DOI：10. 13471/j. cnki. acta. snus. 2022A020

风险敏感马氏决策过程与状态扩充变换*

马帅，夏俐

中山大学管理学院，广东 广州 510275

摘　要：在马氏决策过程中，过程的随机性由策略与转移核决定，优化目标的随机性受随机报酬与随机策略的影响，其中随机报酬往往可通过简化转化为确定型报酬。当优化准则为经典的期望类准则，如平均准则或折扣准则时，报酬函数的简化不会影响优化结果。然而对风险敏感的优化准则，此类简化将影响风险目标值，进而破坏策略的最优性。针对该问题，状态扩充变换将随机信息重组进扩充状态空间，在简化报酬函数的同时保持随机报酬过程不变。本文以三种定义于累积折扣报酬的经典风险测度为例，在策略评价中对比报酬函数简化与状态扩充变换对风险评估的影响。理论验证与数值实验均表明，当报酬函数形式较为复杂时，状态扩充变换可在简化报酬函数的同时保持风险测度不变。

关键词：马氏决策过程；状态扩充变换；风险；报酬函数简化

中图分类号：O177.2　　**文献标志码：**A　　**文章编号：**2097－0137（2023）01－0181－11

Risk-sensitive Markov decision processes and state augmentation transformation

MA Shuai, *XIA Li*

School of Business, *Sun Yat-sen University*, *Guangzhou 510275*, *China*

Abstract: In the theory of Markov decision processes, the randomness of the objective stems from not only the stochasticity of the process but also the randomnesses of the one-step reward and the policy. When the optimality criterion concerns only the risk-neutral expectation of the objective, the reward (function) simplification will not affect the optimization result. However, the simplification will change the stochastic reward sequence, which results in a modification to a risk-sensitive objective, i. e., a risk measure. Since some theoretical methods may require a simple reward function in a practical environment with a complicated one, to bridge this gap, we propose a technique termed state augmentation transformation, which preserves the stochastic reward sequence in a transformed process with a reward function in a simple form. Taking three classical risk measures (variance, exponential utility, and conditional value at risk) for example, the numerical experiment shows that the state augmentation transformation keeps the risk measures intact, while the reward simplification fails.

Key words: Markov decision process; state augmentation transformation; risk; reward simplification

* **收稿日期：**2022－03－09　　**录用日期：**2022－06－12　　**网络首发日期：**2022－10－20
基金项目：国家自然科学基金（62073346，U1811462）
作者简介：马帅（1987年生），男；**研究方向：**马氏决策过程、风险决策；E-mail：mash35@mail.sysu.edu.cn
通信作者：夏俐（1980年生），男；**研究方向：**随机学习与优化、马氏决策过程和强化学习等；
E-mail：xiali5@mail.sysu.edu.cn

马氏决策过程（MDP，Markov decision process），又称马氏控制过程（controlled Markov process）或随机动态规划（stochastic dynamic programming），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，决策者选取一个动作来控制或影响系统的演化，这种状态-动作映射即为一个策略。在无后效性的策略作用下，MDP将产生一个含报酬信号的马氏过程（MRP，Markov reward process）。在随机报酬过程$\{R_t\}$的基础上，MDP的优化准则（optimality criterion）量化了策略的性能。经典的优化准则主要考虑风险中性（risk-neutral）的累计报酬期望，主要分为累积（折扣）准则与长期平均准则。由于期望准则满足全期望公式且具有时间一致性（time-consistency），该准则下的最优策略可通过Bellman最优方程迭代得到。由于风险中性优化准则的良好性质，此类准则已被广泛研究[1-2]。然而经典理论中无风险概念的优化准则无法满足诸如金融、交通、医疗与能源等领域中风险敏感（risk-sensitive）工程问题的实际要求，即决策者难以接受伴有高风险的高收益。

随着人们对风险的愈发重视，针对MDP中风险准则的研究渐受关注。该研究通常包含两类问题，一类是当MDP模型信息不完备，由参数不确定性造成的风险。此类问题通常被称为鲁棒控制（robust control），决策者需针对最坏情况下的参数组合进行优化[3]。本文主要研究由MDP内在随机性引起的风险，此类问题被称为风险敏感MDP（risk-sensitive MDP）。风险敏感MDP是一个重要研究方向，通常对标风险中性MDP，与鲁棒控制和微分博弈（differential game）存在密切的联系，是对传统风险中性MDP的扩展。风险敏感MDP中，决策者需选取一个最优策略，在该策略下可以生成一个“好”的随机报酬过程$\{R_t\}$，其中R_t为$t\in\mathbb{N}$时刻所得一步报酬。对“好”的量化体现于优化准则中，通常用风险测度（risk measures）将一个策略下的$\{R_t\}$转化为标量，并考查该策略是否满足可能存在的约束集。风险敏感MDP中的风险测度ρ可以分为两类，一类着重考查$\{R_t\}$的动态性，通常定义为

$$\rho(M_d) \coloneqq \rho_0\Big(R_0+\rho_1\big(R_1+\rho_2(R_2+\cdots)\big)\Big),$$

