具有贬值资产的强化学习

Taylor Dohmen 和 Ashutosh Trivedi

科罗拉多大学博尔德分校

2023年3月1日

arXiv:2302.14176v1 [cs.AI] 27 Feb 2023

摘要

传统强化学习的一个基本假设是，奖励一旦被智能体获得，其价值便不会改变。本研究摒弃这一假设，考虑奖励价值会随着获得后经过的时间而按比例衰减的情形。我们强调支付时发生的转折点，并使用资产一词，指当前由智能体持有的奖励。采用这一术语，我们在无限时域定量优化的框架下，开启了对折旧资产的研究。具体而言，我们借鉴经典的指数折扣，提出资产折旧的概念：资产被智能体获得后，每经过一个时间步，其价值便乘以一个固定的折扣因子。我们针对这一情境建立了最优性的贝尔曼式等式刻画，并开发了一种无模型强化学习方法，以求得最优策略。

1 引言

时间偏好 [Frederick 等，2002；Loewenstein 和 Jon，1992]指的是理性智能体按照实现某种理想结果 预期所需的时间，对其价值进行评估的倾向。换言之，在其他条件相同的情况下，智能体倾向于尽早而非延后获得未来奖励；同样，智能体也倾向于较晚而非较早经历负面结果。这一现象通常通过折扣 [Shapley， 1953] 纳入数学模型，并应用于经济学 [Heal，2007；Philibert， 1999]，、博弈论 [Filar and Vrieze， 1996]，、控制理论 [Puterman， 1994]，以及强化学习 [Sutton and Barto， 2018]等涉及优化的众多领域。这些模型关注这样一种情形：智能体在离散时间的随机环境中行动，在每个时间步选择一个动作执行，并根据所选动作和环境状态获得即时奖励。具体而言，我们考虑 Shapley 引入的指数折扣 [1953]，：智能体无限期地持续这一过程，从而产生无限的奖励序列 $\langle r_n \rangle_{n=1}^{\infty}$；其目标是针对折扣因子 $\lambda \in (0,1)$，最大化折扣总和 $\sum_{n=1}^{\infty} \lambda^{n-1} r_n$。折扣因子作为一个参数选取，用于量化智能体时间偏好的程度。

上述贴现优化框架的一个显著特点是，它隐含地假设奖励的效用——一旦被学习型智能体获得，便始终保持不变。虽然这一看似无伤大雅的假设简化了模型，使其更易于分析，但在许多情形下，这一假设并不恰当。例如，考虑用于激励人类行为的最基本且最普遍的奖励：货币。货币的价值往往会随时间推移而按照通货膨胀率贬值，这种贬值造成的影响广受关注，也得到了深入研究 [Beckerman，1991；Comley，2015；Fergusson，2010；Hulten 和 Wykoff， 1980]。认识到这种贬值对全球经济系统动态的影响具有根本性作用，我们探讨其对马尔可夫决策过程中优化与强化学习的影响。

1.1 资产折旧

在讨论奖励价值会衰减的情形时，区分潜在未来奖励与已经获得的实际奖励是很有用的。因此，我们引入资产一词，用来指代智能体在过去某个时刻获得的奖励。使用这一术语，本文可以被描述为：在假设资产折旧的条件下，对优化与学习进行探究。折旧一词借自金融与会计领域 [Burt，1972年；Wright， 1964]，它准确描述了某种事物的价值随时间衰减的现象。

我们提出了一种受传统折扣启发的折旧概念，其基础是在智能体的未来之外，也将相同的时间偏好基本原则应用于智能体的历史。更准确地说，我们考虑这样一种情形：根据由无限序列的累积获得资产构成的序列来评估智能体的行为，其中每项资产都按照其获得时间距今的长短进行折扣。也就是说，我们提议根据资产序列上的函数来评估智能体

$$\left\langle \sum_{k=1}^{n} r_k \gamma^{n-k} \right\rangle_{n=1}^{\infty},$$

其中 $\gamma \in (0,1)$ 是折扣因子，而非奖励序列 $\langle r_n \rangle_{n=1}^{\infty}$。为阐明研究折旧的动机并说明其合理性，我们考察以下假设性案例。

示例 1（二手车经销商店）。 设想一家二手车经销商店，其商业模式是在区域市场行情有利的位置购买二手车，将它们开回店里

，然后在当地市场出售以获取利润。假设我们的优化主体是这家汽车经销商的员工，负责管理资本获取。具体而言，每当需要作出选择时，该员工的工作就是决定下一辆车应从哪个地点购入。智能体的目标是在折扣因子为 $\lambda \in (0,1)$ 的情况下，在折扣时间范围内最大化经销商店库存中所有车辆价值的总和。请注意，折扣时间范围问题等价于：在任何时刻以恒定概率 $(1-\lambda)$ 终止运营时，最大化该过程的预期终端收益。

人们早已知道 *[Ackerman, 1973; Wykoff, 1970]* 汽车作为新车售出后，其价值往往会持续贬值，因此，任何合理的库存中所有车辆价值模型都应纳入某种资产贬值机制。假设另一个折扣因子 $\gamma \in (0,1)$ 表示汽车单位时间内的价值损失率。考虑 γ贬值奖励和 λ折扣时域，我们的智能体目标可定义为一个折扣贬值优化问题。或者，也可以寻求优化 γ贬值奖励的长期平均值（平均收益）。

1.2 折扣贬值收益

考虑智能体累积的（绝对）奖励序列 $x = \langle 3,4,5,3,4,5,\ldots \rangle$ 。在存在贬值的情况下，各时点的累计资产价值遵循以下序列

$$3,(3\gamma+4),(3\gamma^2+4\gamma+5),(3\gamma^3+4\gamma^2+5\gamma+3),$$
$$(3\gamma^4+4\gamma^3+5\gamma^2+3\gamma+4),\ldots$$

对于 λ折扣时间范围，可以按如下方式计算资产的价值：

$$\begin{aligned}
&3+\lambda(3\gamma+4)+\lambda^2(3\gamma^2+4\gamma+5)+\lambda^3(3\gamma^3+4\gamma^2+5\gamma+3)+\\
&\quad \lambda^4(3\gamma^4+4\gamma^3+5\gamma^2+3\gamma+4)+\ldots\\
&=(3+3\lambda\gamma+3\gamma^2\lambda^2+\cdots)+(4\lambda+4\lambda^2\gamma+4\lambda^3\gamma^2+\cdots)+\\
&\quad (5\lambda^2+5\lambda^3\gamma+\lambda^5\gamma^2+\cdots)+(3\lambda^3+3\lambda\gamma^4+3\gamma^2\lambda^5+\cdots)+\cdots\\
&=3(1+\lambda\gamma+\gamma^2\lambda^2+\cdots)+4\lambda(1+\lambda\gamma+\lambda^2\gamma^2+\cdots)+\\
&\quad 5\lambda^2(1+\lambda\gamma+\lambda^2\gamma^2+\cdots)+3\lambda^3(1+\lambda\gamma+\gamma^2\lambda^2+\cdots)+\ldots\\
&=\frac{3+4\lambda+5\lambda^2+3\lambda^3+\cdots}{(1-\lambda\gamma)}\\
&=\frac{3+4\lambda+5\lambda^2}{(1-\lambda\gamma)(1-\lambda^3)}.
\end{aligned}$$

