半 Markov 决策过程的研究进展

郭先平*, 黄永辉

中山大学数学与计算科学学院, 广州 510275

E-mail: mcsgxp@mail.sysu.edu.cn, hyongh5@mail.sysu.edu.cn

收稿日期: 2015-01-25; 接受日期: 2015-02-27; * 通信作者

国家自然科学基金 (批准号: 11471341 和 61374067) 资助项目

摘要 本文综述半 Markov 决策过程 (SMDP) 理论的发展现状, 主要介绍 SMDP 无限阶段期望折扣报酬准则、长期平均准则、有限阶段期望报酬准则、首达目标期望报酬准则、概率准则、受约束问题和均值 - 方差准则的研究工作, 着重阐述这些优化准则的背景、意义、主要研究进展及有待研究的问题. 最后, 展望 SMDP 未来的一些潜在研究方向和相关问题.

关键词 半 Markov 决策过程 优化准则 最优方程 最优策略 计算方法

MSC (2010) 主题分类 90C40, 93E20

1 引言

Markov 决策过程 (Markov decision processes, MDP) (又称 Markov 控制过程或随机动态规划) 是研究随机动态系统的最优控制问题, 其主要研究对象是转移结构受控的随机动态系统. 根据系统的状态, 决策者 (如人类或计算机) 选取一个策略来控制或影响系统的转移, 从而每个策略可定义一个随机过程和相应于该过程的目标函数. MDP 的目的是选取一个“好”的控制策略. 在现实生活中, 决策者采取的策略通常有两种影响: (i) 它们会产生效益 (如花费、节约的时间、资金、其他资源或收益); (ii) 通过改变系统的状态转移, 它们对系统将来的运行产生影响. 有许多的情形, 产生很大当前效益的决策用将来观点看时并不是“好”的, MDP 就是模拟这样的情形, 并给出“好策略”的结构、存在性及其计算方法等. 仅含一个赌徒的随机对策即为 MDP, 因此, 许多专家认 MDP 的产生源于文献 [1] 随机对策的工作. 一般来说, 按系统状态逗留时间的分布和采取控制行动的时间, MDP 可分为下列三类: (i) 状态逗留时间为常数且仅在固定时间点上采取控制 (又称决策) 的离散连时间 Markov 决策过程 (discrete-time MDP, DTMDP); (ii) 状态逗留时间服从指数分布但可在任意时间上采取控制的连续时间 Markov 决策过程 (continuous-time MDP, CTMDP); (iii) 状态逗留时间可有任意分布但仅在离散的跳跃点上采取控制的半 Markov 决策过程 (semi-Markov decision processes, SMDP). 文献 [2] 对 MDP 进行了开创性研究. 文献 [3-5] 分别奠定了 DTMDP、CTMDP 和 SMDP 的理论基础. 对 DTMDP 的研究, 成果丰富, 文献众多, 无法例举. 从下列专著及其参考文献, 可对 DTMDP 的理论研究和实际应用有一定了解: 文献 [6-8] 对可数状态 DTMDP 的理论与应用给予了系统分析; 文献 [9, 10] 对一般状态 DTMDP 的理论发展进行了深刻论述; Piunovskiy^[11] 收集了大量 MDP 的例子; Feinberg 和 Shwartz^[12] 介绍了 DTMDP 中的众多研究方向及其在通信网络、水库管理、数理金融和排队系统中的

英文引用格式: Guo X P, Huang Y H. A survey on semi-Markov decision processes (in Chinese). Sci Sin Math, 2015, 45: 477-496, doi: 10.1360/N012015-00041

大量应用; Bauerle 和 Rieder^[13] 概述了 DTMDP 折扣准则和有限阶段准则的理论成果, 尤其是分析了它们在金融与保险中的大量应用; Hu 和 Yue^[14] 讨论了 DTMDP 在管理中的应用; Sigaud 和 Buffet^[15] 总结了 MDP 在人工智能方面的应用; Mahadevan^[16] 分析了 DTMDP 的动态系统学习控制理论、算法和应用等; Simar 和 Parsons^[17] 对 DTMDP 和 BDI (belief-desire-intention) 两种模型及其关系进行了统一分析. 关于 CTMDP 的研究也取得系列进展: 连续时间 MDP 的研究源于 Howard^[2] 的开创性工作. 之后, Gihman 和 Skorohod^[4] 考虑了一般状态有限阶段模型及转移率有界的可数状态折扣模型; Puterman^[7] 和 Sennott^[8] 考虑了状态空间可数的平均和折扣准则; 文献 [18, 19] 对可数 CTMDP 折扣与平均准则的理论和应用进行了系统分析; Guo 等人^[20, 21] 研究了一般状态 CTMDP 的折扣与平均模型. 相对 DTMDP 的研究来说, 连续时间 MDP 的理论研究还不完善, 需要进一步研究. 例如, 从事 Markov 决策过程研究的国际著名学者 Feinberg 教授在 Mathematical Reviews 和 Zentralblatt MATH 上均说: “连续时间 Markov 决策过程的研究领域还未完全开展, 确信将来会进行大量研究”.

SMDP 由 Jewell^[5] 和 Howard^[25] 将 MDP 推广到 Markov 更新 (半 Markov) 过程而建立, 故又称为 Markov 更新规划 (Markov renewal programs). 文献 [26–29] 等对 SMDP 理论作了早期的探索. 由于 SMDP 的状态逗留时间可为任意概率分布, 与逗留时间为固定常数的 DTMDP 和逗留时间服从指数分布的 CTMDP 相比, 更具有一般性和广泛性, 因而在工程技术和社会经济中具有广阔的应用前景和研究潜力. 自 20 世纪 60 年代以来, 伴随着实际问题的推动, 新的优化准则、优化问题不断被提出, SMDP 的研究领域也不断被拓宽, 研究成果日趋丰富, 其实际应用范围已深入到通信网络、设备维修、资源分配、水库发电、排队优化、金融市场和流行病模型等众多领域.

本文将综述 SMDP 的研究进展, 介绍 SMDP 最新的理论成果. 我们在第 3–9 节分别介绍 SMDP 的折扣准则、平均准则、有限阶段准则、首达目标准则、概率准则、受约束问题和均值 - 方差问题的研究工作. 将着重阐述它们的背景、意义、现有主要工作及有待研究的问题. 最后, 讨论 SMDP 未来的一些潜在发展方向.

2 SMDP 模型

SMDP 的模型由如下五元素构成:

$$\{E, A, (A(x), x \in E), Q(dt, dy | x, a), r(x, a)\}, \quad (2.1)$$

其中, E 和 A 分别是状态空间和行动空间, 假定它们均为 Borel 空间并分别赋予 Borel σ -代数 $\mathcal{B}(E)$ 和 $\mathcal{B}(A)$; $A(x) \in \mathcal{B}(A)$ 表示在状态 $x \in E$ 时的允许行动集; $Q(dt, dy | x, a)$ 是半 Markov 核, 它是给定 K 时定义在 $\mathbb{R}_+ \times E$ 的随机核, 其中, $K := \{(x, a) | x \in E, a \in A(x)\}$ 是允许状态 - 行动集, $\mathbb{R}_+ := [0, \infty)$; $r(x, a)$ 是报酬 (费用) 率函数, 且在 K 上可测.

根据 Radon-Nikodym 定理, 半 Markov 核 Q 可以作如下分解:

$$Q(t, D | x, a) = \int_D F(t | x, a, y) p(dy | x, a) \quad \forall t \in \mathbb{R}_+, \quad D \in \mathcal{B}(E), \quad (x, a) \in K, \quad (2.2)$$

其中, $F(\cdot | x, a, y)$ 表示当选取行动 a 且下一个状态转移到 y 时在状态 x 逗留时间的分布, 而 $p(\cdot | x, a)$ 表示系统状态的转移概率.

与 DTMDP 和 CTMDP 类似, 在 SMDP 中, 若系统在某决策时刻处于状态 $x \in E$, 决策者根据某个规则选取一个行动 $a \in A(x)$, 那么将发生下列两个事件 (不管系统以前的历史如何): (i) 以报酬率

$r(x, a)$ 连续地发生报酬直到下一决策时刻; (ii) 系统状态以概率 $Q(dt, dy | x, a)$ 在 x 处停留 t 个时间单位, 然后转移到 $y \in E$, 接着下一决策时刻发生.

SMDP 的动态演化过程可以如下描述: 在初始决策时刻 t_0 , 假设系统处于状态 x_0 , 决策者根据某规则选取一个行动 $a_0 \in A(x_0)$, 由于这一行动的选取, 系统状态依据半 Markov 核 Q 停留在状态 x_0 直至时刻 t_1 , 然后转移到新的状态 x_1 , 接着下一决策时刻发生. 决策者根据某规则, 在 t_1 时刻选取行动 $a_1 \in A(x_1)$, 类似的事件相继发生, 过程如此进行下去, 从而得到 SMDPs 第 n 个跳跃时刻的允许历史 h_n , 即

$$h_n = (t_0, x_0, a_0, \dots, t_{n-1}, x_{n-1}, a_{n-1}, t_n, x_n),$$

其中, $t_{m+1} \geq t_m \geq 0$, $(x_m, a_m) \in K$, $m = 0, 1, \dots, n-1$, 且 $(t_n, x_n) \in \mathbb{R}_+ \times S$. 令 H_n 表示所有允许历史 h_n 组成的集合, 并赋予相应的 Borel σ -代数.

为明确选取行动的规则, 需要引入策略的定义.

定义 2.1 依赖于历史的随机策略 (简称策略) $\pi = \{\pi_n\}$ 是满足以下条件的随机核序列: 每个 π_n 是给定 H_n 时定义在 A 上的随机核, 且

$$\pi_n(A(x_n) | h_n) = 1, \quad \forall h_n \in H_n, \quad n = 0, 1, \dots$$

所有策略构成的集合记为 Π .

在大部分情形下, 控制系统运行的最优策略并不需要在 Π 中寻找, 而往往在 Π 的一些子集中就存在. 为此, 我们引入一些子策略类.

令 Φ 表示给定 E 时定义在 A 上满足 $\varphi(A(x) | x) \equiv 1$ 的所有随机核 φ 组成之集. $\mathbb{F}$ 表示决策函数 f 之集, 其中, $f: E \rightarrow A$ 可测, 且 $f(x) \in A(x)$ 对任意 $x \in E$ 成立.

定义 2.2 (i) 一个策略 $\pi = \{\pi_n\} \in \Pi$ 称为随机 Markov 的, 如果存在随机核序列 $\{\varphi_n\}$ ($\varphi_n \in \Phi$) 使得对每个 $h_n = (t_0, x_0, a_0, \dots, t_{n-1}, x_{n-1}, a_{n-1}, t_n, x_n) \in H_n$ 和 $n \geq 0$, $\pi_n(\cdot | h_n) = \varphi_n(\cdot | x_n)$. 随机 Markov 策略记为 $\pi = \{\varphi_n\}$.

