特里赫特大学，邮政信箱 616，6200 MD，荷兰。电子邮箱：a.predtetchinski@maastrichtuniversity.nl §统计学院，明尼苏达大学，美国明尼苏达州明尼阿波利斯，MN 55455。电子邮箱：bill@stat.umn.edu

¶人工智能、数据与决策科学系，路易斯大学，意大利罗马 Viale Romania 32，00197。电子邮箱：xvenel@luiss.it

我们研究具有有限状态空间和行动空间的马尔可夫决策过程，其阶段集合为 $\mathbb{N} = \{1, 2, 。\}$。在每个阶段，马尔可夫决策过程的运行处于某个状态，决策者在此选择某个行动。根据所选行动，决策者会获得即时奖励，随后按照某个概率分布选定下一个状态，并在下一阶段继续运行。

决策者的每项策略都会为每个阶段定义一个期望奖励，从而确定一个无限的期望奖励序列。我们假设，决策者通过弥散电荷——即定义在阶段集合 $\mathbb{N}$ 上、且每个阶段的权重均为零的有限可加概率测度——将相应的无限期望奖励序列聚合为一个称为收益的实数，并据此评价每项策略。我们将这种弥散电荷称为聚合电荷。

这一框架的一个特例是众所周知的具有长期平均奖励的马尔可夫决策过程类别。当聚合电荷对阶段集合的每个子集赋予其极限频率（只要该极限频率存在）时，就得到这一特例（参见例 1）。

另一个重要特例是聚合电荷满足货币时间价值原理的马尔可夫决策过程类别。Neyman [2023],研究了这一类别，不过在他的论文中，聚合电荷是用线性泛函来表述的（参见第2节）。Neyman [2023] 证明，在每个马尔可夫决策过程中，决策者都存在一个纯平稳策略，该策略对于所有满足货币时间价值原理的聚合电荷都是最优的（参见定理 2）。

当聚合收费不满足货币时间价值原理时，可能出现这样的情况：决策者拥有一个纯最优策略，但没有任何平稳策略能够实现最优性（参见例4）。因此，一个自然产生的问题是：对于每个带有任意聚合电荷的马尔可夫决策过程，决策者是否都拥有一个最优策略？在本文中，我们通过一个反例对此问题给出了否定答案（参见定理3）。我们构造了一个具有确定性转移和精心构造的弥散电荷的马尔可夫决策过程，使得决策者根本不存在最优策略，无论是纯策略还是随机化策略。

本文结构如下。在第2节中，我们讨论一些预备知识。在第3节中，我们定义模型。在第4节中，我们回顾Neyman [2023], 的结果，并陈述我们的主要结果——定理3。最后，在第5节中，我们作出一些结论。

# 2 预备知识

在整篇论文中，我们假设ZFC公理。令 $\mathbb{N} = \{1, 2, .....\}$ 表示自然数集。

**电荷**。定义在 $\mathbb{N}$ 上的电荷，也称为有限可加概率测度，是一个映射$\mu$：$2^{\mathbb{N}} \to [0, 1]$，满足 $\mu(\varnothing) = 0$，$\mu(\mathbb{N}) = 1$，且 $\mu$ 满足有限可加性：对于所有互不相交的 $M_1 \subseteq \mathbb{N}$ 和 $M_2 \subseteq \mathbb{N}$，都有 $\mu(M_1 \cup M_2) = \mu(M_1) + \mu(M_2)$。注意，

电荷 $\mu$ 定义在整个幂集 $2^{\mathbb{N}}$ 上，这样设定不失一般性，因为定义在 $\mathbb{N}$ 的代数上的每个电荷都可以扩展到整个幂集上（通常这种扩展并不唯一）。我们用 $\Delta^f$ 表示 $\mathbb{N}$ 上电荷的集合。关于电荷的详细讨论，请参阅 Rao and Rao [1983]。

电荷 $\mu \in \Delta^f$ 称为弥散，如果[1] $\mu(n) = 0$ 对每个 $n \in \mathbb{N}$ 都成立。我们用 $\Delta^d$ 表示 $\mathbb{N}$ 上弥散电荷的集合。集合 $\Delta^d$ 是 $\Delta_f$ 的真子集。

弥散电荷 $\mu \in \Delta^d$ 称为频率电荷，如果

$$\mu(M) = \lim_{n\to\infty} \frac{|\{1,\ldots,n\} \cap M|}{n} \tag{1}$$

对每个集合 $M \subseteq \mathbb{N}$，右侧的极限都存在。此外，弥散电荷 $\mu \in \Delta^d$ 称为平移不变，如果 $\mu(M) = \mu(M+1) = \mu$ 对所有 $M \subseteq \mathbb{N}$ 均成立，其中 $M+1 = \{n+1 : n \in M\}$。频率电荷确实存在（参见 Kadane and O'Hagan 中的定理3 [1995]），并且它们构成平移不变电荷集合的真子集（参见 Schirokauer and Kadane 中的命题3.3及非平凡例4.8 [2007]）。