其中ρ_t为$t\in\mathbb{N}$时刻的条件风险测度，此类风险测度被称为Markov风险测度（Markov risk measure）[4-5]。另一类测度定义在一个由$\{R_t\}$简化而来的静态随机变量，该静态随机变量通常被定义为累积（折扣）报酬或平均报酬。以无限阶段MDP为例，给定折扣因子$\gamma\in(0,1)$，其累积折扣报酬定义为

$$\Phi \coloneqq \sum_{t\in\mathbb{N}}\gamma^t R_t.$$

该随机变量也被称为收益（return），经典的期望准则与一系列风险测度皆定义于此类静态随机变量。相比于Markov风险测度，基于静态随机变量的风险测度被广泛研究，主要可分为三类：基于方差的测度、基于效用的测度与基于分位数的测度。

方差作为随机变量的中心二阶矩，是一种天然的风险测度。风险敏感MDP中的方差准则包括：

收益方差$\mathbb{V}(\Phi)$，该准则针对收益的方差进行优化。Sobel为带有确定性报酬的MRP收益方差给出了解析解[6]。Mannor和Tsitsiklis证明了有限阶段的均值-方差问题为NP-难[7]。Tamar等[8]为多种基于收益方差的优化准则提出了基于策略梯度的优化方法。Xie等[9]针对均值-方差问题提出了坐标下降法。

极限平均方差（limiting average variance）$\limsup\limits_{T\to\infty}\dfrac{1}{T}\mathbb{V}\left(\sum\limits_{t=0}^{T-1}R_t\right)$，该准则针对无折扣的累积报酬方差进行优化。在该方差准则下，Hernández-Lerma等[10]研究了在最优稳态收益策略空间中的最优策略存在性。Guo和Song为该类方差的有限性提出了“G-条件”（G-condition），并在该条件下证明了该方差准则与平均准则的等价性，该等价性可用于在平均准则下的最优策略集中搜索极限平均方差准则下的最优策略[11]。以上两类方差均针对累积报酬。

稳态方差（steady-state variance）

$$\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left\{\sum_{t=0}^{T-1}\left(R_t-\lim_{T\to\infty}\frac{1}{T}\mathbb{E}\left\{\sum_{t=0}^{T-1}R_t\right\}\right)^2\right\},$$

相比于前两类方差，该准则旨在量化一步报酬的稳定性。Sobel和Chung研究了带有均值约束的单链MDP中稳态方差优化问题[12-13]。Prashanth等[14]应用Actor-Critic算法估计策略梯度，进而优化稳态方差，该方法的局部收敛性可通过常微分方程证明。Gosavi[15]针对稳态方差提出了Q-learning算法，该算法在假设下

可收敛。Xia[16] 针对稳态方差的时变性，提出了“伪方差”的概念，进而提出了高效的策略迭代算法。基于该算法，Ma 等[17]在稳态方差的基础上引入折扣因子，以一步报酬波动性现值的累积为优化目标，针对一类基于该方差的优化准则提出了两层优化算法框架，并在该框架下提出了值迭代算法，并证明其局部收敛性。

由于有着诸多良好性质，基于方差的优化准则被广泛应用于金融、能源、交通与制造业等领域的风险敏感决策问题。金融市场中，Markowitz将方差引入优化目标，在投资组合领域提出了均值-方差优化方法[18]。这种方法被广泛应用于投资组合及对冲等金融问题[19]。能源领域中，当间歇性清洁能源（风电、水电、太阳能等）接入电网，如何借助储能设施，建立合理的充/放电策略，使得电网的稳态负载方差较小，对电网的安全性与经济性至关重要[20]。交通系统中，交通拥堵与安全等问题往往与交通流的波动性直接相关，尤其是在不久的将来，智能网联车逐渐增多，如何调控此类异质交通流将会成为研究热点[21]。工业界中，方差可以作为产品质量控制的优化目标，进而平稳生产流程，减小产品质量波动[22]。当被考查随机变量的分布近似正态分布时，方差是一个良好的风险测度。然而当分布的对称性较差，或随机变量的正/负偏差需要区别对待时，方差不再是一个合适的优化准则。