(ii) 随机 Markov 策略 $\pi = \{\varphi_n\}$ 称为随机平稳的, 如果 φ_n 与 n 无关. 此时, π 记作 φ .

(iii) 随机 Markov 策略 $\pi = \{\varphi_n\}$ 称为确定性 Markov 的, 如果存在决策函数序列 $\{f_n\}$ ($f_n \in \mathbb{F}$) 使得对每个 $x \in E$ 和 $n \geq 0$, $\varphi_n(\cdot | x)$ 是在 $f_n(x)$ 上的 Dirac 测度. 此类策略记为 $\pi = \{f_n\}$.

(iv) 确定性 Markov 策略 $\pi = \{f_n\}$ 称为 (确定性) 平稳的, 若 f_n 与 n 无关. 简记这样的 π 为 f .

分别用 Π_{RM} , Φ , Π_{DM} 和 $\mathbb{F}$ 表示所有随机 Markov 策略、随机平稳策略、确定性 Markov 策略和平稳策略之集. 明显地, $\Phi \subset \Pi_{RM} \subset \Pi$, $\mathbb{F} \subset \Pi_{DM} \subset \Pi$.

对给定的初始状态 $x \in E$ 和策略 $\pi = \{\pi_n\} \in \Pi$, 由熟知的 Ionescu Tulcea 定理, 用半 Markov 核 Q 和策略 π , 我们可构造唯一的概率空间 $(\Omega, \mathcal{F}, P_x^\pi)$, 并定义该空间上的随机过程 $\{(S_n, X_n, A_n), n = 0, 1, \dots\}$, 其中, 随机变量 S_n, X_n 和 A_n 分别表示 SMDP 第 n 个跳跃时刻、第 n 个跳跃时刻系统的状态和所采取的行动. 用 E_x^π 表示相应于概率测度 P_x^π 的期望算子. 利用 $\{(S_n, X_n, A_n)\}$, 定义相应的连续时间半 Markov 决策过程 (即状态 - 行动过程) $\{Z(t), W(t)\}$ (依赖于给定的策略 π , 此处省略):

$$Z(t) = J_n, \quad W(t) = A_n, \quad S_n \leq t < S_{n+1}, \quad t \in \mathbb{R}_+, \quad n \geq 0.$$

一般地, 为了使讨论的问题有意义, 必须保证过程 $\{Z(t), W(t), t \geq 0\}$ 在有限时间内系统发生跳跃 (即状态改变) 的次数不是无穷的. 为此, 需要一个称为“正则性条件”的假设.

假设 2.1 (正则性条件) 存在常数 $\delta > 0$ 和 $1 > \epsilon > 0$, 使得对所有 $(x, a) \in K$, 有

$$Q(\delta, E \mid x, a) \leq 1 - \epsilon.$$

1970 年, Ross^[30] 首先引入了 SMDP 模型的正则性条件, 随后, 正则性条件在 SMDP 文献中被广泛使用, 参见文献 [7, 31–34].

在正则性假设 2.1 下, 对任给初始状态 $x \in E$, 策略 $\pi \in \Pi$, 我们可以根据实际需要定义某种优化准则 $V^\pi(x)$ (准确的定义见下面 (3.1)、(4.1)、(4.2)、(5.1)、(5.2)、(6.1) 和 (7.1)).

定义 2.3 对于给定的准则 $V^\pi(x)$, 称 $V^*(x) = \sup_{\pi \in \Pi} V^\pi(x)$ 或 $V^*(x) = \inf_{\pi \in \Pi} V^\pi(x)$ (何种情形根据是报酬准则还是费用准则而定) 为最优值函数. 对任意给定的 $\epsilon \geq 0$, 一个策略 $\pi^* \in \Pi$ 称为 ϵ -最优策略, 若对任意 $x \in E$, 它满足 $V^{\pi^*}(x) \geq \sup_{\pi \in \Pi} V^\pi(x) - \epsilon$ 或 $V^{\pi^*}(x) \leq \inf_{\pi \in \Pi} V^\pi(x) + \epsilon$. 特别地, 0-最优策略简称为最优策略.

与 DTMDP 和 CTMDP 的一样, 研究 SMDP 模型的主要任务是, 对于给定的某个优化准则, 寻求相应于该准则的 ϵ -最优策略的存在性条件、计算方法及其特征和应用.

我们在第 3–9 节分别对无限阶段期望折扣报酬准则、长期平均报酬准则、有限阶段期望报酬准则、首达目标期望报酬准则、概率准则、受约束问题和均值 - 方差问题, 阐述各准则 (问题) 的背景、意义、研究现状及有待研究的问题. 最后, 介绍可能的研究方向.

3 无限阶段期望折扣报酬准则

给定折扣因子 $\alpha \geq 0$. 对任给初始状态 $x \in E$, 策略 $\pi \in \Pi$, SMDP 的无限阶段期望折扣准则 (用 $V_\alpha^\pi(x)$ 表示) 定义为

$$V_\alpha^\pi(x) := E_x^\pi \left[\int_0^\infty e^{-\alpha t} r(Z(t), W(t)) dt \right]. \quad (3.1)$$

折扣因子 $\alpha > 0$ 的引入有两种解释: 从数学角度看, 期望总报酬 $E_x^\pi [\int_0^\infty r(Z(t), W(t)) dt]$ 不一定存在. 为此, 常常引进一个折扣因子 α , 使得 $E_x^\pi [\int_0^\infty e^{-\alpha t} r(Z(t), W(t)) dt]$ 有意义. 从经济的角度看, 现在的单位资金量在将来会指数增长, 如现在的一元钱, 当存款年利率为 ρ 时, 第 n 年变为 $(1 + \rho)^n$ 元; 等价地, 将来第 n 年时的 1 元钱仅值现在的 $(1 + \rho)^{-n}$ 元. 由于决策的制定是在当下, 将来的资金需要折算到现在来算, 为此引入折扣因子 α , 亦称贴现率.

从 (3.1) 可以看出, 在期望折扣准则中, 愈是周期小的费用看得愈重要. 因此, 期望折扣准则本质上是一个前面有限阶段的费用准则函数, 后面无限多个阶段不起多大作用 (收敛级数的尾数趋于 0). 特别地, 当折扣因子 $\alpha = 0$ 时, 期望折扣准则就成为无限阶段期望总费用准则. 因此, 期望总费用准则可以归结到期望折扣准则中.

无限阶段期望折扣报酬准则是 SMDP 文献中研究最早、成果最完善的准则.

1975 年, Lippman^[35] 考虑了非一致有界即时报酬 (仅在跳跃时刻发生报酬)、状态和行动集均为 Polish 空间的折扣准则 SMDP, 其中, 逗留时间分布和状态转移概率分别由 $F(\cdot \mid x, a)$ 和 $p(\cdot \mid x, a)$ 描述. 他将通常的正则性条件放松为

$$\sup_{(x,a) \in K} \beta_\alpha(x, a) < 1, \quad \beta_\alpha(x, a) := \int_0^\infty e^{-\alpha t} Q(dt, E \mid x, a),$$

并在一定的连续 - 紧条件 (即 $A(x)$ 均为紧的, $r(x, a)$ 和 $Q(dt, dy | x, a)$ 关于 $a \in A(x)$ 分别连续和强连续) 及如下假设: “存在正整数 n , 正实数 M, b 和一个定义在 E 上的权范数 $w \geq 1$, 使得对任意 $(x, a) \in K$, (i) $|r(x, a)| \leq Mw^m(x)$, (ii) $\int_E w^n(y)p(dy | x, a) \leq (w(x) + b)^n$, $n = 1, 2, \dots, m$ ” 下, 利用不动点定理, 证明了值函数 $V_\alpha^*(x)$ 是下列折扣最优方程的唯一解:

$$V_\alpha^*(x) = \sup_{a \in A(x)} \left[r(x, a) + \beta_\alpha(x, a) \int_E V_\alpha^*(y)p(dy | x, a) \right], \quad x \in E,$$

以及达到上述方程右边最大值的平稳策略 f 是最优的. 他还在较弱条件下讨论了 ε - 最优平稳策略的存在性. 但值得注意的是, 作者并未给出当状态空间和行动空间同时为 Polish 空间情形下的最优性结果. 1987 年, Wakuta [36] 完善了 Lippman [35] 的结果, 导出了当状态空间和行动空间同时为 Polish 空间情形下的 Bellman 最优方程和最优平稳策略存在条件.

1975 年, Wessels 和 van Nunen [37] 考虑了有限状态和行动的折扣准则 SMDP, 不同于 Lippman [35] 的不动点和压缩算子方法, 他们通过定义占有测度将折扣准则 SMDP 优化问题转化为了一个线性规划算法.

1977 年, Feldman [38] 考虑了由 SMDP 描述的冲击 (shock) 模型, 其中冲击在随机时间点发生, 并引起相应的系统损坏, 若在系统失效时更换部件, 需要较大的更换费用, 若在系统失效前更换部件, 则更换费用较小, 作者在无限阶段折扣费用准则下寻找到了最优更换策略.

1979 年, Sadjadi 和 Bestwick [39] 运用动态规划的值迭代方法消除折扣 SMDP 的非最优行动, 建立了一个非最优性检验以辨别当前阶段不能为最优的行动, 拓展了 DTMDP 文献的相应结果.

1980 年, Porteus [40] 改进了有限状态折扣 SMDP 最优策略的值迭代和策略迭代的数值算法, 降低了计算复杂度. 1981 年, Porteus [41] 进一步讨论了折扣 SMDP 最优策略的数值计算方法, 通过实验观察, 揭示了 Gauss-Seidel 消去法在一般情形并不优于 Jacobi 消去法.

1985 年, Schweitzer 等人 [42] 利用迭代聚合与解聚方法 (iterative aggregation disaggregation procedure) 求解有限状态 SMDP 的折扣方程. 库存数值实验表明, 当转移矩阵的谱半径接近于 1 时, 此算法比通常的迭代算法 (逐步逼近算法) 显著地快速; 当转移矩阵的谱半径在 0.8 附近时, 此算法会慢一些. 对于大结构问题 (large structured problems), 此算法似乎呈现出较大的优势, 它经常能降低计算时间和储存空间.

1985 年, Hernández-Lerma [43] 考虑可数状态、折扣瞬时报酬的自适应 SMDP, 模型依赖于一个未知参数 θ . 为求解该自适应 SMDP, 作者首先在瞬时报酬有界、正则性条件和连续 - 紧假设等下建立了非平稳 SMDP 的迭代方案 (iterative scheme) 以确定最大总折扣报酬, 并基于此计算得到渐近折扣最优 (自适应) 策略.

1987 年, 董泽清和刘克 [44] 发展了 Lippman [35] 的工作, 研究了无界报酬折扣 SMDP 最优策略的结构. 他们证明了, 任给一策略, 若它是最优的, 则存在随机平稳策略, 对同一折扣因子也是最优的; 任一随机折扣最优平稳策略必可分解为若干个确定性平稳最优策略的凸组合, 较完满地解决了该模型最优策略的结构问题.