关于电荷的积分。设有一个电荷 $\mu \in \Delta^f$。如果对于某个 $m \in \mathbb{N}$，存在 $c_1, \ldots, c_m \in \mathbb{R}$ 和一个将 $\mathbb{N}$ 划分为 $M_1, \ldots, M_m\}$ 的划分{，使得 $s = \sum_{i=1}^m c_i \mathbb{I}_{M_i}$，其中 $\mathbb{I}_{M_i}$ 是集合 $M_i$ 的示性函数，则称函数 $s : \mathbb{N} \to \mathbb{R}$ 为简单函数。关于 $\mu$，$s$ 的积分定义为 $\int_{t\in\mathbb{N}} s(t)\mu(dt) = \sum_{i=1}^m c_i \cdot \mu(M_i)$。

对于有界函数 $f : \mathbb{N} \to \mathbb{R}$，积分 $\int_{t\in\mathbb{N}} f(t)\mu(dt)$ 定义为所有实数 $\int_{t\in\mathbb{N}} s(t)\mu(dt)$ 的上确界，其中 $s$ 是简单函数，且 $s \leq f$。这等价于取那些实数的下确界，其中 $s$ 是简单函数，且 $s \geq f$。由于 $f$ 有界，积分是有限的。

**满足货币时间价值原理的电荷**。现在我们将其与 Neyman [2023] 中的设定联系起来。如下文所述，Neyman [2023] 中定义的耐心估值与满足

$$\liminf_{n\to\infty} \frac{|\{1,\ldots,n\} \cap M|}{n} \leq \mu(M) \leq \limsup_{n\to\infty} \frac{|\{1,\ldots,n\} \cap M|}{n} \tag{2}$$

对每个 $M \subseteq \mathbb{N}$ 的弥散电荷类之间存在自然的一一对应关系。

Neyman [2023, 定义 3] 将估值定义为函数 $w : \ell^\infty \to \mathbb{R}$，其中 $\ell^\infty$ 是 $\mathbb{R}$ 中所有有界无限序列的集合，满足：(i) $w$ 经过归一化：$w(1, 1, \ldots) = 1$；(ii) $w$ 具有可加性：对于所有 $g, g' \in \ell^\infty$，$w(g + g') = w(g) + w(g')$；(iii) $w$ 遵循货币时间价值原理：对于所有 $g = (g_1, g_2, \ldots) \in \ell^\infty$，并且

[1]稍微滥用一下记号，我们用 $\mu(n)$ 表示单点集 $\{n\}$ 的 $\mu$ 测度。[2]这类电荷是真正的频率电荷类的子集。事实上，根据公式（2）可得包含关系，而 Schirokauer and Kadane[2007] 中的例4.5证明了该包含关系是严格的。

$g' = (g'_1, g'_2, \ldots) \in \ell^\infty$，如果部分和满足 $g_1 + \cdots + g_n \geq g'_1 + \cdots + g'_n$ 对每个$n \in \mathbb{N}$均成立，那么 $w(g) \geq w(g')$。

估值$w: \ \ell^\infty \to \mathbb{R}$ 在 Neyman [2023, 定义 6] 中被称为耐心，如果对每个序列$g \in \ell^\infty$，存在某个 $k \in \mathbb{N}$，使得对所有$n \leq k$ 有$g_n = 0$，并且对所有$n > k$有$g_n = 1$，则$w(g) = 1$。根据 Neyman [2023], 中定理 2 的刻画，函数$w: \ \ell^\infty \to \mathbb{R}$ 是耐心估值，当且仅当$w$ 是线性的并且满足

$$\liminf_{n\to\infty} \frac{g_1 + \cdots + g_n}{n} \ \leq \ w(g) \ \leq \ \limsup_{n\to\infty} \frac{g_1 + \cdots + g_n}{n} \tag{3}$$

对每个$g = (g_1, g_2, \ldots) \in \ell^\infty$。由此可知，每个耐心估值关于 $||\cdot||_\infty$-范数是连续的。

对每个耐心估值$w: \ \ell^\infty \to \mathbb{R}$，由于$w$ 关于 $||\cdot||_\infty$-范数是线性且连续的，存在唯一的弥散电荷 $\mu \in \Delta^d$ 使得

$$w(g) \ = \ \int_{n\in\mathbb{N}} g_n \ \mu(dn) \tag{4}$$

对所有$g \in \ell^\infty$；例如参见[3] Rao and Rao [1983]中的定理 4.7.4。注意，该电荷 $\mu$ 由于式 (3)而满足式(2)。

总之，每个耐心估值$w$ 都可以表示为关于满足式 (2) 的弥散电荷 $\mu \in \Delta^d$ 的式 (4) 中的积分。反之亦然：如果 $\mu \in \Delta^d$满足式 (2)，那么由式 (4) 给出的$w$ 就是一个耐心估值。