效用理论始于经济学，最早由 Morgenstern 和 von Neumann 于 1947 年提出[23]。效用理论将随机收益所产生的效用定义为确定性等价物（certainty equivalent），即与该随机收益具有相同效用值的确定性收益，该确定性收益取决于决策者对不同风险情况的主观评价。经典案例有阿莱悖论（Allais Paradox）[24] 与圣彼得堡悖论（St. Petersburg Paradox）[25]。阿莱悖论表示，决策者更愿意选择 100% 的概率得到 100 万元，而非 10% 的概率得到 500 万元，89% 的概率得到 100 万元，1% 的概率无收益，即使前者的期望收益小于后者。该情况出现的原因被归结为确定性效应（certainty effect），即决策者过度重视确定性的收益。圣彼得堡悖论表示，人们不愿意以较大的付出来参与一场收益期望无限大的游戏。该游戏中，参与者需投掷一枚硬币，若第一次投掷为正面，可得收益 2 且游戏结束；若第一次投掷为反面，则继续投掷，若第二次为正面则可得得收益 4，且游戏结束，如此，参与者若投掷不成功则继续投掷，直到成功。若第 n 次投掷成功，则收益为 2^n，游戏结束。人们不愿意以较大的付出来参与该游戏的原因主要被归结于决策者会弱化小概率事件的意义。上述例子中决策者的主观态度可以通过效用函数进行量化。风险敏感 MDP 中，效用函数形式通常为 $U^{-1}\{\mathbb{E}[U(\Phi)]\}$. 指数效用（exponential utility）是效用函数族中的经典形式，被应用于最早的风险敏感 MDP 模型[26]，由于其结构的良好性质，可以构成特殊的乘法形式 Bellman 方程。该效用可表示为

$$\beta^{-1}\log\left\{\mathbb{E}\left[\exp\left(\beta\Phi\right)\right]\right\},$$

即 $U(x)\coloneqq\exp(\beta x)$. Chung 等[27]首次针对收益的指数效用研究了基于收益分布的不动点定理。Bäuerle 等[28]证明 MDP 中指数效用准则可通过定义扩充状态空间进而通过值迭代算法求解。Zhang 为连续时间 MDP 中的指数效用准则建立了最优方程，并证明了最优确定性平稳策略的存在性[29]。实际工程中，指数效用准则被应用于军事[30]、金融[31]与交通[32]等领域。

分位数是对随机变量分布最直接的刻画。风险价值（VaR，value at risk）是一种经典的基于分位数的测度，它起源于金融界，由 JP 摩根（J P Morgan）于 20 世纪 80 年代提出，并于 90 年代被列入到《巴塞尔协议》中。作为商业银行资产风险评估的标准之一，VaR 刻画了在一定的概率水平（α）下收益的最小可能值（τ）。从数学上讲，数值对（τ，α）为随机变量累积分布函数（CDF，cumulative distribution function）上的点，而 α -VaR 即 α 分位点。Filar 等[33]为风险敏感 MDP 中基于 VaR 的研究定义了两类问题：给定 α 下 τ 的优化与给定 τ 下 α 的优化。虽然两个问题都是对收益 CDF 的直接优化，但在风险敏感 MDP 中的方法却不尽相同[34]。VaR 虽然是一种直观的风险测度，但并不具有良好的数学性质（如凸性），不能很好地度量尾部风险，且不满足一致性公理。在 VaR 的基础上，Rockafellar 等[35]于 2000 年提出一种新的风险测度——条件风险价值（CVaR，conditional VaR）。CVaR 又被称为 expected shortfall、average value at risk 或 expected tail loss，它量化了在收益不小于给定 VaR 值的条件下收益的平均值。与 VaR 相比，CVaR 满足次可加性、正齐次性、单调性及传递不变性，因而 CVaR 是一种一致性（coherent）风险测度[36]。由于具有较好的数学性质，CVaR 在风险敏感 MDP 中具有较为广泛的研究。Borkar 和 Jain 针对带有 CVaR 约束的有限阶段 MDP 问题提出了动态规划算法，并证明了算法的收敛性。然而该算法涉及连续变量的积分，在实际应用中难以

实施[37]。Bäuerle 和 Ott 证明了 CVaR 准则下存在最优 Markov 策略，该策略定义在包含了累积报酬的扩充状态空间上[38]。基于该扩充空间，Haskell 和 Jain 为 CVaR 准则下的 MDP 问题提出了基于数学规划的算法，然而该非凸规划需要通过求解一系列的线性规划进行近似求解[39]。Prashanth 针对带有 CVaR 约束的 MDP 问题提出了策略梯度算法，该算法可收敛至局部最优[40]。Chow 等从鲁棒优化的角度分析了 CVaR 准则下的 MDP 问题，证明了其与带约束鲁棒优化问题的等价性，并提出了近似值迭代算法[41]。除了金融领域[42]，CVaR 也被广泛应用于能源[43]、交通[44]与医疗[45]等领域中。针对 CVaR 的综述，见文献[46]。