注意，当即时奖励按因子 $\frac{1}{1-\lambda\gamma}$ 缩放时，这个 γ折旧总和等于 λ 折扣总和。我们证明这并非巧合，并证明这种相等关系对一般的马尔可夫决策过程也成立。

1.3 平均折旧收益

接下来，考虑折旧资产价值的长期平均值，即该序列的下极限

$$3,\frac{3\gamma+4}{2},\frac{3\gamma^2+4\gamma+5}{3},\frac{3\gamma^3+4\gamma^2+5\gamma+3}{4},$$
$$\frac{3\gamma^4+4\gamma^3+5\gamma^2+3\gamma+4}{5},\ldots$$

基于经典的陶伯尔结果 [Bewley和Kohlberg，1976]，人们很容易猜想， λ折扣、 γ折旧的价值会在 $\lambda \to 1$时收敛到该平均值，例如

$$\begin{aligned}
\lim_{\lambda\to 1}(1-\lambda)\frac{3+4\lambda+5\lambda^2}{(1-\lambda\gamma)(1-\lambda^3)} &= \lim_{\lambda\to 1}\frac{3+4\lambda+5\lambda^2}{(1-\lambda\gamma)(1+\lambda+\lambda^2)}\\
&= \frac{3+4+5}{3(1-\gamma)}.
\end{aligned}$$

事实上，我们证明了这一猜想成立。

贡献。 本文的主要亮点如下。

▶ 我们开启了在折旧动态存在的情况下，对折扣收益和平均收益优化的研究。

▶ 我们通过贝尔曼式最优性方程刻画折旧折扣收益的最优值，并据此证明，平稳确定性策略足以实现最优性。此外，我们的刻画使得在规划场景中能够以多项式时间计算最优值和一个最优策略。

▶ 该最优性方程还有助于构建一种与资产折旧兼容的Q学习变体，从而提供一种无模型强化学习方法，以便在学习场景中获得最优策略。

▶ 我们证明，经典的陶伯定理（该定理揭示了折扣目标与平均目标之间的关系）可以扩展到折旧奖励场景。这一结果使我们能够确立：对于平均折旧收益的最优性而言，平稳确定性策略是足够的。

组织结构。 我们首先介绍必要的符号，并回顾相关的技术背景。第3节给出关于折旧折扣收益的结果，第4节则介绍平均折旧目标的相关结果。我们在第5节讨论一些密切相关的工作，并在结论部分总结本文的贡献。

2 预备知识

令 $\mathbb{R}$ 为实数集， $\mathbb{N}$ 为自然数集。对于集合X，我们用 $|X|$ 表示其基数，并用$\mathsf{Dist}(X)$表示X上的所有概率分布构成的集合。X 上的点分布是指将概率 1 分配给X中的某个唯一元素，并将概率 0 分配给其他所有元素的分布。

本文的技术部分是在渐近优化的标准数学框架下完成的

并研究以有限马尔可夫决策过程建模的环境。我们的论述遵循优化与学习领域标准教科书中的惯例 [Feinberg 和 Shwartz, 2012年；Filar and Vrieze，1996；Puterman，1994；Sutton and Barto， 1998]。

2.1 马尔可夫决策过程

一个（有限）马尔可夫决策过程（MDP）M 是一个元组 (S, A,T,R)，其中 S 是有限状态集，A 是有限动作集，$T : (S \times A) \to \text{Dist}(S)$ 是一个随机转移函数，用于指定任意 $s,t \in S$ 和 $a \in A$ 下，在当前状态为 s 且已选择动作 a 的条件下，转移到状态 t 的条件概率 $T(t \mid s,a)$；$R : (S \times A) \to \mathbb{R}$ 是一个实值奖励函数，将每个状态-动作对映射为一个数值。对于任意函数 $f : S \to \mathbb{R}$，即 MDP 状态空间上的任意随机变量，我们用 $\mathbb{E}_T[f(t) \mid s,a]$ 表示在智能体从状态 s 选择动作 a 的条件下，f 在后继状态上的条件期望 $\sum_{t \in S} f(t)T(t \mid s,a)$。M 中的一条路径是由状态和动作交替组成的序列 $s_1a_1s_2 \cdots a_n s_{n+1}$，且在每个索引处都有 $0 < T(s_{k+1} \mid s_k, a_k)$。令 $\mathscr{F}(M)$ 表示 M 中所有有限路径的集合，令 $\mathscr{I}(M)$ 表示 M 中所有无限路径的集合。

收益、策略与最优性。我们关注无限时长的定量优化问题，其中结果可以具体表示为马尔可夫决策过程中的一条无限路径。此类结果依据某个映射到实数 $\mathscr{I}(M) \to \mathbb{R}$ 来评估，该映射称为收益。M 上的策略是一个函数 $\pi : \mathscr{F}(M) \to \text{Dist}(A)$，它根据 M 中的有限路径，从动作集中选择一个分布。固定一个策略 π 后，对于每个状态 s，它都会诱导出一个唯一的概率测度 $\mathbb{P}_s^\pi$，定义在 (M) 的博雷尔子集上的概率空间上。这样，给定收益和初始状态 s，就可以计算策略的期望 $\mathbb{E}_s^\pi$。令 Π^M 表示马尔可夫决策过程 M 上的所有策略。若某个策略使该收益的期望值在所有策略中达到最大，则该策略对于此收益而言是最优的；这个最大期望值称为该收益在 M 上的值。

策略复杂度。 收益的策略复杂度刻画了策略达到最优所需的必要结构。策略复杂度的一个定性方面取决于是否存在这样的环境：在这些环境中，最优策略必然是概率性的（混合）。如果一个策略对每个输入都返回点分布，则称其为确定性 （纯策略）策略。若 $\pi(s_1a_1 \cdots a_{n-1}s_n) = \pi(s_n)$ 在每个时间 n 均成立，则策略是平稳的。确定性平稳策略类特别值得关注，因为在任何有限马尔可夫决策过程中，这类策略的数量都是有限的；我们将这些策略视为函数 $S \to A$。