**电荷集合上的逐点收敛拓扑**。我们将考虑电荷集合 $\Delta_f$上的一种特定拓扑：逐点收敛拓扑。这是[4]在 $\Delta_f$上使得对每个$M \subseteq \mathbb{N}$，由 $\phi_M: \ \Delta^f \to [0,1]$ 定义的映射 $\phi_M(\mu) = \mu(M)$ 连续的最小拓扑。这个拓扑空间是紧致且豪斯多夫的，但不是可度量化的。

$\Delta^f$中的一个网是一个集合 $\{\mu_\alpha\}_{\alpha\in\Omega}$，其中$(\Omega, \succeq)$是一个有向集，并且对于每个 $\alpha \in \Omega$都有 $\mu_\alpha \in \Delta_f$。我们将主要关注 $\Delta_f$中的序列，这是 $\Delta_f$中网的一个特例，其中有向集为$(\mathbb{N}, \geq)$。

考虑 $\Delta_f$中的一个序列 $\{\mu_n\}_{n\in\mathbb{N}}$。电荷 $\mu \in \Delta^f$是$\{\mu_n\}_{n\in\mathbb{N}}$ 的一个聚点，如果对 $\mathbb{N}$ 的子集的每一个有限族 $\{M_1, \ldots, M_k\}$ 以及对每一个 $\epsilon > 0$，存在一个序列$m_1 < m_2 < \cdots$ 使得

$$\forall n \in \mathbb{N}, \ \forall i = 1, \ldots, k: \quad |\mu(M_i) - \mu_{m_n}(M_i)| \ \leq \ \epsilon.$$

[3]根据 Rao and Rao [1983], 中的记号，该定理的应用如下： $\Omega := \mathbb{N}$， $\mathscr{F} := 2^{\mathbb{N}}$，以及 T: =w。于是，根据 Rao and Rao [1983], 中的定义 4.7.1，空间 $\mathscr{C}(\Omega, \mathscr{F})$ 等于 $\ell^\infty$，因为 $\mathscr{F}$ 包含 Ω的所有子集。因此，Rao and Rao [1983] 中定理 4.7.4 的方程表明，对于唯一的有界有限可加符号测度 $\mu$，式 (4) 成立。由于当 g $_n \geq 0$ 时，对所有 n $\in \mathbb{N}$均有 T(g) = w(g) $\geq 0$，定理 4.7.4 的最后一个结论意味着，$\mu$ 只取非负值。又由于根据估值的性质 (i)，T(1,1, . . .) =w(1, 1, . . .) = 1，我们得到 $\mu(\mathbb{N}) = 1$。最后，$\mu$ 是弥散的，因为 T = w 是耐心的。

[4]该拓扑与 $\Delta^f$ 作为紧积空间 $[0,1]^{(2^{\mathbb{N}})}$ 的闭子空间所继承的拓扑相同。

等价地，如果序列 $\{\mu_n\}_{n\in\mathbb{N}}$ 有一个收敛到 $\mu$ 的子网。由于 $\Delta^f$ 是紧致的，$\Delta_f$ 中的每个序列 $\{\mu_n\}_{n\in\mathbb{N}}$ 至少有一个聚点。

# 3 模型

## 3.1 马尔可夫决策过程

**马尔可夫决策过程（MDP）的定义**。MDP 由以下部分构成：

- 一个非空有限状态空间 $S$，其中给定状态 $s^* \in S$ 作为初始状态。
- 对每个状态 $s \in S$，都有一个非空有限行动空间 $A(s)$ 。
- 对每个状态 $s \in S$ 和动作 $a \in A(s)$，都有一个有界奖励 $r(s,a) \in \mathbb{R}$ 。
- 对于状态空间$S$中的每个状态，状态空间$s$ 上的概率测度$p(s,a) = (p(z|s,a))_{z\in}$ e $s \in S$ 以及动作$a \in A(s)$。

马尔可夫决策过程在无限时域上进行，各阶段 $\mathbb{N}$如下：在阶段1，处于给定的初始状态 $s_1 = s^*$时，决策者选择一个动作 $a_1 \in A(s_1)$。随后，她获得奖励 $r(s_1,a_1)$，接着根据概率测度 $p(s_1,a_1)$转移到状态 $s_2 \in S$。在阶段2，处于状态 $s_2$, 时，决策者再次选择一个动作 $a_2 \in A(s_2)$。随后，她获得奖励 $r(s_2,a_2)$，接着根据概率测度 $p(s_2,a_2)$转移到状态 $s_3 \in S$，依此类推。

如果对于每个状态$s \in S$ 和每个动作$a \in A(s)$ ，都存在一个状态$s' \in S$ 使得$p(s'|s,a) = 1$，则称马尔可夫决策过程中的转移是确定性的 。

**策略**。 阶段 $t \in \mathbb{N}$ 的历史 是一个序列$(s_1 = s^*, a_1,\ldots,s_{t-1}, a_{t-1},s_t)$ ，满足$s_i \in S$ 对所有$i = 1,\ldots, t$ 均成立，且 $a_i \in A(s_i)$ 对所有 $i = 1,\ldots, t-1$ 均成立。