由文献综述可见，针对不同的风险测度，学者们提出了诸多理论方法，然而理论方法与工程问题常存有差异。对于风险敏感 MDP 而言，这种差异的主要形式之一就是报酬函数的差异。当系统的不确定性来源复杂时，风险敏感 MDP 中的报酬函数形式将随之变得复杂。理论方法中的 MDP 报酬通常是确定性的、基于当前状态的[47-51]，即 $R_t = r(X_t,\ K_t)$，其中 r 为报酬函数，X_t 与 K_t 分别为 $t \in \mathbb{N}$ 时刻的状态与动作；而工程问题中的报酬可能是随机的、基于状态转移的，如 $R_t \sim r(X_t,\ K_t,\ X_{t+1})$，其中 r 为报酬分布函数。这种报酬函数形式的差异对风险中性的期望准则而言无关紧要，通常方法即将报酬函数进行线性简化（见定义 1）。然而对于风险敏感 MDP 而言，这种对报酬函数的简化将改变随机报酬过程 $\{R_t\}$，进而改变绝大部分风险测度。以累积折扣报酬的方差为例，Sobel 为带有确定性报酬函数的无限阶段离散 MRP 给出了方差评估算法，然而该方法无法直接应用于带有随机报酬的 MRP[6]。针对此类问题，一种解决方案是对报酬函数进行简化，然而该简化将改变 MRP 的 $\{R_t\}$，进而改变累积折扣报酬的方差。另一种方法是针对此类问题开发专门的（ad hoc）算法，但这种算法的设计开发需要工程相关的从业人员对问题本质有着深度的理解。如何从实际问题出发，考虑绝大部分风险测度，将针对简单模型的理论方法与实际中的复杂工程问题合理对接，是风险敏感 MDP 中的一个重要问题，具有一定的理论意义和广泛的应用背景。

状态扩充变换（SAT，state augmentation transformation）针对风险敏感 MDP，将带有复杂报酬函数的 MDP 变换为带有简单报酬函数的 MDP，且保证相同策略（原始策略与对应扩充策略）下 MRP 的随机报酬过程 $\{R_t\}$ 不变。本文针对 MDP 中的策略评价，通过数值实验，在给定策略下的 MDP 中考查三类常用的风险测度：方差、指数效用与条件风险价值，并对比通过 SAT 与报酬函数简化所得三类风险的差异，进而验证 SAT 对带有复杂报酬函数/随机策略的 MDP 中风险敏感策略评价的有效性。理论验证与数值实验均表明，当报酬函数形式较为复杂时，状态扩充变换可在简化报酬函数的同时保持风险测度不变。故而在不确定性来源复杂的风险敏感工程问题中，需通过 SAT 而非简化报酬函数来对 MDP 进行报酬函数形式上的简化。最后，讨论 SAT 的一些潜在发展方向。

1 风险敏感 MDP 模型

1.1 MDP 模型

本文主要研究无限阶段时齐（time-homogeneous）离散 MDP，其状态与动作数量均为有限。一个 MDP 可定义如下：

$$M := \langle S,\ A, r, p, \mu, \gamma \rangle,$$

其中 S 为有限状态空间，令 X_t 为 $t \in \mathbb{N}$ 时刻系统的状态；A 为有限动作空间，通常有 $A = \bigcup\limits_{x \in S} A(x)$，其中 $A(x)$ 为 $x \in S$ 的可行动作空间，令 $K_t \in A(X_t)$ 为 t 时刻决策者在系统状态 X_t 下选择的动作；p 为条件转移概率，给定 x，$y \in S$，$a \in A(x)$，有 $p\left(y \middle| x,\ a\right) = \mathbb{P}\left(X_{t+1} = y \middle| X_t = x,\ K_t = a\right)$；$\mu \in \Delta(S)$ 为系统初始状态分布①；$\gamma \in (0,\ 1)$ 为折扣因子。本文针对不同系统的报酬特性，考虑以下四类报酬：

（i）确定性的、基于状态的报酬 r_{DS}：$S \times A \to \mathbb{R}$；

（ii）确定性的、基于状态转移的报酬 r_{DT}：$S \times A \times S \to \mathbb{R}$；

（iii）随机性的、基于状态的报酬 r_{SS}：$S \times A \to \Delta(\mathbb{R})$；

（iv）随机性的、基于状态转移的报酬 r_{ST}：$S \times A \times S \to \Delta(\mathbb{R})$.