2.2 折扣收益与平均收益

给定马尔可夫决策过程中的一条路径 $s_1a_1s_2 \cdots$，两个经过充分研究的目标是相对于折扣因子 $\lambda \in (0,1)$ 的折扣收益，以及定义如下的平均收益：

$$\sum_{n=1}^{\infty} \lambda^{n-1} R(s_n, a_n), \text{ and} \qquad \text{(Discounted Payoff)}$$

$$\liminf_{n \to \infty} \frac{1}{n} \sum_{k=1}^{n} R(s_k, a_k). \qquad \text{(Average Payoff)}$$

折扣价值函数和平均价值函数定义如下

$$V_\lambda(s) = \sup_{\pi \in \Pi^M} \mathbb{E}_s^\pi \left[\sum_{n=1}^{\infty} \lambda^{n-1} R(s_n, a_n) \right], \qquad \text{(Discounted Value)}$$

$$V(s) = \sup_{\pi \in \Pi^M} \mathbb{E}_s^\pi \left[\liminf_{n \to \infty} \sum_{k=1}^{n} \frac{R(s_k, a_k)}{n} \right]. \qquad \text{(Average Value)}$$

一种更强的最优性概念，专门针对折扣收益，是布莱克威尔最优性。如果存在折扣因子 $\lambda_0 \in (0,1)$，使得 π 对于区间 $[\lambda_0, 1)$ 中的任意折扣因子对应的折扣收益都是最优的，则称策略 π 为布莱克威尔最优。

折扣价值的另一种刻画方式，是将其视为最优性方程的唯一解

$$V_\lambda(s) = \max_{a \in A} R(s,a) + \lambda \mathbb{E}_T[V_\lambda(t) \mid s,a],$$

这为确立以下关于折扣收益与平均收益复杂度的结果奠定了基础 [Feinberg and Shwartz，2012年；Filar 和 Vrieze，1996年；Puterman， 1994]。

定理1. 折扣收益和平均收益都存在确定性平稳最优策略。此外，两种收益的最优值都可以在多项式时间内计算。

2.3 强化学习

强化学习（RL） [Sutton and Barto， 2018] 是一种基于采样的优化范式，它依据从环境中以标量奖励形式获得的反馈进行优化。标准强化学习场景假设采用折扣收益，而无模型方法通常利用状态-动作价值或 Q 值：其定义为在状态 s 下选择动作 a 后所能获得的最优价值，也是下列方程的解

$$Q_\lambda(s,a) = R(s,a) + \lambda \mathbb{E}_T[V_\lambda(t) \mid s,a].$$

Q值为经典的Q学习算法奠定了基础 [Watkins 和 Dayan, 1992], 该算法通过对 Q_λ 进行近似，利用一系列 Q_λ^n 映射，使其渐近收敛到 Q_λ，从而学习最优策略。具体而言，Q_λ^1 初始化为任意值，随后智能体通过从当前状态 s 中选择动作 $a = \text{argmax}_{a \in A} Q_\lambda^n(s,a)$ 来探索环境，并执行以下更新

$$Q_\lambda^{n+1}(s,a) \leftarrow Q_\lambda^n(s,a) + \alpha_n \left(R(s,a) + \lambda V_\lambda^n(t) - Q_\lambda^n(s,a) \right), \quad (1)$$

其中，t 是根据对条件分布 $T(\cdot \mid s,a)$ 进行采样所得的结果确定的下一个状态， $\alpha_n \in (0,1)$ 的族是称为学习率的时变参数，而 $V_\lambda^n(t) = \max_{a \in A} Q_\lambda^n(t,a)$。下面的定理给出了 Q-学习算法渐近收敛的一个充分条件。

定理 2 (Watkins 和 Dayan [1992])。 如果环境决策过程中的每个状态-动作对都被无限次地遇到，并且学习率 $0 \le \alpha_n < 1$ 满足 *Robbins-Monroe* 条件 $\sum_{n=1}^{\infty} \alpha_n = \infty$ 和 $\sum_{n=1}^{\infty} \alpha_n^2 < \infty$，那么 $Q_\lambda^{n+1}(s,a) \to Q_\lambda$ 几乎必然地当 $n \to \infty$ 时成立。

2.4 折旧资产

我们基于这样一种思想来定义折扣收益与平均收益的变体：资产的价值会随着自其作为奖励获得以来经过的时间，按几何比例衰减。也就是说，我们考虑这样一种情形：收益由序列 $\langle R(s_n,a_n)\rangle_{n=1}^{\infty}$ 决定，而是由序列

$$\left\langle \sum_{k=1}^{n} R(s_k,a_k)\gamma^{n-k} \right\rangle_{n=1}^{\infty}$$

智能体资产的指数近期加权平均，其中 $\gamma \in (0,1)$ 为折扣因子。

3 折扣折旧收益

本节中，我们研究资产折旧动态下、 $\lambda \in (0,1)$ 时的折扣优化。此情形下的收益由下式表示

$$\sum_{n=1}^{\infty} \lambda^{n-1} \sum_{k=1}^{n} R(s_k,a_k)\gamma^{n-k}, \quad \text{(Discounted Depreciating Payoff)}$$

它对应的价值函数为

$$V_\lambda^\gamma(s) = \sup_{\pi \in \Pi^M} \mathbb{E}_s^\pi \left[\sum_{n=1}^{\infty} \lambda^{n-1} \sum_{k=1}^{n} R(s_k,a_k)\gamma^{n-k} \right].$$

现在让我们回到二手车经销商店的示例。

示例 2（二手车经销商店，续）。 由于汽车在首次购买后会不断折旧，员工意识到，他们的模型应纳入资产折旧的概念。经过一番市场调研后，员工选取了另一个折扣因子 $\gamma \in$（0,1），以反映汽车在每个时间步内通常的贬值幅度。通过使用折扣因子 λ和 γ，员工可以将这一情境建模为一个折扣折旧优化问题。

为简化起见，假设只有两个位置 s_1 和 s_2 可供选择作为下一个目标市场，并且员工拥有多个可选动作的唯一位置是汽车经销商 s_d （在那里，他们可以选择动作 a_1 前往 s_1，或选择动作 a_2 前往 s_2）。员工意识到，在不考虑意外延误的情况下进行规划是不合理的，因此又引入了两个参数 ρ_1 和 ρ_2，分别表示在 s_1 和 s_2 购买所需车辆的成功率。给定智能体位于位置 s_i，比率 ρ_i 可解释为他们在当天结束前找到卖家并购买车辆的概率，因此 $1-\rho_i$ 是他们未能做到这一点的概率。图*1*将这一情况以有限马尔可夫决策过程的形式进行了图示，其中动作以红色显示，转移概率以蓝色显示，即时奖励（即库存车辆的价值）以绿色显示。如果边标签中省略了某个动作，