（行为）策略 $\sigma$ 是一种映射，它为任意阶段$t \in \mathbb{N}$的每个历史$h = (s_1, a_1,\ldots,s_{t-1}, a_{t-1},s_t)$指定行动空间$A(s_t)$上的一个概率测度 $\sigma(h)$ 。其含义是，若出现历史$h$ ， $\sigma$ 建议按照 $\sigma(h)$ 随机选择一个动作。所有策略的集合记作 $\Sigma$。

如果策略 $\sigma$ 对任意历史$h$都使 $\sigma(h)$ 以概率 1 选择某个动作，则称该策略为纯策略；否则称其为随机化策略。若 $\sigma(h)$ 只通过当前状态$s_t$依赖于历史$h = (s_1, a_1,\ldots,s_{t-1}, a_{t-1}, s_t)$ ，则称策略 $\sigma$ 为平稳策略；也就是说，平稳策略可以看作一个集合$(\sigma_s)_{s\in S}$ ，其中 $\sigma_s$ 是$A(s)$上的概率测度。

## 3.2 聚合电荷

我们在本文中假设，决策者在马尔可夫决策过程中获得的奖励由给定的弥散电荷 $\mu \in \Delta^d$ 聚合，我们也将其称为聚合电荷

电荷。配有聚合电荷 $\mu \in \Delta^d$ 的马尔可夫决策过程记作 MDP($\mu$)。

**策略所诱导的收益**。考虑一个马尔可夫决策过程($\mu$)。Neyman 提出了以下估值 [2023], ，尽管形式略有不同（参见第2节）。

自然的做法是用收益来评估每个策略 $\sigma$，其定义如下:

$$u_\mu(\sigma) := \int_{t\in\mathbb{N}} \mathbb{E}_\sigma[r_t]\ \mu(dt), \tag{5}$$

其中$r_t$ 表示阶段$t \in \mathbb{N}$的奖励随机变量，而 $\mathbb{E}_\sigma$ 表示在 $\sigma$下的期望算子。也就是说，收益$u^\mu(\sigma)$ 是所有阶段期望奖励的 $\mu$加权聚合，或 $\mu$加权平均。注意，对于任意策略 $\sigma$ 和任意聚合电荷 $\mu$，$u^\mu(\sigma)$都有定义，因为对于任意策略 $\sigma$，序列$\{\mathbb{E}_\sigma[r_t]\}_{t\in\mathbb{N}}$ 都有界，因此该序列关于任意电荷 $\mu$ 均可积（参见第2节）。

**例1** 如果 $\mu \in \Delta^d$ 是频率电荷，那么公式（5）中定义的收益$u^\mu$ 对应于具有长期平均奖励的马尔可夫决策过程。例如，若策略 $\sigma$ 在每个奇数阶段诱导的期望奖励为1，在每个偶数阶段诱导的期望奖励为0，则根据公式（1），有$u^\mu(\sigma) = \mu(\{1,3,5,...\}) = \frac{1}{2}$。 ♦

**值与最优策略**。 对于公式（5）中的收益，马尔可夫决策过程($\mu$)的值 定义为

$$v_\mu := \sup_{\sigma\in\Sigma} u_\mu(\sigma).$$

策略 $\sigma \in \Sigma$ 称为最优，如果$u^\mu(\sigma) = v^\mu$。

# 4 最优策略不存在

我们的出发点是以下结果，它可由 Neyman 的定理 6 应用于弥散电荷时得出[2023] 。

**定理 2（Neyman [2023, 定理 6]）**考虑一个具有有限状态空间和行动空间的马尔可夫决策过程。那么，对于所有满足公式（2）的聚合电荷，决策者都有[5] 一个纯平稳策略 $\sigma$，使得 $\sigma$ 在该马尔可夫决策过程($\mu$) 中是最优的 $\mu \in \Delta^d$ 。

请注意，定理 2 中的策略 $\sigma$ 具有稳健性：只要 $\mu$ 满足公式（2）， $\sigma$ 就不取决于聚合电荷 $\mu$的选择。

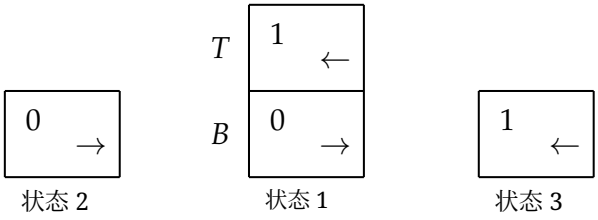
[5]可以为 $\sigma$ 选择任意 Blackwell最优策略，即选择任意一个纯平稳策略，使其在所有充分小的折扣率下均为最优。（在 Neyman 定理 6 的证明中 [2023] ，并采用其中的记号， $\omega_t = 0$ 对所有阶段$t$均成立，因此对任意 $\epsilon > 0$，都可以选择$t_\epsilon = 1$，这意味着所构造的策略 $\tau$ 与一个 Blackwell最优策略一致。）