1989 年, Bhattacharya 和 Majumdar [45] 在 E 为 Borel 空间、 A 为紧的度量空间和报酬无界的情形下证明了折扣 SMDP 最优平稳策略的存在性, 导出了相应的动态规划方程, 总结和推广了较早的已有工作.

1993 年, 胡奇英 [46] 将通常的无限阶段平稳折扣 SMDP 推广到随机终止的非平稳折扣 SMDP, 其模型参数与决策周期数有关, 终止时间 ξ 是一个与过程独立的随机变量, 服从某一分布. 特别地, 当

$P(\xi = \infty) = 1$ 时则退化为无限阶段的情形. 在正则性条件和报酬函数一致有界的条件下, 作者首先把随机终止的非平稳折扣 SMDP 化为无限阶段非平稳折扣 SMDP, 然后再进一步化为无限阶段非平稳折扣 DTMDP, 建立了相应的最优方程, 讨论了 ε -最优策略的存在性.

1994 年, Puterman [7] 总结了有界和无界报酬折扣 SMDP 的一般性结果, 如值函数是最优方程的唯一解和最优平稳策略的存在性条件以及值迭代和策略改进算法等.

1997 年, Hu [47] 研究了半 Markov 环境中的折扣 SMDP, 它的模型参数受半 Markov 环境的影响和确定, 即随着环境状态的转移, SMDP 的参数也改变. 在假设环境的状态为可数、SMDP 的状态和行动空间皆为 Borel 空间、报酬函数无界的情形下作者证明了最优方程的和 ε -最优策略的存在性.

2004 年, Hudak 和 Nollau [48] 考虑了可数状态、有限行动的无限阶段折扣 SMDP 的逼近解. 他们给出了值函数上界和下界的收敛序列, 并导出了消除次优 (非最优) 行动的方法.

2005 年, Luque-Vasquez 和 Minjarez-Sosa [49] 处理了 Borel 状态和行动空间, 费用可能无界和逗留时间分布 H 未知的一类无限阶段折扣准则 SMDP. 在逗留时间分布 H 不依赖于状态和行动的假设下, 作者联合 H 的统计估计方法与 SMDP 理论构造了近似折扣最优策略序列 $\hat{\pi} = \{\hat{\pi}_n\}$ 和一个最优平稳策略 f , 其中, $\hat{\pi}_n$ 收敛于 f . 2011 年, Luque-Vasquez 等人 [50] 进一步考虑了逗留时间分布依赖于一个随阶段变化的可能不能观察的未知参数的情形, 他们将此 SMDP 化为一个随机对策 (game), 得到了折扣和平均费用准则最优策略的存在性.

2008 年, Yin 等人 [51] 研究了有限状态和紧致行动空间的折扣准则 SMDP 的性能优化问题. 他们首先定义 SMDP 的无穷小生成元矩阵, 基于此导出 SMDP 的折扣 Poisson 方程和 α -势能 (α -potential), 建立了不同于 Lippman [35] 的最优方程, 讨论了折扣模型与平均模型的关系, 得到了平均准则下的相应结果. 作者还给出了寻找 ε -最优策略的值迭代和策略迭代算法, 并证明了其收敛性.

2010 年, 黄永辉和郭先平 [52] 讨论了可数状态和行动、非负费用率的折扣 SMDP, 他们假设费用函数仅有下界而无上界, 利用不同于不动点理论的最小非负解方法证明了最优值满足最优方程且存在 ε -最优平稳策略, 并在允许行动集有限的条件下, 进一步证明了存在最优平稳策略, 且值函数是最优方程的最小非负解. 他们还给出了计算值函数和最优策略的值迭代算法.

2012 年, Rosas-Rosas [53] 考虑了当逗留时间分布 $F(\cdot | x, a)$ 未知时折扣 SMDP 最优值函数的估计误差, 即若记 $\hat{F}(\cdot | x, a)$ 是真值 $F(\cdot | x, a)$ 的估计值, 则对估计值 $\hat{F}(\cdot | x, a)$ 可以求解一个折扣最优策略 $\hat{\pi}^*$, 对真值 $F(\cdot | x, a)$ 也可以求解一个相应的折扣最优策略 π^* , 作者致力于研究偏差 $D(x, \hat{\pi}^*) = V_{\alpha}^{\hat{\pi}^*}(x) - V_{\alpha}^{\pi^*}(x)$, 在正则性假设和连续 - 紧条件下, 证明了 $D(x, \hat{\pi}^*) \leq \Psi(d(\hat{F}, F))$, 其中, $\Psi: [0, \infty) \rightarrow [0, \infty)$ 是满足 $\lim_{s \rightarrow 0} \Psi(s) = 0$ 的一个函数, d 是关于分布函数 $\hat{F}$ 与 F 在某种意义下的距离. 基于 $F(\cdot | x, a)$ 的相合估计 $\hat{F}_n(\cdot | x, a)$ (与样本容量 n 有关), 该结果还应用到构造近似折扣最优策略.

注 3.1 就我们所知, 还未见折扣最优策略特征和结构 (如单调性最优策略和 (s, S) -最优策略) 的讨论.

4 长期平均期望报酬准则

类似于文献 [30, 33] 的定义, SMDP 的平均期望报酬准则有以下两个形式:

$$V_1^{\pi}(x) := \liminf_{T \rightarrow \infty} \frac{E_x^{\pi}[\int_0^T r(Z(t), W(t))dt]}{T}, \quad x \in E, \quad \pi \in \Pi, \quad (4.1)$$

$$V_2^\pi(x) := \liminf_{n \rightarrow \infty} \frac{E_x^\pi[\int_0^{S_n} r(Z(t), W(t)) dt]}{E_i^\pi[S_n]}, \quad x \in E, \quad \pi \in \Pi, \quad (4.2)$$

其中, 准则 $V_1^\pi(x)$ 是通常意义下的平均期望报酬, 准则 $V_2^\pi(x)$ 表示直到第个 n 跳跃时刻的单位时间平均报酬的下极限, 在一定条件下, 这两个准则是一致的, 参见文献 [7, 30].

由于有限阶段平均期望报酬的极限不一定存在, 所以, 上述平均期望准则的定义考虑其上 (下) 极限, 此时表示在策略 $\pi \in \Pi$ 下可能得到的最“小”平均期望报酬, 说明决策者采取一种“保守”态度. 与期望折扣准则相反, 平均期望准则注重的则是长期的稳定费用, 而近期费用起很小作用. 由此, 平均期望准则适用于相对较快进入“稳态”运行的系统, 如通信网络和排队系统等, 但不适用于金融领域, 因为对资金来说, 现在花的钱比将来花的钱更值钱.

平均准则同折扣准则一样, 是 SMDP 文献研究最多的一种准则. 较早的研究是 1970 年 Ross^[30] 的工作, 他考虑状态和行动空间皆为 Borel 空间的 SMDP 模型, 在正则性条件和“首次返回初始状态的期望时间为有限”的条件下, 证明了上述两种形式的平均准则等价, 进一步, 在费用函数有界和行动空间有限的条件下, 运用消失折扣方法 (the vanishing discount approach), 建立如下的平均最优方程: 存在一个有界可测函数 $h(x)$ 和一个常数 g , 使得

$$h(x) = \inf_{a \in A(x)} \left[\bar{C}(x, a) + \int_E p(dy | x, a) h(y) - g\tau(x, a) \right], \quad x \in E,$$

其中, $\bar{C}(x, a) = \int_0^\infty \int_E c(x, a, t) Q(dt, dy | x, a)$ 是相邻两次跳跃之间的期望费用, $\tau(x, a) := \int_0^\infty \int_E t Q(dt, dy | x, a)$ 是期望逗留时间, 并证明了达到上式最优方程右边最小值的平稳策略 $f(x)$ 是最优的, 且最优值就等于常数 g , 由此得到了最优平稳策略的存在性.

1971 年, Lippman^[54] 研究了一般状态和行动空间 SMDP 期望平均报酬最大化问题, 在假设期望报酬、期望逗留时间一致有界和存在一个期望回访时间一致有界的状态条件下, 他证明了对每个 $\epsilon > 0$, 存在一个 ϵ -最优平稳策略, 且在状态可数和行动有限的假设下, 建立了最优平稳策略的存在性. 其证明方法是强大数定理和动态规划.

1973 年, Lippman^[33] 考虑了一般行动空间而状态空间是非负整数的平均 SMDP. 他假设 $\{0, 1, 2, \dots, n+N\}$ 是状态 n 一步能转移到的状态集合, 其中, N 与 n 独立, 另外, 一步报酬不要求一致有界, 仅假设由一个 n 的多项式控制. 作者在一套较弱的条件下运用消失折扣方法得到了 ϵ -最优平稳策略和最优平稳策略存在性.

1975 年, Lippman^[35] 进一步推广了他的前期工作, 考虑一般状态空间平均最优问题. 作者放弃了文献 [33] 中报酬函数由一个 n 的多项式控制的条件, 代替之的是权函数条件, 即报酬函数可以由一个权函数控制, 转移概率的限制也代替为由这个权函数控制. 在这些更宽松的条件下, 作者也建立了平均最优方程和最优策略的存在性. 不过, 这些结果是在状态空间 E 可数的条件下得到的.

1978 年, Federgruen 和 Tijms^[55] 讨论了可数状态空间 SMDP 的平均费用准则, 其模型主要特征是, (i) 瞬时费用 $c(x, a)$ 一致有界, (ii) 平均逗留时间 $\tau(x, a)$ 一致大于零且有界; (iii) 允许行动空间为紧集, (iv) $c(x, a)$ 、 $\tau(x, a)$ 和 $p(y | x, a)$ 关于 a 连续. 在一系列不同于 Ross 和 Lippman 的常返性条件下, 作者利用数据转化方法 (data transformation) 将平均 SMDP 转化为平均 DTMDP, 然后利用平均 DTMDP 的消失折扣方法建立了最优方程有界解的存在性, 他们还讨论并证明了策略迭代和值迭代算法及其收敛性.

1979 年, Federgruen 等人^[56] 在文献 [55] 基础上拓展到无界费用的情形. 不过, 他们的条件和方法不同于 Lippman^[35] 处理无界报酬平均 SMDP 的条件和方法, 他们假设 SMDP 的嵌入链只有一个

遍历集合, 在一些有限性的条件 (如存在一个有限状态集使得从任意状态出发首次返回这个状态集的期望时间和绝对期望费用有限) 下, 通过分析 SMDP 的嵌入链并运用正动态规划 (positive dynamic programming) 等方法, 证明了平均最优方程有限解的存在性, 以及平均费用最优平稳策略的存在性.