$r \in \{r_{\mathrm{DS}},\ r_{\mathrm{DT}},\ r_{\mathrm{SS}},\ r_{\mathrm{ST}}\}$ 为系统的报酬函数或报酬分布函数，令 $R_t \in [-C,\ C]$ 为 t 时刻的一步报酬，其中

① $\Delta(S)$ 为定义在 S 上的概率分布空间。

$C \in \mathbb{R}$为一步报酬绝对值的上确界。简洁起见，相同报酬函数表述也被使用于MRP。对于随机性报酬，本文仅考虑离散随机报酬分布。

策略描述了决策者如何选择动作。针对无限阶段MDP，本文仅考查平稳Markov策略，即当前动作的选择仅依赖于当前状态而非整个历史，且策略不随时间改变。用D表示平稳Markov策略空间，其可进一步分为确定性策略空间D_d与随机性策略空间D_r。M在策略$d \in D_d$的作用下将构成$M_d = \langle S,\ r_d,\ p_d,\ \mu,\ \gamma \rangle$②。需注意的是，$M$在策略$d \in D_r$的作用下构成的$M_d$不能直接表述为$\langle S,\ r_d,\ p_d,\ \mu,\ \gamma \rangle$，这是因为该表述暗示了报酬函数的部分简化，进而改变$\{R_t\}$。这也是下文中，情况3无法与情况2建立等价性的原因。

定义1 (报酬函数线性简化） 给定一个M与策略$d \in D$，若所得MRP的报酬(分布)函数r_d非r_{DS}型，则可通过计算条件期望将r_d简化为r_{DS}。考虑最一般化的形式，以一个带有r_{ST}的M在随机策略$d \in D_r$下所生成的M_d为例，其报酬函数可作如下线性简化：

$$r'_d(x) = \sum_{a \in A(x)} d(a|x) \sum_{y \in S} p(y|x,a) \sum_{j \in \mathrm{supp}\{r_d(\cdot|x,\ a,\ y)\}} j\, r_d(j|x,a,y),$$

其中$\mathrm{supp}\{r_d(\cdot|x,\ a,\ y)\}$表示分布$r_d(\cdot|x,\ a,\ y)$的支集(support)。

当优化准则为风险中性的平均准则或折扣准则时，报酬函数的线性简化不会影响策略的最优性。然而优化目标为风险测度时，报酬函数的线性简化将改变M_d的$\{R_t\}$，进而改变策略的最优性。下文将介绍三种常用风险测度的计算或估计。

1.2　风险测度

本部分内容主要介绍三种经典风险测度：方差、指数效用与CVaR。针对MRP的收益，三种风险测度可定义如下。

方差　方差作为随机变量的中心二阶矩，是最具代表性的风险测度之一。MRP中收益的方差定义为

$$\rho_v(\Phi) \coloneqq \mathbb{V}_\mu(\Phi) = \mathbb{E}_\mu\left\{\left[\Phi - \mathbb{E}_\mu(\Phi)\right]^2\right\},$$

其中$\mathbb{E}_\mu$与$\mathbb{V}_\mu$为给定系统初始状态分布μ时的期望与方差。Sobel基于Bellman方程，为带有确定性报酬的MRP中收益的方差提供了一种高效计算方法。

定理1（收益方差[6]） 对于任意MRP $M_d = \langle S,\ r_d,\ p_d,\ \mu,\ \gamma \rangle$，其中$r_d$为$r_{\mathrm{DS}}$形式，令$P$为转移概率矩阵，即$P(x,\ y) = p_d(y|x) \coloneqq \sum_{a \in A(x)} d(a|x)\, p(y|x,\ a)$，$x,\ y \in S$. 令$v$为(期望)值函数(向量)，则$v(x) = \mathbb{E}_x(\Phi)$；令$\psi$为方差值函数(向量)，则$\psi(x) = \mathbb{V}_x(\Phi)$，其中$\mathbb{E}_x$与$\mathbb{V}_x$为给定系统初始状态$x \in S$时的期望与方差。令$\theta$对应$\psi$的报酬函数（向量），有$\theta(x) = \sum_{y \in S} P(x,\ y)\left[r_d(x) + \gamma v(y)\right]^2 - v^2(x)$. 则有

$$v = r_d + \gamma P v = (I - \gamma P)^{-1} r_d,$$

$$\psi = \theta + \gamma^2 P \psi = (I - \gamma^2 P)^{-1} \theta.$$

定理1为MRP收益的方差给出了一种类Bellman方程的高效算法，但该算法仅针对带有确定性报酬的MRP。

指数效用　给定一个风险敏感系数$\beta \in \mathbb{R}$，MRP的指数效用为

$$\rho_e(\Phi, \beta) \coloneqq \frac{1}{\beta} \ln\left[\mathbb{E}_\mu\left\{\exp(\beta\Phi)\right\}\right].$$