若允许使用不满足公式（2）的聚合收费 $\mu$，不难找到这样的马尔可夫决策过程：决策者在马尔可夫决策过程$(\mu)$中存在纯最优策略，但没有任何平稳策略是最优的；参见例4。

自然会产生这样的问题：对于每个马尔可夫决策过程和每个聚合电荷 $\mu$，决策者在马尔可夫决策过程$(\mu)$中是否有最优策略？这引出了本文的主要贡献。

**奇偶 马尔可夫决策过程**( $\mu$ $\mu \in$ )：
首先定义该马尔可夫决策过程，然后定义聚合电荷。

考虑如下马尔可夫决策过程：(i) 状态空间为$S = \{1,2,3\}$；(ii) 状态 1 中有两个动作$T$和$B$，状态 2 和状态 3 中各只有一个动作；(iii) 状态 1 中，动作$T$ 的奖励为 1，动作$B$的奖励为 0；状态 2 中的奖励为 0，状态 3 中的奖励为 1；(iv) 动作$T$ 会转移到状态 2，动作$B$ 会转移到状态 3，而从状态 2 和状态 3 都会转移到状态 1。该马尔可夫决策过程如下图所示：




假设初始状态是状态 1。本质上，这些阶段可以两两分组：阶段 1 和 2、阶段 3 和 4、阶段 5 和 6，等等。在每个奇数阶段，决策者选择当前阶段和下一个阶段的收益，其选项只有两种：现在获得 1 而下一阶段获得 0，或者相反，现在获得 0 而下一阶段获得 1。注意，这些转移是确定性的。

现在我们定义电荷 $\mu$。$\mu$ 的构造相当复杂，其细节对我们的证明至关重要。

设 $\phi \in \Delta^d$ 为一个频率电荷（参见第2节）；注意此时 $\phi$ 是平移不变的。设 $E_0 = \{1,3,5,\ldots\}$，并且对每个 $n \in \mathbb{N}$，设 $E_n = \{k \cdot 2^n : k \in \mathbb{N}\}$。因此，例如，$E_1 = \{2,4,6,8,10,12,14,16,\ldots\}$，$E_2 = \{4,8,12,16,\ldots\}$，$E_3 = \{8,16,\ldots\}$；一般地，对于 $n \geq 1$，集合 $E_{n+1}$ 由集合 $E_n$中每隔一个元素组成。因此，序列 $E_1, E_2, E_3, \ldots$ 关于集合包含关系是递减的，并且 $E_1 \cap E_2 \cap E_3 \cap \cdots = \emptyset$。

定义弥散电荷 $\mu_0 \in \Delta^d$ 如下：对于每个集合 $W \subseteq \mathbb{N}$，令 $\mu_0(W) = 2 \cdot \phi(W \cap E_0)$。这本质上是将电荷 $\phi$ 应用于奇数集；特别地，$\mu_0(E_0) = 2 \cdot \phi(E_0) = 1$。类似地，对于每个 $n \in \mathbb{N}$，定义弥散电荷 $\mu_n \in \Delta^d$ 如下：对于每个集合 $W \subseteq \mathbb{N}$，令$\mu_n(W) = 2^n \cdot \phi(W \cap E_n)$。因此，电荷 $\mu_n$ 就是将电荷 $\phi$ 应用于集合 $E_n$；特别地，$\mu_n(E_n) = 2^n \cdot \phi(E_n) = 1$。

我们赋予 $\mathbb{N}$ 上的电荷集合 $\Delta^f$逐点收敛拓扑（参见第2节）。由于 $\Delta_f$是一个紧空间，因此序列 $\{\mu_n\}_{n=1}^{\infty}$作为一个网，存在一个

聚点[6] $\mu_* \in \Delta^f$。由于该序列的每个元素都是弥散电荷，因此我们有 $\mu_* \in \Delta^d$。

对每个$n \in \mathbb{N}$，都有 $\mu_*(E_n) = 1$。事实上，只要$m \geq n$，就有$E_n \cap E_m = E_m$，因此 $\mu_m(E_n) = 2^m \cdot \phi(E_n \cap E_m) = 2^m \cdot \phi(E_m) = 1$。

现在令聚合电荷为 $\mu = \frac{1}{2}\mu_0 + \frac{1}{2}\mu_*$。显然，$\mu \in \Delta^d$，因为 $\mu_0 \in \Delta^d$且 $\mu_* \in \Delta^d$。

**直观解释：** 这个马尔可夫决策过程的思路如下。如上所述，在每个奇数阶段，决策者必须在动作$T$（当前阶段获得奖励 1，下一阶段获得奖励 0）和动作$B$（当前阶段获得奖励 0，下一阶段获得奖励 1）之间作出选择。

针对电荷 $\mu_0$ 和电荷 $\mu_*$采取良好策略之间存在矛盾。电荷 $\mu_0$ 集中在奇数阶段集合上，而 $\mu_*$ 集中在偶数阶段集合上。为了针对 $\mu_0$采取良好策略，决策者需要尽可能频繁地选择动作$T$。相比之下，为了针对电荷 $\mu_*$采取良好策略，决策者只需以正频率选择动作$B$即可。事实上，对于$n \geq 1$，任一集合$E_n$的频率均为$\frac{1}{2^n}$，但其 $\mu_*$-测度为 1。