1983 年, Federgruen 等人^[57] 将文献 [56] 中单链的条件放松到一个较弱的常返性条件, 但费用函数加强为有上界而可无下界, 他们的方法与通常的消失折扣因子法不同, 他们将最优方程转化为可以利用 Tychonoff 不动点定理的形式, 从而运用 Tychonoff 不动点定理证明转化后模型最优方程解的存在性, 由此得到原模型平均最优方程解的存在性以及平均最优平稳策略的存在性.

1999 年, Luque-Vasquez 和 Hernández-Lerma^[58] 考虑了 Borel 状态和行动空间, 无界费用率和瞬时费用的平均 SMDP, 他们在文献 [35] 中的条件基础上, 进一步引入了适当的关于一般空间 Markov 过程的不可约性、常返性和几何遍历性条件, 利用数据转化方法 (data transformation) 将平均 SMDP 转化为平均 DTMDP, 分析了平均最优方程解的存在性和强平均最优策略的存在性, 实质地推广了文献 [35] 中关于可数状态条件下平均最优的结果.

2000 年, Vega-Amaya 和 Luque-Vasquez^[59] 处理了 SMDP 的样本轨道平均费用最优性 (sample path average cost optimality) 问题, 其中状态和行动空间都是 Borel 空间, 费用函数无界, 且平均逗留时间也无界 (放弃了通常的平均逗留时间有界的假设), 在适当的关于费用函数和平均逗留时间的增长条件 (growth conditions) 以及嵌入 Markov 链的稳定性条件下, 证明了三种样本轨道平均准则的等价性, 以及关于这三种样本轨道平均准则最优平稳策略的存在性.

2004 年, Luque-Vasquez 和 Vega-Amaya^[60] 及 Jaskiewicz^[61] 致力于讨论 Borel 状态和行动空间、无界支付函数的 SMDP 两种平均准则定义的等价性. Luque-Vasquez 和 Vega-Amaya^[60] 考虑了平均逗留时间无界的情形; Jaskiewicz^[61] 考虑了平均逗留时间一致有界的情形.

2004 年, Liu 和 Zhao^[62] 研究了有限状态和行动的 SMDP 平均准则, 将 Federgruen 等人^[56] 关于单链的结果推广到多链的情形, 分析方法是数据转化技巧, 在适当的条件下证明了转化后的 DTMDP 的平均最优策略也是原 SMDP 的平均最优策略.

2007 年, Jaskiewicz^[63] 讨论了一般 Borel 空间和 Feller (亦即弱连续的) 转移概率的 SMDP 平均准则, 基于不动点方法, 给出了平均最优方程存在下半连续解和连续解的充分条件.

2010 年, Luque-Vasquez 等人^[64] 考虑了状态和行动空间为 Borel 空间、费用/报酬函数无界和逗留时间 H 未知的一类平均准则 SMDP. 在假定 H 不依赖于状态和行动的条件下, 作者联合运用期望逗留时间的统计估计方法和平均 SMDP 的结果, 构造了一个平均费用最优 Markov 策略.

2012 年, Wei 和 Guo^[65] 给出了一般状态和行动空间 SMDP 平均最优的一些新条件, 提出了用“折扣最优值函数相对差可由一个 Lyapunov 型函数控制”的新条件代替已有的 w -几何遍历条件, 并用平均最优双不等式的新技术证明了最优平稳策略的存在性.

注 4.1 平均最优策略的特征及其计算的复杂性等还需要探讨.

5 有限阶段期望报酬准则

设 $T < \infty$ 是一个固定阶段数, SMDP 的有限阶段准则有以下两种形式定义:

$$V_1^\pi(t, x) := E_x^\pi \left[\sum_{n=0}^{N(t)} R(J_n, A_n) \right], \quad x \in E, \quad \pi \in \Pi, \quad 0 \leq t \leq T, \quad (5.1)$$

$$V_2^\pi(t, x) := E_x^\pi \left[\int_t^T r(Z(s), W(s)) ds \right], \quad x \in E, \quad \pi \in \Pi, \quad 0 \leq t \leq T, \quad (5.2)$$

其中, $N(t) := \sup\{n \geq 0 \mid S_n \leq t\}$, 即 $N(t)$ 是 $[0, t]$ 时间内的跳跃 (决策) 次数, $R(x, a)$ 是在状态 x 作出决策时的即时报酬函数, $r(x, a)$ 是模型 (2.1) 中的报酬率函数.

显然, 准则 $V_1^\pi(t, x)$ 表示系统从状态 x 出发使用策略 π , 在 $[0, t]$ 时间内的所有跳跃时刻点上产生的即时报酬之和 (不考虑两决策时刻之间的情形). 准则 $V_2^\pi(t, x)$ 表示系统从状态 x 出发使用策略 π , 在 $[t, T]$ 时间内产生的累积报酬之和 (考虑两决策时刻之间的情形).

即时报酬函数 $R(x, a)$ 和报酬率函数 $r(x, a)$ 是两种不同的报酬结构, 许多文献经常一起考虑这两种报酬, 如文献 [7, 66]; 也有很多文献仅考虑单一的报酬结构, 如文献 [34, 57] 只考虑了即时报酬 (费用) 函数 $R(i, a)$, 而文献 [31, 32, 67–69] 只考虑了报酬率 (费用率) 函数 $r(x, a)$. 实际上, 采用何种报酬结构本质上没有差别, 只有应用意义的不同, 我们可以根据实际问题需要选择合适的报酬结构.

众所周知, 有限阶段准则是 MDP 最基本的准则, 经过几十年的发展, DTMDP 的有限阶段准则理论已完善并有诸多应用 (参见专著 [7, 9, 13, 70] 等), 但关于 SMDP 有限阶段期望准则的研究却较少, 参见文献 [32, 34, 71, 72] 等. 与离散时间模型相比, 连续时间模型的有限阶段优化问题更加复杂且困难. 一方面, 不同于 DTMDP 的情形, 连续时间模型在有限阶段内的跳跃次数是随机的; 另一方面, 与连续时间模型的无限阶段折扣准则和长期平均准则不同, 连续时间模型的有限阶段最优策略不仅依赖于系统状态, 也依赖于当前时间, 增加了时间维度.

1986 年, Mamer^[34] 研究了第一种有限阶段期望准则 $V_1^\pi(t, x)$. 作者仅考虑了确定性 Markov 策略类、一般状态和行动空间以及无界即时报酬函数, 在正则性条件和文献 [35] 中涉及权函数 w 的假设框架内, 运用逐步逼近 (successive approximation) 方法建立了最优方程和最优策略的特征, 证明了值函数是下列有限阶段最优方程的唯一解:

$$V_1^*(t, x) = \sup_{a \in A(x)} \left[R(x, a) + \int_E \int_0^t V_1^*(t-s, y) F(ds \mid x, a, y) p(dy \mid x, a) \right],$$

以及达到上述方程右边最大值的策略 $f(t, x)$ 是最优的, 指出有限阶段 SMDP 的最优策略是非平稳的, 依赖于时间和状态, 导出了逐步逼近的值迭代算法, 最后通过分析最优资产清算 (liquidation of an asset) 的例子说明了他的方法.

2011 年, Huang 和 Guo^[32] 研究了第二种有限阶段期望准则 $V_2^\pi(t, x)$. 他们考虑了可数状态、有限行动、非负费用率 $c(x, a)$ 和策略依赖历史的 SMDP, 从直观出发引入依赖于剩余阶段数的策略, 在正则性条件而不需要文献 [34] 中涉及权函数 w 的假设下, 利用最小非负解方法证明了值函数是下列最优方程的最小非负解:

$$V_2^*(t, x) = \inf_{a \in A(x)} \left[c(x, a) \int_t^T (1 - Q(s, E \mid x, a)) ds + \sum_{y \in E} \int_t^T V_2^*(s, y) Q(ds, y \mid x, a) \right],$$

由此导出了最优策略的存在性、性质和计算方法, 并将结果应用到机器维护 - 维修问题有限阶段最优策略的数值计算.

2013 年, Ghosh 和 Saha^[72] 将通常的 SMDP 推广到了非平稳且在跳跃点之间也受控情形, 讨论了此类 SMDP 的有限阶段期望准则, 并在正则性假设和费用有界的条件下, 通过构造等价的 DTMDP, 建立了最优逐段开环松弛控制 (piecewise open loop relaxed control) 的存在性.

注 5.1 有限阶段最优策略的特征及其计算的复杂性等需要探讨, 而报酬 (费用) 函数等依赖时间的非平稳 SMDP 有限阶段模型还未涉及.

6 首达目标期望报酬准则

无限阶段期望折扣准则和平均期望准则关注的是系统“无限时间长期”的一种行为特征, 有限阶段期望准则关注的是系统“有限时间内”的性能特征. 与此同时, 在许多实际问题中, 由问题本身的特性或决策者的偏好, 一些人更关心系统首次到达某些状态集合前这段随机时间内的行为特性, 例如, 对可修设备系统, 往往关心设备发生故障前所产生的效益或故障修复前所花费的费用; 对通信网络, 更关注的是出现延迟、堵塞和信号损失等的时间, 总希望把它们控制得尽可能迟; 再如电子元件首次失效时间和排队系统顾客数首次超过容量的时间, 在实际中都受到许多关注, 并尽可能避免这些时间出现, 等等.

首达目标准则即是一种关注系统在首次到达目标状态集 B 前的某种性能指标的准则. 在实际问题中, 目标状态集 B 通常相应于设备系统的故障状态或修复状态, 通信网络的延迟、堵塞、信号损失状态、电子元件的失效状态和排队系统的容量等, 而首达目标准则的优化指标通常是首达目标状态集的期望时间和期望总报酬等. 具体地, 对目标集 $B \in \mathcal{B}(E)$, 首先引入过程首达 B 的时间,

$$\tau_B := \inf\{t \geq 0 \mid Z(t) \in B\}, \quad \inf \emptyset := \infty.$$

首达目标期望时间准则定义为

$$V_B^\pi(x) := E_x^\pi[\tau_B], \quad x \in E/B, \quad \pi \in \Pi.$$

首达目标期望折扣报酬准则定义为

$$V_B^\pi(x) := E_x^\pi \left[\int_0^{\tau_B} e^{-\alpha t} r(Z(t), A(t)) dt \right], \quad x \in E/B, \quad \pi \in \Pi. \quad (6.1)$$

显然, 这两个准则都具有清楚的实际意义, 例如, 当 B 代表系统“好”状态集时, 上述准则分别表示系统的期望“修复”时间和系统“修复”前的期望总费用; 当 B 代表系统“坏”状态集时, 上述准则分别表示系统的期望“工作”时间和系统“失效”前的期望总报酬. 值得注意的是, 当 $B = \emptyset$ 时, $\tau_B = +\infty$, 从而首达目标折扣报酬退化为无限阶段折扣报酬, 在这样一个意义下, 首达目标准则可以看作是无限阶段折扣准则的一种推广.