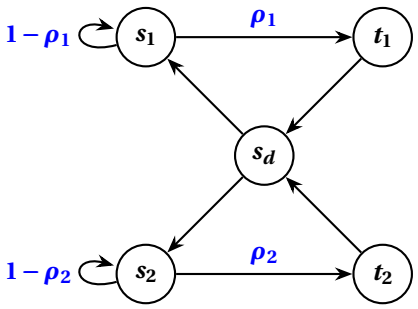

图1：用于汽车经销商折扣折旧优化问题的马尔可夫决策过程。

则表示只有一个可用动作 a。如果省略了转移概率，则转移是确定性的，即以概率 *1* 发生。如果省略了奖励值，则获得的奖励为 *0*。

在传统的折扣优化中，折扣因子 λ会带来某种权衡。例如，假设ρ_1 较大而 r_1 较小，并且 ρ_2 较小而 r_2 较大。那么，较小的折扣因子表明，采取动作 a_1 可能更有收益，因为采取动作 a_2 很可能会导致显著延迟，从而降低最终奖励 r_2的价值。另一方面，如果折扣因子接近 *1*，那么智能体或许值得接受较高的延迟概率，因为最终的折扣价值会更接近 r_2。

加入折扣因子γ下的折旧动态后，这种权衡依然存在，但折旧在多大程度上会改变特定环境和策略的动态，尚不明确。直觉上，引入折旧进行折扣优化，或许只会使风险—收益权衡更加尖锐；进一步还可以推测，当 γ接近*0*时，累计资产价值更快的衰减率会促使智能体采取更冒险的行为。另一方面，折旧因子接近 *1*时，也可能让智能体更愿意采取类似的冒险动作，因为随着资产积累得越来越多，这种行为的机会成本会随之降低。在接下来的折扣递减收益分析中，我们将尝试阐明这类问题，并深入探究折旧在此语境中的本质。

我们的首个主要结果给出了折扣递减价值的贝尔曼型方程刻画。

定理3(最优性方程)。 折扣递减价值是下列方程的唯一解

$$V_\lambda^\gamma(s) = \max_{a \in A} \frac{R(s,a)}{1-\lambda\gamma} + \lambda \mathbb{E}_T\left[V_\lambda^\gamma(t) \,\middle|\, s,a \right]. \tag{2}$$

证明。 将折旧折扣收益定义中出现的项 λ^{n-1} 拆分为乘积 $\lambda^{n-k}\lambda^{k-1}$，并将这些因子分配到内层求和中，我们得到表达式

$$\sum_{n=1}^{\infty} \sum_{k=1}^{n} \lambda^{k-1} R(s_k,a_k) \lambda^{n-k} \gamma^{n-k}. \tag{3}$$

证明的下一步依赖于实分析中的以下经典结果（参见 Rudin [1976] 的定理 3.50）。

⌜ MERTENS' THEOREM。 设 $\sum_{n=1}^{\infty} x_n = X$ 和 $\sum_{n=1}^{\infty} y_n = Y$ 是两个收敛的实数级数。若给定的级数中至少有一个绝对收敛，则它们的柯西乘积收敛于其极限的乘积：

$$\left(\sum_{n=1}^{\infty} x_n\right)\left(\sum_{n=1}^{\infty} y_n\right) = \sum_{n=1}^{\infty}\sum_{k=1}^{n} x_k y_{n-k} = XY.$$

⌟

级数 (3) 可分解为柯西乘积

$$\left(\sum_{n=1}^{\infty} (\lambda\gamma)^{n-1}\right)\left(\sum_{n=1}^{\infty} \lambda^{n-1} R(s_n, a_n)\right), \tag{4}$$

由于这个柯西乘积中的两项都绝对收敛，因此梅尔滕斯定理适用。注意到左侧级数是几何级数，表达式 (4) 等价于

$$\frac{1}{1-\lambda\gamma}\sum_{n=1}^{\infty} \lambda^{n-1} R(s_n, a_n).$$

因此，折扣递减价值可以写成

$$\begin{aligned} V_\lambda^\gamma(s) &= \sup_{\pi\in\Pi^M} \mathbb{E}_s^\pi\left[\frac{1}{1-\lambda\gamma}\sum_{n=1}^{\infty} \lambda^{n-1} R(s_n, a_n)\right] \\ &= \frac{1}{1-\lambda\gamma}\sup_{\pi\in\Pi^M} \mathbb{E}_s^\pi\left[\sum_{n=1}^{\infty} \lambda^{n-1} R(s_n, a_n)\right] \\ &= \frac{V_\lambda(s)}{1-\lambda\gamma}. \end{aligned} \tag{5}$$

折扣价值V_λ的等式刻画使我们能够进一步推导出所需的折扣递减价值V_λ^γ的等式刻画，如下所示

$$\begin{aligned} V_\lambda^\gamma(s) &= \frac{1}{1-\lambda\gamma}\left(\max_{a\in A} R(s,a) + \lambda\mathbb{E}_T\left[V_\lambda(t) \mid s, a\right]\right) \\ &= \max_{a\in A}\frac{R(s,a)}{1-\lambda\gamma} + \lambda\mathbb{E}_T\left[V_\lambda^\gamma(t)\,\middle|\, s, a\right]. \end{aligned} \tag{6}$$

□

定理3的一个直接推论是对折扣递减收益的策略复杂度的刻画。

推论 1 (策略复杂度)。 对于任意有限马尔可夫决策过程上的折扣递减收益，都存在一个平稳且确定性的最优策略。

定理3使得许多经过广泛研究的算法技术能够调整后用于折扣递减收益。具体而言，折扣递减价值的方程刻画表明，它是压缩映射 [Banach, 1922], 的唯一不动点；这进而有助于基于价值迭代和线性规划等基础方法，制定合适的规划算法变体。因此，我们可以根据环境马尔可夫决策过程的规模和给定的折扣因子，对求解折扣递减价值的计算复杂度进行界定。

定理4（计算复杂度）。 折现的贬值资产价值及其对应的最优策略可在多项式时间内计算。

证明。 令 $\delta_{i,j} = \begin{cases} 1 \\ 0 \end{cases}$ 若$i = j$则为克罗内克δ函数。假设对于环境M中的每个状态 s，都有一个与之关联的实数 $0 < x_s$，其取值任意。以下线性规划的唯一解是M各状态的价值向量。

$$\begin{aligned} &\text{minimize} \sum_{s\in S} x_s v_s \quad \text{subject to} \\ &\frac{R(s,a)}{1-\lambda\gamma} \le \sum_{t\in S} v_t\left(\delta_{s,t} - \frac{\lambda T(t\mid s,a)}{1-\lambda\gamma}\right) \quad \forall (s,a)\in S\times A \end{aligned} \tag{7}$$

根据式（7）的解 v^*，可以得到一个最优策略：

$$\pi(s) = \arg\max_{a\in A}\frac{R(s,a)}{1-\lambda\gamma} + \lambda\mathbb{E}_T\left[v_t^* \mid s, a\right].$$