—

为了在 $\mu$上实现最优，决策者必须就其两个分量 $\mu_0$和 $\mu_*$都表现良好。对于任意 $\epsilon \in (0, 1)$，以频率 $1 - \epsilon$ 采取动作$T$ 会为 $\mu_0$带来收益 $1 - \epsilon$，为 $\mu_*$带来收益1，从而为 $\mu$带来收益 $1 - \frac{1}{2}\epsilon$。然而，以频率1采取动作$T$ 会忽略 $\mu_*$，并且仅为 $\mu$带来收益$1\bar{2}$。这说明获得收益1是不可能的。

**定理3** 在偶数或奇数马尔可夫决策过程$(\mu)$ 中，决策者没有最优策略，既没有纯策略也没有随机化策略。

**证明：** 该陈述由下面两个断言得出。

断言1：马尔可夫决策过程$(\mu)$的值等于1：$v_\mu = 1$。

断言1的证明：给定$n \in \mathbb{N}$，考虑如下策略 $\sigma_n$：在$E_n^- = \{k \cdot 2^n - 1 : k \in \mathbb{N}\}$ 中的每一阶段采取动作$B$，并在所有其他奇数阶段，即$E_0 \setminus E_n^-$中的阶段，采取动作$T$。于是，决策者在$E_n$ 的每一阶段以及集合$E_0 \setminus E_n^-$的每一阶段都获得奖励1。因此，总收益为

$$u_\mu(\sigma_n) = \mu\big((E_0 \setminus E_n^-) \cup E_n\big) = \tfrac{1}{2} \cdot \mu_0(E_0 \setminus E_n^-) + \tfrac{1}{2} \cdot \mu_*(E_n).$$

由于$E_0 \supseteq E_n^-$，我们有

$$\begin{aligned} \mu_0(E_0 \setminus E_n^-) &= \mu_0(E_0) - \mu_0(E_n^-) \\ &= 1 - \mu_0(\{2^n - 1, 2 \cdot 2^n - 1, 3 \cdot 2^n - 1, \ldots\}) \\ &= 1 - 2 \cdot \phi(\{2^n - 1, 2 \cdot 2^n - 1, 3 \cdot 2^n - 1, \ldots\}) \\ &= 1 - \tfrac{1}{2^{n-1}}, \end{aligned}$$

[6]这样一个累积聚点是电荷 $\mu_*$，使得 $\mu_*(W) = \int_{n \in \mathbb{N}} \mu_n(W)\, \phi(dn)$ 对所有 $W \subseteq \mathbb{N}$ .

其中，最后一个等式是由 $\phi$ 为频率电荷这一事实得出的。我们已经看到，$\mu_*(E_n) = 1$。因此，策略 $\sigma_n$ 下的收益为

$$u_\mu(\sigma_n) = \frac{1}{2} \cdot (1 - \frac{1}{2^{n-1}}) + \frac{1}{2} \cdot 1 = 1 - \frac{1}{2^n}.$$

由于 $n \in \mathbb{N}$是任意的，该断言成立。

断言 2：对于任意策略 $\sigma$，我们有 $u^\mu(\sigma) < 1$。

断言 2 的证明：假设相反情形：某个策略 $\sigma$ 给出收益 1，即 $u^\mu(\sigma) = 1$。设 $W$ 为期望奖励大于 12 的阶段集合，即 $W = \{t \in \mathbb{N} : \mathbb{E}_\sigma[r_t] > \frac{1}{2}\}$。我们必有 $\mu(W) = 1$，从而 $\mu_0(W) = 1$ 且 $\mu_*(W) = 1$。

注意，对于任意 $n \in \mathbb{N}$，我们有 $(W \cap E_n) - 1 \subseteq E_0 \setminus W$。事实上，设 $k \in (W \cap E_n) - 1$。则 $k + 1 \in E_n$，因此 $k + 1$ 为偶数且 $k$ 为奇数，从而证明 $k \in E_0$。此外，$k + 1 \in W$。回想 $r_k + r_{k+1} = 1$，因此 $\mathbb{E}_\sigma[r_k] + \mathbb{E}_\sigma[r_{k+1}] = 1$，从而阶段 $k$ 与 $k + 1$ 的期望奖励不可能都大于 12。这证明了 $k \notin W$。