自从 1970 年 Derman^[73] 首次提出首达目标模型以后, 近几十年来已有很多文献涉及, 它们主要以首达目标期望总报酬 (一阶矩) 为优化目标. 林元列^[74] 研究了优化指标更为精确的 k -阶矩最优和 L 最优模型. Liu 等人^[75] 提出了首达目标期望报酬和首达时间依分布最优模型. 他们规定目标状态为失效状态, 考虑的是寻找策略使系统状态首次达到失效集的时间超过某给定值的概率达到最大, 给出了首达目标依分布最优策略的基本性质、结构和算法. 相关的文献还有文献 [76–80] 等. 不过, 上述文献都是考虑 DTMDP 或 CTMDP 模型的首达目标准则.

关于 SMDP 首达目标准则的工作, Huang 和 Guo^[68] 讨论了 SMDP 首达目标期望折扣费用准则, 他们假设状态和行动可数、费用率函数 $c(x, a)$ 非负、正则性条件成立, 利用最小非负解方法, 证明了在 “ $\inf_{a \in A(x)} c(x, a) > 0, x \in E/B$ ” 条件下存在 ϵ -最优平稳策略, 在允许行动集 $A(x)$ 有限的条件下存在最优平稳策略, 并证明了值函数 V_B^* 是下列首达最优方程的最小非负解:

$$V_B^*(x) = \inf_{a \in A(x)} \left[c(x, a) \int_0^\infty e^{-\alpha s} (1 - Q(s, E \mid x, a)) ds + \sum_{y \in E/B} \int_0^\infty e^{-\alpha t} Q(dt, y \mid x, a) V_B^*(y) \right], \quad x \in E/B,$$

同时还给出了计算值函数和最优策略的值迭代算法.

另外, Huang 和 Guo^[67] 考虑了 SMDP 的首达目标概率准则, Huang 和 Guo^[69] 研究了受约束 SMDP 的首达目标期望准则, 详见下面概率准则和受约束问题部分.

注 6.1 多目标准则的 SMDP 平均、有限阶段和首达目标模型似乎未曾研究.

7 概率准则

在许多实际问题中, 人们关注的是系统的可靠性或失效的风险. 此时, 经典的期望 (报酬/费用) 准则不再适用. 在这样背景下, 一种新准则 — 概率准则被提出, 并得到越来越多的研究. 概率准则的一般形式为

$$F_1^\pi(x, \lambda) := P_{(x, \lambda)}^\pi(Y \leq \lambda); \quad F_2^\pi(x, \lambda) := P_{(x, \lambda)}^\pi(Y \geq \lambda), \quad x \in E, \quad \lambda \in \mathbb{R}, \quad (7.1)$$

其中, Y 是某一随机变量指标, λ 是任意的水平. 显然, 使得随机报酬期望值相同的两个策略不一定使得其概率值一致, 故概率准则是比期望准则更高级的准则, 它能更加全面地刻画策略的性能, 更加灵敏地将期望相同的策略进一步区别开来. 在实际中, 前者 F_1^π 通常刻画的是风险, 后者 F_2^π 通常刻画的是可靠性. 这里主要以第一种形式的概率准则说明其意义.

当 Y 取首达时间, 即 $Y = \tau_B$ 时, $F_1^\pi(x, \lambda)$ 表示系统首次到达目标状态集 B 的时间不超过 λ 的概率. 在应用中, 目标集 B 通常代表系统的失效状态集, 这样, $F_1^\pi(x, \lambda)$ 就表示系统在 $[0, \lambda]$ 时间区间内发生失效的概率, 它是系统失效的风险度量, 自然, 我们要最小化这个概率. 当然, 若目标集 B 代表系统“好”的状态集, $F_1^\pi(x, \lambda)$ 就表示系统在 $[0, \lambda]$ 时间区间内被修复的概率, 自然, 我们要最大化这个概率.

当 Y 分别取无限阶段折扣总报酬、首达目标总报酬或有限阶段总报酬时, 即

$$Y = \int_0^\infty e^{-\alpha s} r(Z(s), A(s)) ds, \quad \int_0^{\tau_B} e^{-\alpha s} r(Z(s), A(s)) ds \quad \text{或} \quad \int_0^T r(Z(s), A(s)) ds, \quad (7.2)$$

则 $F_1^\pi(x, \lambda)$ 表示系统无限阶段折扣总报酬、首达目标集 B 前 (如失效前) 产生的总报酬或有限时间 $[t, T]$ 内产生的总报酬不超过水平 λ 的概率, 自然, 我们要最小化这些概率. 当然, 若考虑“费用”情形, 则应最大化这些概率.

由上可见, 概率准则有许多实际背景, 因而得到了许多学者的研究, 如文献 [79–85] 等. 不过这些研究都致力于 DTMDP 模型. 关于 MDP 概率准则的综述, 可参见文献 [86].

2009 年, Huang 和 Guo^[31] 研究了 SMDP 首次到达目标状态集 B 的时间不超过给定值 λ 的风险概率, 将 Liu 和 Huang^[79] 关于 DTMDP 模型的结果推广到 SMDP 的情形. 他们假设状态空间 E 可数, 行动集 $A(x)$ 有限, 报酬率函数非负, 在正则性条件下, 证明了值函数 $F_1^*(x, \lambda) := \inf_\pi F_1^\pi(x, \lambda)$ (其中 $Y = \int_0^{\tau_B} r(Z(s), A(s)) ds$) 是下列首达概率最优方程的最大解:

$$F^*(x, \lambda) = \inf_{a \in A(x)} \left[Q(\lambda, B \mid x, a) + \sum_{y \in B^c} \int_0^\lambda Q(dt, y \mid x, a) F^*(y, \lambda - t) \right], \quad x \in B^c, \quad \lambda \in \mathbb{R}_+,$$

讨论了最优策略的性质、最优方程解的唯一性及最优策略的存在性, 导出了计算最优值函数和最优策略的值迭代、策略改进算法, 给出了存在与阈值 λ 无关的最优策略的充要条件.

2010 年, Sakaguchi 和 Ohtsubo^[87] 考虑了无折扣首达目标 SMDP 的阈值概率最小化问题, 即要使首次到达目标状态集 B 前的累积总报酬不超过某给定值 λ 的概率达到最小. 其模型状态空间可数,

允许行动集 $A(x)$ 有限, 报酬函数 $r(x, a, t)$ 可正可负但有界且非平稳. 在正则性条件、 B 是常返类和无报酬 (reward-free) 的假设下, 他们将该问题转化为带常返类的无限阶段的情形, 然后证明了最优值函数是下列最优方程的唯一解:

$$F^*(x, \lambda) = \inf_{a \in A(x)} \left[\sum_{y \in E} \int_0^\infty Q(dt, y | x, a) F^* \left(y, \lambda - \int_0^t r(x, a, s) ds \right) \right],$$

并存在一个最优平稳策略, 给出了若干值迭代和策略改进算法. 作者还进一步研究了累积总报酬的阈值概率与期望的关系.

2011 年, Huang 和 Guo^[67] 也考虑了 SMDP 首达目标报酬的概率准则. 他们假设状态空间 E 和行动集 A 均为 Borel 空间, 在正则性条件和连续 - 紧条件下, 证明了值函数 $F_1^*(x, \lambda)$ 是下列概率准则最优方程的最大解:

$$F^*(x, \lambda) = \inf_{a \in A(x)} \left[Q(\lambda/r(x, a), B | x, a) + \int_{B^c} \int_0^{\lambda/r(x, a)} Q(dt, y | x, a) F^*(y, \lambda - r(x, a)t) \right],$$

$$x \in B^c, \quad \lambda \in \mathbb{R}_+,$$

进一步在“存在一个常数 $\gamma > 0$, 使得对 $x \in B^c$ 和 $a \in A(x)$, $p(B | x, a) \geq \gamma$ ”条件下, 证明了值函数是上述方程的唯一解, 给出了最优平稳策略的存在性、迭代算法及其数值算例.

2013 年, Huang 和 Guo^[88] 将文献 [83–85] 中关于 DTMDP 有限阶段报酬的概率准则推广到了 SMDP 的情形, 即优化问题是要最小化 SMDP 有限阶段报酬不超过某给定值 λ 的概率 $F^\pi(t, x, \lambda) := P_{(x, \lambda)}^\pi(\int_0^t r(Z(s), A(s)) ds \leq \lambda)$. 他们考虑了非负无上界的报酬函数、可数状态空间 E 和有限行动集 $A(x)$, 仅在正则性条件下证明了值函数 $F^*(t, x, \lambda) := \inf_{\pi \in \Pi} F^\pi(t, x, \lambda)$ 是下列有限阶段概率准则最优方程的唯一解,

$$F^*(t, x, \lambda) = \inf_{a \in A(x)} \left[\mathbb{1}_{[0, \lambda]}(r(x, a)t)(1 - Q(t, E | x, a)) + \sum_{y \in E} \int_0^t Q(du, y | x, a) F^*(t - u, y, \lambda - r(x, a)u) \right],$$

$$t \in [0, T], \quad x \in E, \quad \lambda \in \mathbb{R}_+,$$

进而建立了最优策略的存在性以及迭代算法.

注 7.1 目前还未发现关于概率准则最优策略结构和特征的讨论, 也未发现基于“受控 Itô 过程”概率准则的文献.

8 受约问题

受约束问题产生于实际中资源的有限性和多样性, 一般形式是在使得期望费用不超过某上限值的策略中, 去寻找获得最大期望报酬的策略. 受约束的 MDP 应用非常广泛, 例如, 在排队系统控制中, 保证其他类顾客一定输出量的约束下, 最大化某类顾客的输出量问题; 在库存控制中, 保证供给满足一个概率上界的最小费用问题, 以及很多经济问题都能用受约束的 MDP 来刻画和描述.

早在 1970 年, Derman^[73] 就已经提出了有约束条件的 MDP, 随后, 受约束 MDP 引起了很多学者的研究兴趣, 出现了大量文献, 例如, Chen 和 Feinberg^[89] 考虑了 DTMDP 受约束的折扣准则; Altman^[90, 91] 研究了 DTMDP 受约束的总费用准则; 文献 [92, 93] 对 DTMDP 受约束的平均准则进行

了讨论; 文献 [20, 94] 处理了 CTMDP 受约束的折扣准则; Guo 等人^[21] 考虑了 CTMDP 受约束的平均准则.

关于受约束 SMDP 的工作, Beutler 和 Ross^[95] 探讨了受约束长期平均准则, 他们假设状态空间为有限空间, 行动空间是紧致度量空间, 运用 Lagrange 乘子方法将约束优化问题与无约束问题联系起来, 从而证明了半简单 (semi-simple) 最优约束策略的存在性, 即最优策略至多在一个状态随机采取两个行动, 而在其他所有状态都是确定性的选取行动. Feinberg^[96] 进一步对受约束 SMDP 的平均准则进行了讨论, 他们局限于有限状态和行动, 发展了一种状态 - 行动更新率 (类似占有测度) 技术, 导出了计算约束最优策略的线性规划算法.