或者，也可以根据以下对偶线性规划的解推导出一个最优策略。

$$\begin{aligned} &\text{maximize} \sum_{(s,a)\in S\times A}\frac{R(s,a)}{1-\lambda\gamma} y_{s,a} \quad \text{subject to} \\ &x_s = \sum_{(t,a)\in S\times A}\left(\delta_{s,t} - \frac{\lambda T(t\mid s,a)}{1-\lambda\gamma}\right) \qquad \forall s\in S \\ &0 \le y_{s,a} \qquad \forall (s,a)\in S\times A \end{aligned} \tag{8}$$

具体而言，如果 y^* 是式（8）的一个解，那么任何在每个状态都满足不等式 $0 < y^*$ s，$\pi(s)$ 的策略 π 都是最优的。这些线性规划的正确性可由定理3的证明得出。由于线性规划可以在多项式时间内求解，定理得证。

□

定理3允许构造相应的Q值

$$Q_\lambda^\gamma(s,a) = \frac{R(s,a)}{1-\lambda\gamma} + \lambda\mathbb{E}_T\left[V_\lambda^\gamma(t)\,\middle|\, s, a\right],$$

可用于构建折扣递减收益的Q学习迭代方案：

$$Q_\lambda^{\gamma,n+1}(s,a) \leftarrow Q_\lambda^{\gamma,n}(s,a) + \alpha_n\left(\frac{R(s,a)}{1-\lambda\gamma} + \lambda V_\lambda^{\gamma,n}(t) - Q_\lambda^{\gamma,n}(s,a)\right). \tag{9}$$

定理5。 如果环境中的每个状态-动作对都被无限次遇到，且学习率满足*Robbins-Monro*收敛准则

$$\sum_{n=0}^{\infty}\alpha_n = \infty \quad \text{and} \quad \sum_{n=0}^{\infty}\alpha_n^2 < \infty,$$

那么迭代(9) 几乎必然收敛到折扣折旧Q值，当$n \to \infty$时：

$$\lim_{n\to\infty} Q_\lambda^{\gamma,n} = Q_\lambda^\gamma.$$

证明。 方程（5）和（6）表明，折扣递减价值的最优性方程可化为折扣价值的最优性方程，只是相差一个依赖于 λ 和 γ的乘法因子。因此可以推出，在相同条件下，折扣递减Q学习通过迭代（9）最终收敛于最优Q_λ^γ，而标准折扣Q学习通过迭代（1）最终收敛于最优Q_λ。因此，我们得出结论：只要每个状态-动作对都被无限次遇到，且定理陈述中的收敛条件对于学习率均得到满足，折扣递减Q学习就会渐近收敛。

□

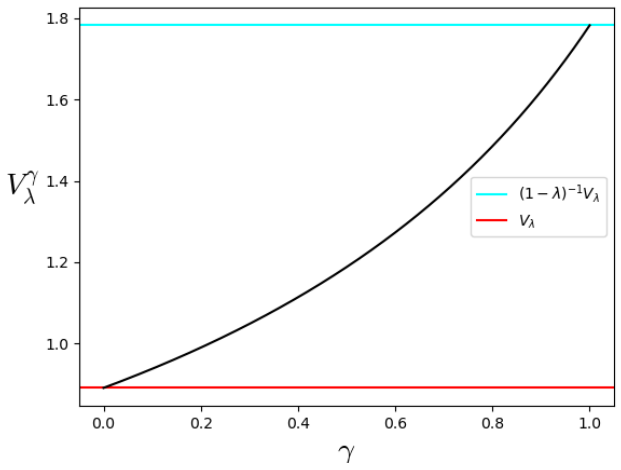

图 2：汽车经销商示例中折扣递减价值随 γ 变化的曲线图，变化区间为 (0,1)，且 $\lambda = \frac{1}{2}$ 固定。该图的参数值为 $\rho_1 = \frac{1}{2}$, $\rho_2 = 1\,4$, $r_1 = 5$, $r_2 = 7$。

3.1 讨论

除定理3的技术含义外，其证明还揭示了折扣与折旧之间相互作用的一些信息。无限期优化理论中的一个基础性结果 [Bewley和Kohlberg，1976] 表明，在一个共同的马尔可夫决策过程中，当 λ 从下方趋近于1时，折扣价值渐近地趋近于平均价值，二者相差一个$(1-\lambda)$的乘法因子：

$$\lim_{\lambda\to 1}(1-\lambda)V_\lambda = V.$$

沿用这一方法，我们考察折扣因子的类似极限下，折扣递减价值的渐近行为。以方程 (5) 中的恒等式 $V_\lambda^\gamma = V_{1-\frac{\lambda}{\lambda\gamma}}$ 为起点求这些极限，可得以下方程：

$$\lim_{\lambda\to 1}(1-\lambda)V_\lambda^\gamma = \frac{V}{1-\gamma}, \tag{10}$$

$$\lim_{\gamma\to 1} V_\lambda^\gamma = \frac{V_\lambda}{1-\lambda}, \tag{11}$$

$$\lim_{\gamma\to 0} V_\lambda^\gamma = V_\lambda. \tag{12}$$

方程（12）和（11）所描述的关系如图2所示。我们可以用一种简单的概念性解释来论证这些关系，这有助于直观理解折旧折扣收益的行为。可以将标准折扣收益视为折旧折扣收益的一个特例，其中 $\gamma = 0$。也就是说，致力于最大化折扣收益的优化智能体，在任何时刻都完全不考虑其资产的价值；从它的角度来看，唯一需要关注的是流入的奖励流。若将 γ 解释为衡量智能体对过去结果的记忆程度，那么很自然地可以得出，折旧折扣收益会在智能体完全没有任何回忆时简化为

折扣收益。联系这一观点与折旧的概念，可以认为，从智能体的角度来看，外部驱动的资产折旧，在道德意义上等同于基于对过去事件的不完整回忆而产生的内部折旧感知。

反过来，采用折扣收益的完美记忆智能体，最终会在累积资产序列 $\langle \sum_{k=1}^{n} R(s_k, a_k)\rangle_{n=1}^{\infty}$ 上最大化该收益，而不是在即时奖励序列 $\langle R(s_n,a_n)\rangle_{n=1}^{\infty}$ 上最大化。假设即时奖励为正，则得到的价值会高于直接基于奖励序列所能获得的价值，图 2 展示了这一点。由于标准折扣所产生的压缩性质，即使累积资产流可能没有界，整体总和仍会收敛。

4 平均折旧收益

现在来考虑资产发生折旧时的渐近平均评估准则。在此情境下，结果的收益定义为

$$\liminf_{n\to\infty} \sum_{k=1}^{n}\sum_{i=1}^{k} \frac{R(s_i,a_i)\gamma^{k-i}}{n}, \quad \text{(Average Depreciating Payoff)}$$