我们区分两种情形，并分别推出矛盾。

首先，假设 $\mu_n(W) > 0$ 对某个$n \in \mathbb{N}$成立。那么，根据 $\mu_n$的定义，我们有$\phi(W \cap E_n) > 0$，因此 $\phi$ 的平移不变性意味着

$$0 < \phi(W \cap E_n) = \phi((W \cap E_n) - 1).$$

因此

$$\phi(W \cap E_0) = \phi(E_0) - \phi(E_0 \setminus W) \leq \phi(E_0) - \phi((W \cap E_n) - 1) < \phi(E_0) = \frac{1}{2},$$

于是 $\mu_0(W) < 1$成立，这就产生了矛盾。

现在假设对所有$n \in \mathbb{N}$，$\mu_n(W) = 0$ 均成立。由于 $\mu_*$ 是序列 $\{\mu_n\}_{n=1}^\infty$的聚点，因此有 $\mu_*(W) = 0$，这再次产生矛盾。证明完毕。 ■

# 5 结论

**I. 平稳最优策略的不存在性**。对于第4节中定义的偶数或奇数马尔可夫决策过程，不难找到一种聚合电荷 $\mu' \in \Delta^d$，使得在马尔可夫决策过程$(\mu')$中，尽管决策者拥有纯最优策略，但没有任何平稳策略是最优的。

**例4** 考虑第4节中定义的偶数或奇数马尔可夫决策过程。令

$$Q = \{4n - 3 : n \in \mathbb{N}\} \cup \{4n : n \in \mathbb{N}\} = \{1, 4, 5, 8, 9, 12, 13, \ldots\}.$$

令 $\alpha : \mathbb{N} \to Q$为双射，其中 $\alpha(n)$是$Q$中的第$n$个元素。令 $\mu' \in \Delta^d$为一个弥散电荷，使得 $\mu'(\{4n - 3 : n \in \mathbb{N}\}) = \mu'(\{4n : n \in \mathbb{N}\}) = \frac{1}{2}$，因此尤其有 $\mu'(Q) = 1$。

在马尔可夫决策过程($\mu'$)中，存在一种可获得收益1、因而是最优的纯策略：每当对局处于状态1时，交替选择动作$T$ 和$B$ ，从而在 $\{4n-3\colon n \in \mathbb{N}\}$ 中的每个阶段都在状态1获得奖励1，并在 $\{4n\colon n \in \mathbb{N}\}$中的每个阶段都在状态3获得奖励1。

然而，不存在平稳最优策略。事实上，假设该平稳策略以概率 $T$ $q$选择动作。那么，在$\{4n-3n \in \mathbb{N}\}$的每个阶段，期望奖励都等于$q$；而在 $\{4nn \in \mathbb{N}\}$的每个阶段，期望奖励都等于 $1-q$。因此，该策略下的收益等于 $\frac{1}{2}\cdot q+\frac{1}{2}\cdot(1-q)=\frac{1}{2}$，严格小于 1。 ♦

**II. 收益中的积分次序**。策略 $\sigma$ 下的收益（定义见公式（5））可以改写为累次积分。事实上，为了简化记号，假设每个状态下的行动空间都是同一个集合$A$。那么，

$$u_\mu(\sigma) = \int_{t\in\mathbb{N}}\int_{x\in X} r_t(x)\,\mathbb{P}_\sigma(dx)\,\mu(dt), \tag{6}$$

其中，$X = (S\times A)^{\mathbb{N}}$ 表示马尔可夫决策过程中的无限次对局集合，该集合带有由乘积拓扑诱导的博雷尔 σ-代数； $\mathbb{P}_\sigma$ 是由 $\sigma$诱导的$X$ 上的概率测度；而$r_t(x)$ 是当无限次对局为$x \in X$时，第$t$ 阶段奖励的随机变量。我们指出，尽管反转公式（6）中的积分次序颇有意义，但并非总是定义良好，因为积分 $\int_{t\in\mathbb{N}} r_t(x)\,\mu(dt)$ 未必是无限次对局$x$的可测函数。

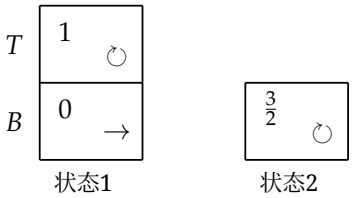
**III. 非弥散聚合收费**。在第 3.2 节中，我们假设聚合电荷是弥散的，这是最具挑战性的情形。当聚合电荷为非弥散电荷 $\mu \in \Delta^f\setminus\Delta^d$ 时，我们仍可用相同的方式定义模型。由于 $\Delta^f$ 中的每个电荷都可以唯一地写成 $\mathbb{N}$ 上的可列可加概率测度与 $\Delta^d$ 中的弥散电荷的凸组合（例如，参见 Rao and Rao [1983] 中的定理 10.2.1），因此有两种情形：(C1) $\mu$ 是 $\mathbb{N}$ 上的可列可加概率测度；或 (C2) $\mu$ 是 $\mathbb{N}$ 上的可列可加概率测度与弥散电荷的凸组合，且可列可加部分的权重为正。