1997 年, 胡光华和张升^[97] 研究了无限阶段折扣准则 SMDP 的单一约束最优问题, 他们假定状态集可数, 行动集为紧的非空 Borel 集, 报酬函数和费用函数一致有界, 通过引入 Lagrange 乘子, 把约束问题转化为无约束问题, 证明了在连续 - 紧条件和 Slater 条件下存在约束最优策略, 还给出了求解约束最优策略线性规划算法.

2002 年, Feinberg^[98] 讨论了一般状态和行动空间、报酬函数有上界的多约束折扣准则 SMDP, 作者将它转化为相应的多约束 DTMDP 问题来处理, 推广了 DTMDP 的一些结果到 SMDP 的情形, 得到了约束最优策略的存在性和求解方法. 2004 年, Feinberg^[66] 通过引入占有测度将一般状态和行动空间的 SMDP 无限阶段折扣准则受约束问题转化为一个线性规划, 并用同样的方法研究了有限次跳跃 SMDP 的受约束问题, 说明了其实际等价于一个有限阶段 DTMDP 的受约束问题, 也导出了计算约束最优策略的线性规划算法. 2013 年, Gonzalez-Hernandez 和 Villarreal-Rodriguez^[99] 讨论了文献 [66, 98] 中的类似问题, 不过他们考虑的是非负可测费用函数的情形.

2012 年, Huang 和 Guo^[69] 考虑了 SMDP 中的首达目标准则的单一约束问题, 即

$$\begin{aligned} \text{maximize } V_{\nu}(r, \pi) &:= \int_E E_x^{\pi} \left[\int_0^{\tau_B} e^{-\alpha t} r(Z(t), W(t)) dt \right] \nu(dx), \\ \text{s.t. } V_{\nu}(c, \pi) &:= \int_E E_x^{\pi} \left[\int_0^{\tau_B} e^{-\alpha t} c(Z(t), W(t)) dt \right] \nu(dx) \leq b, \end{aligned}$$

其中, ν 为初始分布, $V_{\nu}(r, \pi)$ 表示系统失效前产生的期望总报酬, 而 $V_{\nu}(c, \pi)$ 则表示系统失效前需要的期望总费用. 显然, 当 $B = \emptyset$ 时, $\tau_B = +\infty$, 从而, 首达目标准则约束问题退化为经典的无限阶段折扣准则受约束问题. 作者假设状态空间可数, 行动空间为 Borel 空间, 报酬/费用率可能既无上界也无下界, 运用 Lagrange 乘子法, 在连续 - 紧条件和 Slater 条件下证明了约束最优策略存在, 并说明约束最优策略至多在一个状态随机化.

2014 年, Huang 和 Guo^[100] 提出并研究了下列 SMDP 有限阶段准则的多约束问题:

$$\text{maximize } V_{\nu}(r, \pi) := \int_E E_x^{\pi} \left[\int_0^T r(Z(t), W(t)) dt \right] \nu(dx), \quad \text{在 } U \text{ 上,}$$

其中,

$$U := \left\{ \pi \in \Pi : \int_E E_x^{\pi} \left[\int_0^T c_k(Z(t), W(t)) dt \right] \nu(dx) \leq b_k, k = 1, \dots, N \right\},$$

这里 $c_k(\cdot, \cdot)$ 表示第 k 种费用率函数, b_k 是第 k 个约束常数. 目标是寻求约束最优策略 $\pi^* \in U$, 其满足 $V_{\nu}(r, \pi^*) = \sup_{\pi \in U} V_{\nu}(r, \pi)$. 作者通过定义新的含有时间维度的占有策略, 即

$$\eta_{\nu}^{\pi}(\Gamma, C, D) := \sum_{n=0}^{\infty} P_{\nu}^{\pi}(T_n \in \Gamma, X_n \in C, A_n \in D), \quad \Gamma \in \mathcal{B}([0, T]), \quad C \in \mathcal{B}(E), \quad D \in \mathcal{B}(A),$$

证明了历史依赖策略与随机 Markov 策略等价 (具有相同的有限阶段报酬/费用), 进而将有限阶段 SMDP 的多约束问题转化为一个关于占有测度的线性规划:

$$\begin{aligned} \text{LP:} \quad & \text{maximize} \quad \int_{[0,T] \times E \times A} \tilde{r}(t, x, a) \eta(dt, dx, da), \\ & \text{s.t.} \quad \hat{\eta}(B \times C) = \nu(B \times C) + \int_{[0,T] \times E \times A} \mathbb{Q}(B \times C \mid t, x, a) \eta(dt, dx, da), \\ & \quad \forall B \times C \in \mathcal{B}([0, T] \times E), \quad \int_{[0,T] \times E \times A} \tilde{c}_k(t, x, a) \eta(dt, dx, da) \leq d_k, \quad k = 1, \dots, N, \end{aligned}$$

其中, $\mathbb{Q}$, $\tilde{r}(t, x, a)$ 和 $\tilde{c}_k(t, x, a)$ 定义为

$$\begin{aligned} \mathbb{Q}(B \times C \mid t, x, a) &:= \int_0^{T-t} I_B(t+s) Q(ds, C \mid x, a), \quad B \in \mathcal{B}([0, T]), \quad C \in \mathcal{B}(E), \\ \tilde{r}(t, x, a) &:= r(x, a) \int_0^{T-t} (1 - Q(s, E \mid x, a)) ds, \\ \tilde{c}_k(t, x, a) &:= c_k(x, a) \int_0^{T-t} (1 - Q(s, E \mid x, a)) ds, \quad (t, x, a) \in [0, T] \times K, \quad k = 0, \dots, N. \end{aligned}$$

他们给出了保证约束最优策略存在的两套条件 (经典的连续 - 紧假设和一些涉及权范数的条件), 在这些条件下, 上述 LP 被证明是可解的, 从而得到最优策略的存在性, 并刻画了约束最优策略的特征和性质. 不过, 最优策略的数值计算 (特别是连续状态和行动的情形) 需要进一步研究.

注 8.1 受约束的概率准则有明显的实际意义, 但还无文献涉及.

9 均值 - 方差问题

均值 - 方差问题最早由美国经济学家 Markowitz 于 1959 年提出. 由于其兼顾了金融市场中的收益和风险两大要素, 自提出以来就得到了广泛关注和研究. 均值 - 方差问题的经典形式是在保证随机收益的期望不少于给定水平的条件下, 最小化某随机收益的方差, 即要最大化随机收益的稳定性. 其数学表述是

$$\text{minimize } \text{Var}^\pi[Y], \quad \text{s.t. } E^\pi[Y] \geq \lambda,$$

其中, Y 是金融市场中的某随机收益, λ 是决策者预定的收益水平, Var^π 和 E^π 分别是依赖策略 π 的方差和期望算子 (如 E_x^π , 不仅依赖策略 π , 还依赖初始状态 x). 1959 年, Markowitz^[101] 提出均值 - 方差问题时考虑的是单周期的静态的随机收益, 其解法是利用 Lagrange 乘子法求解一个二次规划. 2000 年, Li 和 Ng^[102] 将单周期的均值 - 方差问题推广到了多周期的动态随机收益. 后来, 均值 - 方差问题被引入到 MDP 中. 在 MDP 中, 常见的均值 - 方差问题是在期望折扣报酬或平均报酬最优策略中再寻找方差最小策略, 此类问题通常称为方差最小问题. 近年, 在期望收益为一定值 (不一定是最优值) 的策略中再寻找方差最小策略的问题 (即均值 - 方差问题) 也引起不少学者的兴趣和研究, 如文献 [103–105].

在目前 MDP 文献中, 大部分工作考虑的是 DTMDP 或 CTMDP 中方差最小问题或方差惩罚 (variance-penalized) 问题, 例如, Collins^[106] 处理了 DTMDP 有限阶段终端报酬的方差惩罚问题; Van Dijk 和 Sladký^[107] 建立了 CTMDP 有限阶段折扣报酬的方差公式及其计算方法; Filar 等人^[108] 处理了 DTMDP 无限阶段折扣报酬的方差惩罚问题; Guo 等人^[103] 处理 CTMDP 无限阶段折扣报酬的均

值 - 方差问题; 文献 [109, 110] 考虑了 CTMDP 极限平均准则的方差最小化问题; 文献 [7, 10, 111, 112] 处理了 DTMDP 的极限平均方差问题.

现有工作中关于 SMDP 均值 - 方差问题的相对较少. Sobel^[113] 考虑了 SMDP 无限阶段折扣报酬的方差问题, 他利用无限阶段折扣报酬的分布函数公式 (formula) 导出了无限阶段折扣报酬的方差及更高阶矩满足的方程, 但未讨论最优性问题. Wei 和 Guo^[105] 研究了 SMDP 无限阶段折扣报酬的方差最小问题, 其中的状态空间为一般的 Borel 空间, 报酬函数无界, 他们首先通过折扣最优方程刻画期望折扣最优策略, 接着将最优策略下的方差转化为一阶矩, 然后建立方差最优方程, 刻画方差最小策略.

Huang 和 Guo^[104] 考虑了如下 SMDP 有限阶段报酬的均值 - 方差问题:

$$\text{minimize } \text{Var}_x^f \left[\int_0^T r(Z(s), W(s)) ds \right], \quad f \in F_g,$$

其中,

$$F_g := \left\{ f \in F \mid E_x^f \left[\int_0^T r(Z(s), W(s)) ds \right] = g(x), x \in E \right\},$$

这里状态空间和行动空间都是 Borel 空间, 报酬率 r 可为无界函数, g 是给定的期望报酬. 他们致力于寻找均值 - 方差最优策略 $f^* \in F_g$, 即其满足

$$\text{Var}_x^{f^*} \left[\int_0^T r(Z(s), W(s)) ds \right] = \inf_{f \in F_g} \text{Var}_x^f \left[\int_0^T r(Z(s), W(s)) ds \right]$$

对任意 $x \in E$ 成立. 如前面指出, 有限阶段报酬的均值 - 方差问题会比无限阶段折扣报酬的均值 - 方差问题复杂和困难一些, 此时必须考虑含有时间维度形如 $f(t, x)$ 的策略. 在正则性条件和一些权范数有界的条件下, 作者首先刻画 F_g 中的策略, 即 $f \in F$ 是 F_g 中的策略当且仅当对每个 $(t, x) \in [0, T] \times E$, $f(t, x) \in A_g(t, x)$, 其中,

$$A_g(t, x) := \left\{ a \in A(x) \mid g(t, x) = r(t, x, a) + \int_E \int_0^{T-t} Q(ds, dy \mid x, a) g(t+s, y) \right\};$$

然后将方差 (二阶矩) 转化为一个新报酬函数的期望 (一阶矩), 即

$$\text{Var}_x^f \left[\int_0^T r(Z(s), W(s)) ds \right] = E_x^f \left[\sum_{k=0}^{\infty} c_g(T_k \wedge T, X_k, f) \right],$$

其中, 新报酬函数 $c_g(t, x, a)$ 定义为

$$c_g(t, x, a) := 2r^2(x, a) \int_0^{T-t} s(1 - Q(s, E \mid x, a)) ds + 2r(x, a) \int_E \int_0^{T-t} sg(t+s, y) Q(ds, dy \mid x, a),$$

$$(t, x, a) \in [0, T] \times K,$$

由此利用已知的关于 DTMDP 无限阶段期望报酬准则的结果解决原问题.