相应的平均折旧价值函数为

$$V^\gamma(s) = \sup_{\pi\in\Pi^M} \mathbb{E}_s^\pi \left[\liminf_{n\to\infty} \sum_{k=1}^{n}\sum_{i=1}^{k} \frac{R(s_i,a_i)\gamma^{k-i}}{n}\right].$$

本节的主要结果在渐近意义下建立了平均折旧价值与折现折旧价值之间的联系。

定理 6 （陶伯定理）。当 $\lambda \to 1$ 从下方趋近时，折现折旧价值乘以（$1-\lambda$）后收敛于平均折旧价值：

$$\lim_{\lambda\to 1}(1-\lambda)V_\lambda^\gamma = V^\gamma.$$

定理 6 的证明用到以下两个引理。

引理 1。对于环境 *MDP* 中的任意有限路径，

$$\sum_{k=1}^{n}\sum_{i=1}^{k} \frac{R(s_i,a_i)\gamma^{k-i}}{n} = \sum_{k=1}^{n} \frac{R(s_k,a_k)(1-\gamma^{n+1-k})}{n(1-\gamma)}. \tag{13}$$

证明。我们对n进行归纳法证明。

基例。假设 $n = 1$。则式 (13) 中出现的两个表达式都求值为 $R(s_1,a_1)$。

归纳步骤。假设 (13) 对 $n-1$ 成立。将 (13) 左侧的求和拆分后，我们得到如下表达式

$$\sum_{k=1}^{n-1}\sum_{i=1}^{k} \frac{R(s_i,a_i)\gamma^{k-i}}{n} + \sum_{k=1}^{n} \frac{R(s_k,a_k)\gamma^{n-k}}{n}.$$

从该表达式的双重求和中提取$n-1$ n，得到

$$\frac{n-1}{n}\sum_{k=1}^{n-1}\sum_{i=1}^{k}\frac{R(s_i,a_i)\gamma^{k-i}}{n-1}+\sum_{k=1}^{n}\frac{R(s_k,a_k)\gamma^{n-k}}{n}.$$

现在，应用归纳假设，可将其改写为

$$\frac{n-1}{n}\sum_{k=1}^{n-1}\frac{R(s_k,a_k)(1-\gamma^{n-k})}{(n-1)(1-\gamma)}+\sum_{k=1}^{n}\frac{R(s_k,a_k)\gamma^{n-k}}{n}.$$

从整个表达式中提取公因子 $\frac{1}{n(1-\gamma)}$，得到

$$\frac{\sum_{k=1}^{n-1}R(s_k,a_k)(1-\gamma^{n-k})+(1-\gamma)\sum_{k=1}^{n}R(s_k,a_k)\gamma^{n-k}}{n(1-\gamma)}.$$

对分子进行展开，得到表达式

$$\frac{\sum_{k=1}^{n-1}R(s_k,a_k)-R(s_k,a_k)\gamma^{n-k}+\sum_{k=1}^{n}R(s_k,a_k)\gamma^{n-k}-R(s_k,a_k)\gamma^{n+1-k}}{n(1-\gamma)}$$

去掉相加后抵消的项，得到

$$\frac{\sum_{k=1}^{n}R(s_k,a_k)-\sum_{k=1}^{n}R(s_k,a_k)\gamma^{n+1-k}}{n(1-\gamma)}.$$

最后，再次分解分子，得到式 (13)：

$$\sum_{k=1}^{n}\frac{R(s_k,a_k)(1-\gamma^{n+1-k})}{n(1-\gamma)},$$

由此证明，若式 (13) 对长度为 $n-1$ 的路径成立，则对长度为 n 的路径也成立。□

引理 2。 对于环境 *MDP* 中的任意无限路径，

$$\lim_{n\to\infty}\sum_{k=1}^{n}\frac{R(s_k,a_k)\gamma^{n+1-k}}{n(1-\gamma)}=0.$$

证明。 将所述等式左侧分母中的常数项提取出来，可得等价表达式

$$\frac{1}{1-\gamma}\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}R(s_k,a_k)\gamma^{n-k}.$$

由于假设环境 MDP 是有限的，因此可能的奖励值只有有限种，我们可以将上述表达式中的求和界定为

$$\frac{r_{\downarrow}(1-\gamma^{n-1})}{1-\gamma}\le\sum_{k=1}^{n}R(s_k,a_k)\gamma^{n-k}\le\frac{r_{\uparrow}(1-\gamma^{n-1})}{1-\gamma}$$

其中 $r_{\downarrow}=\min_{(s,a)\in S\times A}R(s,a)$ 和 $r_{\uparrow}=\max_{(s,a)\in S\times A}R(s,a)$。最后，注意到

$$\lim_{n\to\infty}\frac{r_{\downarrow}(1-\gamma^{n-1})}{n(1-\gamma)}=\lim_{n\to\infty}\frac{r_{\uparrow}(1-\gamma^{n-1})}{n(1-\gamma)}=0,$$

由此可得

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}R(s_k,a_k)\gamma^{n-k}=0.$$

□

现在我们可以证明定理 6 了。

定理 *6* 的证明。 鉴于公式（10），只需证明恒等式 $V^{\gamma}=V^{1-\gamma}$。应用引理 1，平均折旧收益可改写为

$$\liminf_{n\to\infty}\sum_{k=1}^{n}\frac{R(s_k,a_k)(1-\gamma^{n+1-k})}{n(1-\gamma)}.$$

将分子中的乘积展开，再将求和拆分为两个求和之差，可得表达式

$$\liminf_{n\to\infty}\left(\sum_{k=1}^{n}\frac{R(s_k,a_k)}{n(1-\gamma)}-\sum_{k=1}^{n}\frac{R(s_k,a_k)\gamma^{n+1-k}}{n(1-\gamma)}\right).$$

根据引理 2，当$n\to\infty$时，该差值中的右侧项趋于 0，因此上述表达式等价于

$$\liminf_{n\to\infty}\sum_{k=1}^{n}\frac{R(s_k,a_k)}{n(1-\gamma)}.$$

将分母中的常数项提取出来后，剩余的极限项正是平均收益的定义，因此我们得出：对于任意状态s，有

$$V^{\gamma}(s)=\frac{V(s)}{1-\gamma}.$$

□

作为定理 6 的直接推论，存在一个 Blackwell 最优策略：当 λ 充分接近 1 时，该策略对于V_{λ}^{γ}是最优的，同时对于V^{γ}也是最优的。

推论 2。 存在一个折扣因子 $\lambda_0\in(0,1)$ 和一个策略 π，使得对于所有 $\lambda\in[\lambda_0,1)$ 以及每个状态 s，都有

$$V_{\lambda}^{\gamma}(s)=\mathbb{E}_s^{\pi}\left[\sum_{n=1}^{\infty}\lambda^{n-1}\sum_{k=1}^{n}R(s_k,a_k)\gamma^{n-k}\right],$$

$$V^{\gamma}(s)=\mathbb{E}_s^{\pi}\left[\liminf_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n}\sum_{i=1}^{k}R(s_i,a_i)\gamma^{k-i}\right].$$