指数效用与方差联系紧密，其Taylor展开形式为

$$\rho_e(\Phi, \beta) = \mathbb{E}_\mu(\Phi) + \frac{\beta}{2}\mathbb{V}_\mu(\Phi) + \mathcal{O}(\beta^2), \tag{1}$$

其中$\mathcal{O}(\cdot)$为无穷小渐近。由此可知，当$\beta < 0$时，该准则为一种风险规避准则。当β足够小时，该准则可以用收益的期望与方差近似估计。

CVaR　CVaR是当收益值超过某置信度下的VaR情况时的条件数学期望，VaR是收益在给定置信度

② 此处忽略策略对状态空间的可能影响。

$\alpha\in(0,\ 1)$下的最小收益值。给定一个置信度α，MRP的VaR定义为：

$$\mathrm{VaR}_{\alpha}(\Phi)\coloneqq\inf\{x\in\mathbb{R}|\mathbb{P}(\Phi\leqslant x)\geqslant\alpha\}.$$

则MRP在给定置信度α下的CVaR可定义为

$$\rho_c(\Phi,\ \alpha)\coloneqq\frac{1}{1-\alpha}\int_{\alpha}^{1}\mathrm{VaR}_t(\Phi)\mathrm{d}t=\mathbb{E}_{\mu}\{\Phi\,|\,\Phi\geqslant\mathrm{VaR}_{\alpha}(\Phi)\}.$$

虽然CVaR作为一类一致性风险测度，具有良好的数学性质，但难以在MRP中被高效评估。本文通过假设收益的分布近似服从正态分布，进而对指数效用与CVaR进行近似估计。

假设1 MRP的收益近似服从正态分布。

在假设1下，CVaR可如下估计：

$$\rho_c(\Phi,\alpha)\approx\mu+\sigma\frac{g\left[G^{-1}(\alpha)\right]}{1-\alpha},$$

其中g与G分别表示标准正态分布$\mathcal{N}(\mu,\ \sigma^2)$的概率密度函数和累积分布函数，该式被称为逆米尔斯比率（inverse Mills ratio）。更多常用常见分布的CVaR表达式可见文献[52]。

当一个带有r_{ST}的MDP/MRP需要应用一种针对带有r_{DS}模型的理论方法时，该如何处理方法与模型在报酬函数上的差异？一种方法是为特定问题开发新算法，但这种方法需要工程相关的从业人员对问题本质有着深度的理解。另一种方法是应用SAT将其变换为一个带有确定性报酬的MDP/MRP。

2 状态扩充变换

针对理论方法与实际问题由于报酬函数的差异而引起的风险测度优化与评估的问题，本文研究了状态扩充变换（以下简称SAT）[53]。该方法针对上述问题，从策略优化与评价两个角度为两类MDP/MRP建立等价形式，即对于一个带有复杂报酬函数的MDP/MRP，SAT可以将其转换为一个带有简单报酬函数的MDP/MRP，且两者的$\{R_t\}$相同。本文针对MDP中的策略评价，考查三类不确定性来源：由状态转移导致的不确定性、报酬本身的随机性与策略的随机性。将不确定性来源依次扩展，定义如下三种情况。

情况1：带有r_{DT}的M_d；

情况2：带有r_{ST}的M_d；

情况3：带有r_{ST}的M和一个$d\in D_r$.

其中情况1为早期SAT考虑的问题，因其针对带有基于状态转移报酬函数的MRP，故又称状态转移变换[54]。情况2为情况1的拓展，考虑了更一般化的报酬函数。情况3将问题进一步扩展，将由策略引起的随机性考虑进来。三种情况中前者为后者的特殊形式，若以“$<$”表示此种关系，则有