注 9.1 上述文献大都考虑的是期望收益等于某一水平的均值 - 方差问题, 其处理方法似乎不适合分析期望收益大于等于某一水平的均值 - 方差问题. 而当约束策略集 F_g 取为

$$\left\{ f \in F \mid E_x^f \left[\int_0^T r(Z(s), W(s)) ds \right] \geq g(x), x \in E \right\}$$

的情形更有意义, 此时的均值 - 方差问题还无人探讨.

10 展望

SMDP 是一类广泛而又重要的随机动态系统, 目前研究活跃, 取得了丰富的成果, SMDP 未来的潜在研究内容, 除了在注 3.1–9.1 提及的之外, 还有下列两个方面:

一方面, 模型向一般化发展, 考虑连续受控 SMDP. 作为一类连续时间模型, 现有文献研究的 SMDP 模型大都是在跳跃点受控, 能否考虑像 CTMDP 那样连续受控的 SMDP 呢? 答案是肯定的. 早在 1977 年, Yushkevich^[114, 115] 就已将 CTMDP 模型推广到了连续受控 SMDP, 为与通常 SMDP 相区别, 研究者一般称 Yushkevich 推广的 SMDP 为 YSMDP. 在 YSMDP 中, 在跳跃点选取的不是一个通常行动而是一个依赖于时间的行动函数, 这为处理 YSMDP 增加了许多困难. 目前, 关于 YSMDP 的工作很少且具挑战性, 需要发展.

另一方面, SMDP 新的优化问题的研究. 伴随着实际情形的不断变化, 新的优化准则和问题必将不断提出, 这也为推动 SMDP 的发展注入了动力. 例如, 经济金融中具有重要意义的期望效用准则、风险价值 (VaR) 准则和平均风险价值 (AVaR) 准则等都可以引入到 SMDP 模型中, 是 SMDP 理论有待研究的课题之一.

另外, 通过引入决策变量和优化目标 (如折扣准则、平均准则、有限阶段准则、首达目标准则和概率准则等), 研究所谓的 Markov 骨架决策过程, 即将 MDP 的思想和主要结果拓展到 Markov 骨架过程的情形, 研究 Markov 骨架过程的最优控制问题. 此类优化问题有丰富的实际背景和大量的研究内容, 值得探索.

参考文献

- 1 Shapley L S. Stochastic games. *Proc Natl Acad Sci USA*, 1953, 39: 1095–1100
- 2 Howard R A. *Dynamic Programming and Markov Processes*. Cambridge: MIT Press, 1960
- 3 Dynkin E B, Yushkevich A A. *Controlled Markov Processes*. New York: Springer-Verlag, 1979
- 4 Gihman I I, Skorohod A V. *Controlled Stochastic Processes*. New York/Heidelberg: Springer-Verlag, 1979
- 5 Jewell W S. Markov renewal programming I: Formulation, finite return models; Markov renewal programming II: Infinite return models, example. *Oper Res*, 1963, 11: 938–971
- 6 Altman E. *Constrained Markov Decision Processes*. Boca Raton, FL: Chapman and Hall/CRC, 1999
- 7 Puterman M L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New York: John Wiley & Sons Inc, 1994
- 8 Sennott L I. *Stochastic Dynamic Programming and the Control of Queueing Systems*. New York: Wiley, 1999
- 9 Hernández-Lerma O, Lasserre J B. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. New York: Springer-Verlag, 1996
- 10 Hernández-Lerma O, Lasserre J B. *Further Topics on Discrete-Time Markov Control Processes*. New York: Springer-Verlag, 1999
- 11 Piunovskiy A. *Examples in Markov Decision Processes*. London: Imperial College Press, 2012
- 12 Feinberg E A, Shwartz A. *Handbook of Markov Decision Processes*. Boston/Dordrecht/London: Kluwer Academic Publishers, 2002
- 13 Bauerle N, Rieder U. *Markov Decision Processes with Applications to Finance*. Heidelberg: Springer, 2011
- 14 Hu Q Y, Yue W Y. *Markov Decision Processes with Their Applications*. New York: Springer, 2008
- 15 Sigaud O, Buffet O. *Markov Decision Processes in Artificial Intelligence*. London: Wiley-IST, 2010
- 16 Mahadevan S. *Learning Representation and Control in Markov Decision Processes*. Hanover: Now Publishers Inc, 2009
- 17 Simari G I, Parsons S D. *Markov Decision Processes and the Belief-Desire-Intention Model*. New York: Springer-Verlag, 2011
- 18 Guo X P, Hernández-Lerma O. *Continuous-Time Markov Decision Processes: Theory and Applications*. Berlin: Springer-Verlag, 2009

- 19 Prieto-Rumeau T, Hernandez-Lerma O. Selected Topics in Continuous-Time Controlled Markov Chains and Markov Games. London: Imperial College Press, 2012
- 20 Guo X P, Song X Y. Discounted continuous-time constrained Markov decision processes in Polish spaces. *Ann Appl Probab*, 2011, 21: 2016–2049
- 21 Guo X P, Huang Y H, Song X Y. Linear programming and constrained average optimality for general continuous-time Markov decision processes in history-dependent policies. *SIAM J Control Optim*, 2012, 50: 23–47
- 22 Arapostathis A, Borker V S, Fernández-Gaucherand E, et al. Discrete-time controlled Markov processes with average cost criterion: A survey. *SIAM J Control Optim*, 1993, 31: 282–344
- 23 Guo X P, Hernández-Lerma O, Prieto-Rumeau T. A survey of recent results on continuous-time Markov decision processes. *TOP*, 2006, 14: 177–246
- 24 董泽清. 马尔可夫决策规划的现状和展望. *运筹学杂志*, 1987, 6: 1–17
- 25 Howard R A. Semi-Markovian decision processes. *Bull Inst Internat Statist*, 1963, 40: 625–652
- 26 Fox B L. Markov renewal programming by linear fractional programming. *SIAM J Appl Math*, 1966, 14: 1418–1432
- 27 Fox B L. (g, w) -optimal in Markov renewal programmes. *Management Sci*, 1968, 15: 210–212
- 28 Denardo E V, Fox B. Multichain Markov renewal programmes. *SIAM J Appl Math*, 1968, 16: 468–487
- 29 Schweitzer P J. Iterative solution of the functional equations of undiscounted Markov renewal programming. *J Math Anal Appl*, 1971, 34: 495–501
- 30 Ross S M. Average cost semi-Markov decision processes. *J Appl Probab*, 1970, 7: 649–656
- 31 Huang Y H, Guo X P. Optimal risk probability for first passage models in semi-Markov decision processes. *J Math Anal Appl*, 2009, 359: 404–420
- 32 Huang Y H, Guo X P. Finite horizon semi-Markov decision processes with application to maintenance systems. *European J Oper Res*, 2011, 212: 131–140
- 33 Lippman S A. Semi-Markov decision processes with unbounded rewards. *Management Sci*, 1973, 19: 717–731
- 34 Mamer J W. Successive approximations for finite horizon semi-Markov decision processes with application to asset liquidation. *Oper Res*, 1986, 34: 638–644
- 35 Lippman S A. On dynamic programming with unbounded rewards. *Management Sci*, 1975, 21: 1225–1233
- 36 Wakuta K. Arbitrary state semi-Markov decision processes with unbounded rewards. *Optimization*, 1987, 18: 447–454
- 37 Wessels J, van Nunen J A E E. Discounted semi-Markov decision processes: Linear programming and policy iteration. *Statist Neerlandica*, 1975, 29: 1–7
- 38 Feldman R M. Optimal replacement with semi-Markov shock models using discounted costs. *Math Oper Res*, 1977, 2: 78–90
- 39 Sadjadi D, Bestwick P F. A stagewise action elimination algorithm for the discounted semi-Markov problem. *J Oper Res Soc*, 1979, 30: 633–637
- 40 Porteus E. Improved iterative computation of the expected discounted return in Markov and semi-Markov chains. *Z Oper Res*, 1980, 24: 155–170
- 41 Porteus E. Computing the discounted return in Markov and semi-Markov chains. *Naval Res Logist Quart*, 1981, 28: 567–577
- 42 Schweitzer P J, Puterman M L, Kindle K W. Iterative aggregation-disaggregation procedures for discounted semi-Markov reward processes. *Oper Res*, 1985, 33: 589–605
- 43 Hernández-Lerma O. Nonstationary value-iteration and adaptive control of discounted semi-Markov processes. *J Math Anal Appl*, 1985, 112: 435–445
- 44 董泽清, 刘克. 无界报酬折扣半马氏模型最优策略的结构. *中国科学 A 辑*, 1987, 11: 975–985
- 45 Bhattacharya R N, Majumdar M. Controlled semi-Markov models-the discounted case. *J Statist Plann Inference*, 1989, 21: 365–381
- 46 胡奇英. 随机终止的非平稳折扣半马氏决策规划. *应用数学学报*, 1993, 16: 566–570
- 47 Hu Q Y. Discounted semi-Markov decision process in a semi-Markov environment. *Optimization*, 1997, 39: 367–382
- 48 Hudak D, Nollau V. Approximation solution and suboptimality for discounted semi-Markov decision problems with countable state space. *Optimization*, 2004, 53: 339–353
- 49 Luque-Vasquez F, Minjarez-Sosa J A. Semi-Markov control processes with unknown holding times distribution under a discounted criterion. *Math Methods Oper Res*, 2005, 61: 455–468
- 50 Luque-Vasquez F, Minjarez-Sosa J A, Rosas-Rosas L C. Semi-Markov control models with partially known holding times distribution: Discounted and average criteria. *Acta Appl Math*, 2011, 114: 135–156
- 51 Yin B Q, Li Y J, Zhou Y P, et al. Performance optimization of semi-Markov decision processes with discounted-cost criteria. *Eur J Control*, 2008, 14: 213–222