这进而对平均折旧收益的策略复杂度给出了如下结果。

推论 3 (策略复杂度)。 对于任意有限马尔可夫决策过程上的平均折旧收益，存在一个平稳且确定性的最优策略。

5 相关工作

折扣收益与平均收益在最优控制和强化学习理论中一直发挥着核心作用。已有大量深刻成果将这些目标联系起来[Andersson和Miltersen，2009年；Bewley和Kohlberg，1976年、1978年；Chatterjee和Majumdar，2012年；Chatterjee等，2011年；Mertens和Neyman，1981年；Ziliotto，2016年（a）、（b），2018] 此外，关于相关优化问题的算法及其复杂度，也已有广泛的研究 [Chatterjee和Ibsen-Jensen，2015年；Chatterjee等，2008年；Filar和Schultz，1986年；Raghavan和Filar，1991年；Raghavan和Syed，2003]。

折旧资产的价值定义为奖励过去折扣和。有限序列的过去折扣和曾在优化领域得到研究 [Alur等，2012] 并且与指数近期加权平均密切相关；指数近期加权平均是一种用于非平稳多臂老虎机问题的技术 [Sutton和Barto， 2018] 通过赋予近期结果更高的权重，来估计不同动作的平均奖励。然而，据我们所知，折旧资产尚未作为收益函数得到正式研究。

折扣目标已在程序验证和综合领域得到广泛应用 [Cerný 等，2011年；de Alfaro 等， 2003]。尽管过去时算子的概念由来已久 [Lichtenstein 等， 1985],，但直到相对近期，包括 LTL 和 CTL 等时序逻辑以及模态 μ-演算在内的多种经典形式，才扩展出过去时算子和折扣定量语义 [Almagor 等，2014、2016年；de Alfaro 等，2005年；Littman 等， 2017]。一项特别重要的结果 [Markey， 2003]关于采用经典布尔语义的 LTL：尽管带过去时算子的 LTL 表达能力不强于标准 LTL，但其表达简洁度高出指数级。对于其他逻辑及其过去时算子扩展，在折扣定量语义下进行解释时，是否也存在这种关系，仍是一个悬而未决的问题 [Almagor 等， 2016]。

6 结论

在随机最优控制和强化学习的设定中，智能体选择动作，以最大化与由此产生的标量奖励序列相关的折扣收益。这种交互模式刻画了多巴胺驱动的生物体如何凭借延迟满足（折扣）的能力来最大化其奖励序列。对于即时奖励流而言，这一范式提供了一个自然的模型；然而，当估值和目标是以会折旧的资产来定义时，便无法直接用经典框架对该问题建模。我们开启了对折旧资产的优化与学习研究，并揭示了这类问题与传统折扣问题之间一个出人意料的联系。我们的结果使得人们能够通过调整过去几十年来为经典优化问题广泛开发的算法基础设施，来求解具有折旧动态的优化问题。

我们认为，折旧资产可以为许多相关问题提供一种有用的抽象。以下几点概述了其中一些研究方向，并指出若干尚未解决的问题。

▶ 遗憾最小化 [Cesa-Bianchi 和 Lugosi， 2006] 是在线学习场景中一种广泛采用的标准。在该场景中，决策者选择自己的动作，以最小化平均遗憾——即实际获得的奖励与本可获得的奖励之间的差值。我们认为，不完美的决策者可能会以折旧的方式看待自己的遗憾，因为近期的一个次优动作往往比遥远过去同样次优的动作造成更多遗憾。我们希望，本研究的结果能够推动人们进一步关注在线学习与优化中对过去进行折扣的遗憾刻画基础。

▶ 在求解多智能体优化问题时，一种实际的假设是：通过假定任何对手的能力都受到限制，即其对交互历史的记忆有限，来对其能力设定上界；这可以通过对过去的结果进行折扣来建模。根据我们的结果，具有折旧动态的双人零和博弈，无论采用折扣收益还是平均收益，都可以归约为经典优化博弈，只需对即时奖励进行某种缩放。

▶ 基于状态的折扣因子这一概念已在经典优化和学习的背景下得到研究。是否可以将本文的结果扩展到具有依赖于状态的折旧因子的情形？这一结果无法直接由本文所开发的工具推出，目前仍是一个开放问题。

▶ 连续时间马尔可夫决策过程为离散时间马尔可夫决策过程提供了稠密时间的对应形式，针对这类系统的优化和强化学习算法已经得到了充分研究。是否可以针对具有折旧资产的连续时间马尔可夫决策过程解决优化和学习问题?

参考文献

Susan Rose Ackerman。二手车作为折旧资产。经济探询，11(4):463，1973。

Shaull Almagor、Udi Boker 和 Orna Kupferman。LTL 中的折扣。在*系统构造与分析工具和算法，TACAS*中，收录于第 8413 卷《计算机科学讲义》，第 424–439 页。施普林格，2014 年。网址：https://doi.org/10.1007/978-3-642-54862-8_37。

Shaull Almagor、Udi Boker 和 OrnaKupferman。关于质量的形式化推理。《美国计算机协会会刊》，63(3):24:1–24:56，2016年。网址https://doi.org/10.1145/2875421。

Rajeev Alur、Loris D'Antoni、Jyotirmoy V. Deshmukh、Mukund Raghothaman 和 Yifei Yuan。正则函数、代价寄存器自动机与广义最小代价问题，2012年。

DanielAnderssonandPeterBro Miltersen。求解图上随机博弈的复杂度。载于算法与计算 *ISAAC*，第5878卷《计算机科学讲义》，第112–121页。施普林格，2009年。网址 *https://doi.org/10.1007/ 978-3-642-10631*-6_13。

Stefan Banach。关于抽象集合上的运算及其在积分方程中的应用。*Fund. math*，3 (1):133–181，1922年。

Paul Beckerman。《高通货膨胀经济学》。施普林格，1991年。

Truman Bewley 和 Elon Kohlberg。随机博弈的渐近理论。《运筹学数学》，1(3): 197–208，1976。

Truman Bewley 和 Elon Kohlberg。《具有平稳最优策略的随机博弈》。运筹学数学，3(2):104–125，1978年。ISSN 0364765X, 15265471。网址http://www.jstor.org/stable/3689337。