$$\text{情况1}<\text{情况2}<\text{情况3}.$$

针对情况3，SAT定义如下。

定义2（情况3下的SAT） 对于在策略$d\in D_r$下的$M=\langle S,\ A,\ r,\ p,\ \mu,\ \gamma\rangle$，其中$r$为$r_{\mathrm{ST}}$形式，其经SAT变换所得MRP可表示为$M_d^{\dagger}=\langle S^{\dagger},\ r_d^{\dagger},\ p_d^{\dagger},\ \mu^{\dagger},\ \gamma\rangle$，其中$S^{\dagger}=S^2\times A\times J$为扩充状态空间，$J\coloneqq\bigcup\limits_{x,\ y\in S,\ a\in A(x)}\mathrm{supp}\{r_d(\cdot\,|\,x,\ a,\ y)\}$为所有可能报酬值的集合；$r_d^{\dagger}\colon S^{\dagger}\to\mathbb{R}$为定义在$S^{\dagger}$上的$r_{\mathrm{DS}}$形式报酬函数，对任意$x^{\dagger}=(x,\ a,\ y,\ j)\in S^{\dagger}$，$x,\ y\in S$，$a\in A(x)$，$j\in\mathrm{supp}\{r_d(\cdot\,|\,x,\ a,\ y)\}$，有$r_d^{\dagger}(x^{\dagger})=j$；$p_d^{\dagger}(y^{\dagger}\,|\,x^{\dagger})\coloneqq d(a'\,|\,y)p_d(z\,|\,y,\ a')r_d(j'\,|\,y,\ a',\ z)$为定义在$S^{\dagger}$上的条件转移概率，其中$x^{\dagger}$，$y^{\dagger}=(y,\ a',\ z,\ j')\in S^{\dagger}$，$z\in S$，$a'\in A(y)$，$j'\in\mathrm{supp}\{r_d(\cdot\,|\,y,\ a',\ z)\}$；$\mu^{\dagger}(x^{\dagger})\coloneqq\mu(x)d(a\,|\,x)p_d(y\,|\,x,\ a)r_d(j\,|\,x,\ a,\ y)$为定义在$S^{\dagger}$上的初始状态分布。

对于情况3下的SAT有如下定理。

定理2（SAT作用下的随机报酬过程等价性） 对于任意MDP $M=\langle S,\ A,\ r,\ p,\ \mu,\ \gamma\rangle$，其中$r$为$r_{\mathrm{ST}}$形式，在策略$d\in D_r$下所产生的$M_d$与SAT变换所得$M_d^{\dagger}$的$\{R_t\}$相同。

证明 考虑M_d下任意样本路径$\omega=(s_0,\ a_0,\ s_1,\ j_1,\ a_1,\ s_2,\ j_2,\ a_2,\ \cdots)$。对任意$t\in\mathbb{N}$，令$\omega(t)=(s_0,\ a_0,\ s_1,\ j_1,\ a_1,\ s_2,\ j_2,\ a_2,\ \cdots,\ s_t,\ a_t,\ s_{t+1},\ j_{t+1})$及其概率$\mathbb{P}(\Omega(t)=\omega(t))$。对应该样本路径，在$M_d^{\dagger}$下

有 $\omega^{\dagger}=\left(s_0^{\dagger},\ s_1^{\dagger},\ \cdots\right)=((s_0,\ a_0,\ s_1,\ j_1),\ (s_1,\ a_1,\ s_2,\ j_2),\ \cdots)$ 与 $\omega^{\dagger}(t)=\left(s_0^{\dagger},\ s_1^{\dagger},\ \cdots,\ s_t^{\dagger}\right)=((s_0,\ a_0,\ s_1,\ j_1),\ (s_1,\ a_1,\ s_2,\ j_2),\ \cdots,\ (s_t,\ a_t,\ s_{t+1},\ j_{t+1}))$，则易证 $\mathbb{P}\left(\Omega^{\dagger}(t)=\omega^{\dagger}(t)\right)=\mathbb{P}\left(\Omega(t)=\omega(t)\right)$，由于该式对任意 $t\in\mathbb{N}$ 成立，可知 M_d 与 $M_d^{\dagger}$ 的 $\{R_t\}$ 相同。　　证毕

该定理描述了情形 3 中两个带有不同类型报酬函数的 MRP 关于 $\{R_t\}$ 的等价性，而当两个 MRP 的 $\{R_t\}$ 相同时，其风险测度必然相同。针对 MDP 的 SAT 被证明于文献[53]，并于文献[55]从概率空间的角度被进一步补充。值得注意的是，当直接将 SAT 应用于 MDP 进行策略优化时，由于状态空间的扩充，对应策略空间也需要扩充。应在扩充策略空间上增加相应约束，进而保证其与原策略空间的一一映射关系，详见文献[55]。由定理 2 出发，可得针对情况 1 与 2 的推论，此处以情况 2 为例给出相应推论。

推论 1（情况 2 下的 SAT）　对于任意 $M_d=\left\langle S,\ r_d,\ p_d,\ \mu,\ \gamma\right\rangle$，其中 r_d 为 r_{ST} 形式，存在 $M_d^{\dagger}=\left\langle S^{\dagger},\ r_d^{\dagger},\ p_d^{\dagger},\ \mu^{\dagger},\ \gamma\right\rangle$，其中 $r_d^{\dagger}$ 为 r_{DS} 形式，使得 M_d 与 $M_d^{\dagger}$ 的 $\{R_t\}$ 相同。