- 52 黄永辉, 郭先平. 非负费用折扣半马氏决策过程. 数学学报, 2010, 53: 503–514
- 53 Rosas-Rosas L C. Estimation of the optimality deviation in discounted semi-Markov control models. In: Optimization, Control, and Applications of Stochastic Systems, Systems Control Found Appl. New York: Birkhäuser/Springer, 2012, 253–264
- 54 Lippman S A. Maximal average-reward policies for semi-Markov decision processes with arbitrary state and action space. Ann Math Statist, 1971, 42: 1717–1726
- 55 Federgruen A, Tijms H C. The optimality equation in average cost denumerable state semi-Markov decision problems: Recurrency conditions and algorithms. J Appl Probab, 1978, 15: 356–373
- 56 Federgruen A, Hordijk A, Tijms H C. Denumerable state semi-Markov decision processes with unbounded costs, average cost criterion. Stochastic Process Appl, 1979, 9: 223–235
- 57 Federgruen A, Schweitzer P J, Tijms H C. Denumerable undiscounted semi-Markov decision processes with unbounded rewards. Math Oper Res, 1983, 8: 298–313
- 58 Luque-Vasquez F, Hernández-Lerma O. Semi-Markov control models with average costs. Appl Math, 1999, 26: 315–331
- 59 Vega-Amaya O, Luque-Vasquez F. Sample-path average cost optimality for semi-Markov control processes on Borel spaces: Unbounded costs and mean holding times. Appl Math, 2000, 27: 343–367
- 60 Luque-Vasquez F, Vega-Amaya O. Time and ratio expected average cost optimality for semi-Markov control processes on Borel spaces. Semi-Markov processes: Theory and applications. Comm Statist Theory Methods, 2004, 33: 715–734
- 61 Jaskiewicz A. On the equivalence of two expected average cost criteria for semi-Markov control processes. Math Oper Res, 2004, 29: 326–338
- 62 Liu J Y, Zhao X B. On average reward semi-Markov decision processes with a general multichain structure. Math Oper Res, 2004, 29: 339–352
- 63 Jaskiewicz A. A fixed point approach to solve the average cost optimality equation for semi-Markov decision processes with Feller transition probabilities. Comm Statist Theory Methods, 2007, 36: 2559–2575
- 64 Luque-Vasquez F, Minjarez-Sosa J A, Rosas-Rosas L C. Semi-Markov control processes with unknown holding times distribution under an average cost criterion. Appl Math Optim, 2010, 61: 317–336
- 65 Wei Q D, Guo X P. New average optimality conditions for semi-Markov decision processes in Borel spaces. J Optim Theory Appl, 2012, 153: 709–732
- 66 Feinberg E A. Continuous time discounted jump Markov decision processes: a discrete-event approach. Math Oper Res, 2004, 29: 492–524
- 67 Huang Y H, Guo X P, Song X Y. Performance analysis for controlled semi-Markov systems with application to maintenance. J Optim Theory Appl, 2011, 150: 395–415
- 68 Huang Y H, Guo X P. First passage models for denumerable semi-Markov decision processes with nonnegative discounted costs. Acta Math Appl Sin, 2011, 27: 177–190
- 69 Huang Y H, Guo X P. Constrained optimality for first passage criteria in semi-Markov decision processes. In: Optimization, Control, and Applications of Stochastic Systems, Systems Control Found Appl. New York: Birkhauser/Springer, 2012, 181–202
- 70 Hinderer K. Foundations of Non-Stationary Dynamic Programming with Discrete Time Parameter. Lecture Notes in Operations Research. New York: Springer-Verlag, 1970
- 71 Ghosh M K, Saha S. Continuous-time controlled jump Markov processes on the finite horizon. In: Optimization, Control, and Applications of Stochastic Systems, Systems & Control: Foundations & Applications. New York: Birkhäuser/Springer, 2012, 99–109
- 72 Ghosh M K, Saha S. Non-stationary semi-Markov decision processes on a finite horizon. Stoch Anal Appl, 2013, 31: 183–190
- 73 Derman C. Finite State Markov Decision Processes. New York: Academic Press, 1970
- 74 林元列. 连续时间首达目标模型 (I) — 折扣矩最优模型. 应用数学学报, 1991, 14: 115–124
- 75 Lin Y L, Wu C B, Kang B D. Optimal models with maximizing the probability of first achieving target value in preceding stages. Sci China Ser A, 2003, 46: 396–414
- 76 Lin Y L, Tomkins R J, Wang C L. Optimal models for the first arrival time distribution function in continuous time-with a special case. Acta Math Appl Sin, 1994, 10: 194–212
- 77 Liu J Y, Liu K. Markov decision programming — the first passage model with denumerable state space. Systems Sci Math Sci, 1992, 5: 340–351
- 78 Liu J Y, Liu K. Markov decision programming — the moment optimal problem for the first-passage model. J Aust Math Soc Ser B, 1997, 38: 542–562
- 79 Liu J Y, Huang S M. Markov decision processes with distribution function criterion of first-passage time. Appl Math

- Optim, 2001, 43: 187–201
- 80 Ohtsubo Y. Optimal threshold probability in undiscounted Markov decision processes with a target set. *Appl Math Anal Comp*, 2004, 149: 519–532
 - 81 Bouakiz M. *The Target Level Criterion in Markov Decision Processes*. Florence, SC: Francis Marion College, 1986
 - 82 White D J. Minimizing a threshold probability in discounted Markov decision processes. *J Math Anal Appl*, 1991, 173: 634–646
 - 83 Bouakiz M, Kebir Y. Target-level criterion in Markov decision processes. *J Optim Theory Appl*, 1995, 88: 1–15
 - 84 Wu C B, Lin Y L. Minimizing risk models in Markov decision processes with policies depending on target values. *J Math Anal Appl*, 1999, 231: 47–67
 - 85 Ohtsubo Y, Toyonaga K. Optimal policy for minimizing risk model in Markov decision processes. *J Math Anal Appl*, 2002, 271: 66–81
 - 86 姜玉双. 关于马氏决策中概率准则的概述. 第五届中国青年运筹与管理学者大会论文集. 香港: Global-Link 出版公司, 2003, 195–204
 - 87 Sakaguchi M, Ohtsubo Y. Optimal threshold probability and expectation in semi-Markov decision processes. *Appl Math Comput*, 2010, 216: 2947–2958
 - 88 Huang Y H, Guo X P, Li Z F. Minimum risk probability for finite horizon semi-Markov decision processes. *J Math Anal Appl*, 2013, 402: 378–391
 - 89 Chen R C, Feinberg E A. Non-randomized policies for constrained Markov decision processes. *Math Methods Oper Res*, 2007, 66: 165–179
 - 90 Altman E. Constrained Markov decision processes with total cost criteria: Occupation measures and primal LP. *Math Methods Oper Res*, 1996, 43: 45–72
 - 91 Altman E. Constrained Markov decision processes with total cost criteria: Lagrangian approach and dual linear program. *Math Methods Oper Res*, 1998, 48: 387–417
 - 92 Kurano M, Huang Y Q. Average reward Markov decision processes with multiple cost constraints. *J Inf Optim Sci*, 1995, 16: 403–415
 - 93 Serin Y, Kulkarni V. Markov decision processes under observability constraints. *Math Methods Oper Res*, 2005, 61: 311–328
 - 94 Guo X P, Piunovskiy A. Discounted continuous-time Markov decision processes with constraints: Unbounded transition and loss rates. *Math Oper Res*, 2011, 36: 105–132
 - 95 Beutler F J, Ross K W. Time-average optimal constrained semi-Markov decision processes. *Adv Appl Probab*, 1986, 18: 341–359
 - 96 Feinberg E A. Constrained semi-Markov decision processes with average rewards. *Z Oper Res*, 1994, 39: 257–288
 - 97 胡光华, 张升. 约束折扣半马氏决策规划. *应用数学学报*, 1997, 20: 187–195
 - 98 Feinberg E A. Constrained discounted semi-Markov decision processes. In: *Markov Processes and Controlled Markov Chains* (Changsha, 1999). Dordrecht: Kluwer Acad Publ, 2002, 233–244
 - 99 Gonzalez-Hernandez J, Villarreal-Rodriguez C E. Optimal solutions of constrained discounted semi-Markov control problems. *Bol Soc Mat Mexicana*, 2013, 19: 111–124
 - 100 Huang Y H, Li Z F, Guo X P. Constrained optimality for finite horizon semi-Markov decision processes in Polish spaces. *Oper Res Lett*, 2014, 42: 123–129
 - 101 Markowitz H M. *Portfolio Choice: Efficient Diversification of Investment*. New York: Wiley, 1959
 - 102 Li D, Ng W L. Optimal dynamic portfolio selection: Multiperiod mean-variance formulation. *Math Finance*, 2000, 10: 387–406
 - 103 Guo X P, Ye L, Yin G. A mean-variance optimization problem for discounted Markov decision processes. *European J Oper Res*, 2012, 220, 423–429
 - 104 Huang Y H, Guo X P. Mean-variance problems for finite horizon semi-Markov decision processes. *Appl Math Optim*, doi: 10.1007/s00245-014-9278-9, 2014
 - 105 Wei Q D, Guo X P. Semi-Markov decision processes with variance minimization criterion. *4OR*, doi: 10.1007/s10288-014-0267-2, 2014
 - 106 Collins E J. Finite-horizon variance penalised Markov decision processes. *OR Spektrum*, 1997, 19: 35–39
 - 107 van Dijk N M, Sladký K. On the total reward variance for continuous-time Markov reward chains. *J Appl Probab*, 2006, 43: 1044–1052
 - 108 Filar J A, Kallenberg L C M, Lee H M. Variance-penalized Markov decision processes. *Math Oper Res*, 1989, 14: 147–161
 - 109 Guo X P, Song X Y. Mean-variance criteria for finite continuous-time Markov decision processes. *IEEE Trans Automat*

- Control, 2009, 54: 2151–2157
- 110 Prieto-Rumeau T, Hernández-Lerma O. Variance minimization and the overtaking optimality approach to continuous-time controlled Markov chains. *Math Meth Oper Res*, 2009, 70: 527–540
- 111 Hernández-Lerma O, Vega-Amaya O, Carrasco G. Sample-path optimality and variance-minimization of average cost Markov control processes. *SIAM J Control Optim*, 1999, 38: 79–93
- 112 White D J. Computational approaches to variance-penalised Markov decision processes. *OR Spektrum*, 1992, 14: 79–83
- 113 Sobel M J. The variance of discounted Markov decision processes. *J Appl Probab*, 1982, 19: 794–802
- 114 Yushkevich A A. Controlled Markov models with countable state space and continuous time. *SIAM Theory Probab Appl*, 1977, 22: 215–235
- 115 Yushkevich A A. On reducing a jump controllable Markov model to a model with discrete time. *SIAM Theory Probab Appl*, 1980, 25: 58–69

A survey on semi-Markov decision processes

GUO XianPing & HUANG YongHui

Abstract This paper is a survey on semi-Markov decision processes (SMDPs). We present the background, the significance, and the research actuality of the infinite horizon expected discounted reward criterion, the long-run expected average reward criterion, the finite horizon expected reward criterion, the expected first passage reward criterion, the probability criterion, constrained problems, and mean-variance problems in SMDPs. At the same time, some issues to be studied in the future for these criteria or problems are pointed out. We also discuss potential research directions for SMDPs.

Keywords semi-Markov decision processes, optimality criteria, optimality equation, optimal policy, algorithm

MSC(2010) 90C40, 93E20

doi: 10.1360/N012015-00041