Oscar R Burt。《折旧统一理论》。会计研究期刊，第28–57页，1972年。

Pavol Cerný、Krishnendu Chatterjee、Thomas A. Henzinger、Arjun Radhakrishna 和 Rohit Singh。《并发程序的定量综合》。载于计算机辅助验证 *CAV*，收录于第 6806 卷《计算机科学讲义》，第 243–259 页。施普林格，2011年。网址https://doi.org/10.1007/978-3-642-22110-1_20。

Nicolò Cesa-Bianchi 和 Gábor Lugosi。《预测、学习与博弈》。剑桥大学出版社，2006年。ISBN 978-0-521-84108-5。网址 https://doi.org/10.1017/CBO9780511546921。

Krishnendu Chatterjee 和 Rasmus Ibsen-Jensen。并发平均收益博弈的定性分析。信息与计算，242：2–24，2015年。网址 https://doi.org/10.1016/j.ic.2015.03.009。

Krishnendu Chatterjee 和 Rupak Majumdar。跨时间尺度博弈中的折扣与平均。国际计算机科学基础期刊，23(3):609–625，2012年。网址https://doi.org/10.1142/S0129054112400308。

Krishnendu Chatterjee、Rupak Majumdar 和 Thomas A. Henzinger。随机极限平均博弈属于指数时间复杂度。国际博弈论期刊，37(2):219–234，2008年。网址 https://doi.org/10.1007/s00182-007-0110-5。

Krishnendu Chatterjee、Laurent Doyen 和 Rohit Singh。关于无记忆定量目标。载于计算理论基础 *FCT*，第 6914 卷《计算机科学讲义》，第 148–159 页。施普林格，2011年。网址 https://doi.org/10.1007/978-3-642-22953-4_13。Pete Comley。《style id='8'>《通胀至关重要：通胀波理论、其对过去和现在通货膨胀的影响……以及即将到来的通货紧缩。Pete Comley，2015年。Luca de Alfaro、Thomas A. Henzinger 和 Rupak Majumdar。在系统理论中折扣未来。载于自动机、语言与编程 *ICALP*，第 2719 卷《计算机科学讲义》，第 1022–1037 页。施普林格，2003年。网址 https://doi.org/10.1007/3-540-45061-0_79。Luca de Alfaro、Marco Faella、Thomas A. Henzinger、Rupak Majumdar 和 Mariëlle Stoelinga。折现时序性质的模型检测。理论计算机科学，345(1):139–170，2005年。网址 https://doi.org/10.1016/j.tcs.2005.07.033。Eugene A Feinberg 和 Adam Shwartz。《马尔可夫决策过程手册：方法与应用》，第 40 卷。施普林格科学与商业媒体，2012年。Adam Fergusson。《style id='23'>《当货币消亡时。Old Street Publishing，2010年。Jerzy Filar 和 Koos Vrieze。《style id='25'>竞争性马尔可夫决策过程。施普林格出版社，1996年。

Jerzy A. Filar 和 Todd A. Schultz。随机博弈中的非线性规划与平稳策略。数学规划，34(2):243–247，1986年。网址 https://doi.org/10.1007/BF01580590。

Shane Frederick、George Loewenstein 和 Ted O'donoghue。时间贴现与时间偏好：一篇批判性综述。经济文献期刊，40(2):351–401，2002年。

Geoffrey Heal。折扣：基础经济学综述。芝加哥大学法律评论，74:59，2007年。

CharlesR Hultenand Frank C Wykoff。《经济折旧的测量》。华盛顿城市研究所，1980年。

Orna Lichtenstein、Amir Pnueli 和 Lenore D. Zuck。《过去的荣耀》。载于《程序逻辑会议》，第193卷《计算机科学讲义》，第196–218页。施普林格，1985年。网址：https://doi.org/10.1007/3-540-15648-8_16。

Michael L. Littman、Ufuk Topcu、Jie Fu、Charles Isbell、Min Wen 和 James MacGlashan。基于广义线性时序逻辑的环境无关任务规范，2017年。

George Loewenstein 和 Elster Jon，编者。《跨时选择》。罗素·塞奇基金会，1992年。

Nicolas Markey。带过去时态的时序逻辑表达更简洁，且简洁度呈指数级，并发专栏。《欧洲理论计算机科学协会会刊》，79：122–128，2003年。

J.F. Mertens 和 Abraham Neyman。随机博弈。国际博弈论期刊，10(2):53–66，1981年。Cédric Philibert。气候变化经济学与折扣理论。《能源政策》，27(15):913–927，1999年。Martin L. Puterman。《马尔可夫决策过程：离散随机动态规划》。《概率与统计丛书》。威利，1994年。ISBN 978-0-47161977-2。doi: 10.1002/9780470316887。URLhttps://doi.org/10.1002/9780470316887。T. E. S. Raghavan 和 Jerzy A. Filar。随机博弈算法——综述。*ZOR方法、模型与运筹研究*，35(6):437–472，1991年。URLhttps://doi.org/10.1007/BF01415989。T. E. S. Raghavan 和 Zamir Syed。一种用于求解完全信息双人零和随机博弈的策略改进型算法。数学规划，95(3):513–532，2003年。URL https://doi.org/10.1007/s10107-002-0312-3。Walter Rudin。《数学分析原理》。麦格劳-希尔，纽约，1976。Lloyd S Shapley。随机博弈。《美国国家科学院院刊》，39(10):1095–1100，1953年。R. S. Sutton 和 A. G. Barto。《强化学习：导论》。麻省理工学院出版社，第二版，2018年。Richard S. Sutton 和 Andrew G. Barto。强化学习——导论。自适应计算与机器学习。麻省理工学院出版社，1998年。ISBN 978-0-262-19398-6。URLhttps://www.worldcat.org/oclc/37293240。

Christopher J. C. H. Watkins 和 Peter Dayan。技术札记：Q学习。机器学习，8:279–292，1992年。URL https://doi.org/10.1007/BF00992698。

F Kenneth Wright。论折旧的一般理论。会计研究杂志，第80–90页，1964年。

Frank C. Wykoff。战后时期的资本折旧：汽车。《经济学与统计学评论》，52(2)：168–172，1970年。

Bruno Ziliotto。非扩张算子的陶伯定理及其在零和随机博弈中的应用。《运筹学数学》，41(4)：1522–1534，2016年（a）。URLhttps://doi.org/10.1287/moor.2016.0788。

Bruno Ziliotto。零和随机博弈中的一般极限值。《国际博弈论杂志》，45(1-2)：353–374，2016年（b）。URLhttps://doi.org/10.1007/s00182-015-0509-3。

Bruno Ziliotto。算子一般迭代的陶伯定理：在零和随机博弈中的应用。《博弈与经济行为》，108：486–503，2018年。URL https://doi.org/10.1016/j.geb.2018.01.009。