在情形（C1）中，有 $\sum_{t=1}^{\infty}\mu(t) = 1$，并且对于任意策略 $\sigma$，收益等于 $u^\mu(\sigma) = \sum_{t=1}^{\infty}\mu(t)\,\mathbb{E}_\sigma[r_t]$。因此，由于行动空间 $A$ 是有限的，已知决策者拥有纯最优策略：在每个阶段，根据当前状态，在动态规划方程中采取最优动作。[7]

[7]这种策略中，动作选择仅取决于当前阶段和当前状态，称为马尔可夫策略。纯马尔可夫最优策略的存在性也可由许多其他一般性结论推出，例如 Fudenberg 和 Levine [1983] 中的推论 4.2。事实上，为了使记号更简洁，假设每个状态下的行动空间都是同一个集合 $A$。有限期限为 $T$、估值为 $u_\mu^T(\sigma) = \sum_{t=1}^{T}\mu(t)\mathbb{E}_\sigma[r_t]$ 的马尔可夫决策过程通过逆向归纳可得一个纯马尔可夫最优策略 $\sigma^T$。赋予纯策略集合 $A^H$ 乘积拓扑，其中 $H$ 是历史集合，该集合是紧致且可度量化的（因为 $H$ 是可数集），因此序列 $\{\sigma^T\}_{T=1}^{\infty}$ 存在一个聚点 $\sigma$，它是一个纯马尔可夫策略。由连续性论证可知， $\sigma$ 是最优的。

在情形（C2）中，以下示例表明，最优策略并不总是存在。当决策者试图最大化奖励的折扣和与长期平均奖励的凸组合时，这类示例是众所周知的，下面的示例正属于此类。

**例5** 考虑以下马尔可夫决策过程。状态集为 $S = \{1,2\}$。在状态1中，决策者有两个动作：$T$ 和$B$。如果她选择$T$，则奖励为1，且过程仍留在状态1。如果她选择$B$，则奖励为0，下一个状态为状态2。在状态2中，决策者只有一个动作，该动作带来32的奖励，并使过程继续留在状态2。




最后，定义 $\mu \in \Delta^f$ 为

$$\mu = \frac{1}{2}\mu' + \frac{1}{2}\mu''$$

其中，$\mu'$是定义在 $\mathbb{N}$ 上的可列可加概率测度，其概率为$(12, 14, 18, \ldots)$，而 $\mu'' \in \Delta^d$ 是任意一个弥散电荷。[8]

直观而言，决策者希望在某个阶段选择动作$B$，但最好尽可能晚地选择，这样因奖励为0而产生的损失就会降到最低。正如我们将要证明的，这将导致她不存在最优策略。

我们只讨论纯策略；证明可以扩展到非纯策略。令$T^\infty$ 表示始终选择$T$的纯策略；对于每个$n \in \mathbb{N}$，令$B^n$ 表示在阶段$n$ 之前选择$T$，并在阶段$n$选择$B$ 的纯策略。

考虑公式（5）中的收益 $u_\mu(\cdot)$。我们有 $u_\mu(T^\infty) = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 1 = 1$ 且 $u_\mu(B^1) = \frac{1}{2}\cdot(\frac{1}{2}\cdot\frac{3}{2}) + \frac{1}{2}\cdot\frac{3}{2} = \frac{9}{8}$。因此，$T^\infty$ 不是最优的。此外，对于任意 $n \in \mathbb{N}$，策略 $B^n$ 和 $B^{n+1}$ 仅在阶段 $n$（奖励分别为 0 和 1）以及阶段 $n+1$（奖励分别为 3/2 和 0）产生不同的奖励。因此，$\mu''(B^n) = \mu''(B^{n+1})$，由此得到

$$u_\mu(B^n) - u_\mu(B^{n+1}) = \frac{1}{2}\cdot(\mu'(B^n) - \mu'(B^{n+1})) = \frac{1}{2}\cdot\left(\frac{1}{2^{n+1}}\cdot\frac{3}{2} - \frac{1}{2^n}\cdot 1\right) < 0.$$

因此，对于任意$n \in \mathbb{N}$，策略$B^{n+1}$ 严格优于$B^n$。由此可知不存在纯最优策略，结论得证。♦

[8]这里，$\mu'$ 是折扣因子为 12 的折扣评估，而 $\mu''$ 可以是代表平均奖励评估的任意频率电荷。

# 参考文献

Fudenberg D、Levine D [1983]：有限期与无限期博弈的子博弈精炼均衡。《经济理论杂志》31，251-268。

Kadane JB, O’Hagan A [1995]：使用有限可加概率：自然数上的均匀分布。《美国统计学会杂志》90，626-631。

Neyman A [2023]：满足货币时间价值原理的收益流的可加估值：一种刻画与鲁棒优化。《理论经济学》，18(1)，303-340。

Puterman ML [1990]：马尔可夫决策过程。《运筹学与管理科学手册》，2，331-434。

Rao KPSB, Rao MB [1983]：荷理论。学术出版社。

Schirokauer O, Kadane JB [2007]：自然数上的均匀分布。《理论概率杂志》，20，429-441。