针对该推论的证明详见文献[53]。依据推论 1，以一个带有 r_{SS} 报酬函数的二状态 MRP 为例，SAT 的作用如图 1 所示。图中圆圈表示随机过程的状态，箭头表示状态转移，其上方的数字表示对应的转移概率，状态旁的方框表示报酬，随机性报酬表示为报酬值与括号中的概率。该图示直观地解释了 SAT 在简化报酬函数的同时保持 $\{R_t\}$ 不变的原理，即将对一步报酬有影响的因素综合为一个扩充状态，该扩充状态可以被理解为与报酬对应的“情况”。SAT 作用下产生的随机过程保留了原过程的 Markov 性，且新的转移核可由原 MRP 的转移核与报酬/策略的分布计算而得。图 1 中，带有随机报酬的状态 y 被扩充为两个状态：y_1 与 y_2，分别代表了状态为 y 时，获取报酬值为 −1 与 1 的两种“情况”。基于扩充状态空间，该 MRP 的转移概率可由原转移概率与状态 y 上的报酬分布计算而得。

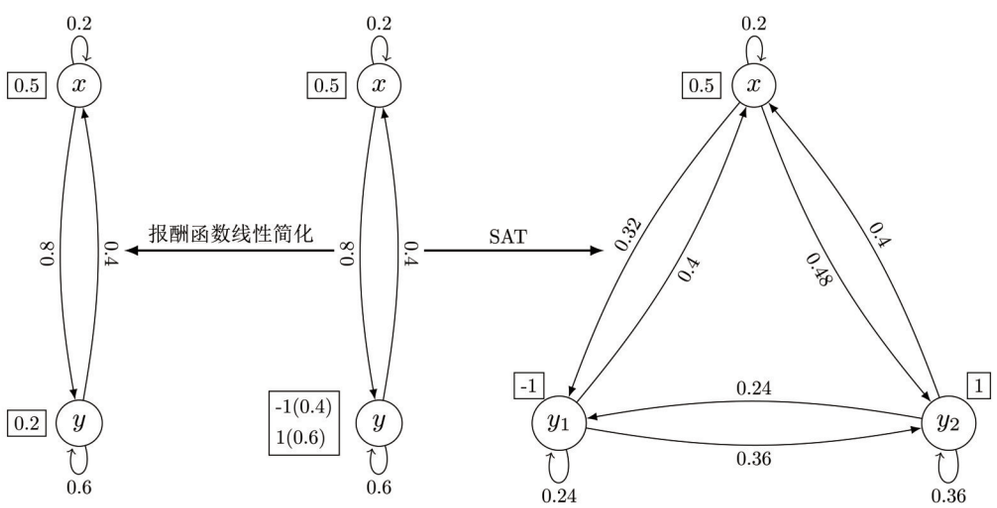

图 1　一个 MRP 在报酬函数线性简化与 SAT 作用下的两种变换

Fig. 1　The linear reward simplification and the SAT on an MRP

3　数值实验

本部分内容以图 1 所示 MRP 为例，通过数值实验考查报酬函数简化对三种风险测度的影响，同时验证 SAT 的有效性。由于指数效用与 CVaR 的估计均基于假设 1，首先验证该假设对于此 MRP 是否成立，该验证可量化为近似分布与真实分布的误差分布的尾部概率。

3.1　近似分布的误差

近似分布（ACDF，Approximated CDF，以 $\tilde{F}$ 表示）即在假设 1 下通过计算收益的期望与方差所得的正态分布。近似经验分布（AECDF，Approximated Empirical CDF，以 $\hat{F}_N$ 表示）即通过 Monte Carlo 仿真所得经验分布。ACDF 与 AECDF 间的差异可由仿真结果测量而得。设数值实验共进行 M 组，每组进行 N 次仿真，每次仿真 H 阶段，每组仿真可得 N 个近似收益值。在第 $i\in\{1,\ \cdots,\ M\}$ 组中，令 $Z_{(k)}^{i}$ 表示第

$k \in \{1, \cdots, N\}$小的收益值，则 $\hat{Z}_{(k)} = \sum_{i \in \{1, \cdots, M\}} Z^i_{(k)}/M$ 可作为近似分位点以表示 AECDF。则 ACDF 与 AECDF 间的差异δ可通过Kolmogorov-Smirnov统计量[56]量化

$$\delta = \left\| \tilde{F} - \hat{F}_N \right\| := D_{KS} = \sup_{x \in \mathbb{R}} \left| \tilde{F}(x) - \hat{F}_N(x) \right| = \max_{x \in \left\{ \hat{Z}_{(k)} \right\}_{k \in \{1, \cdots, N\}}} \left| \tilde{F}(x) - \hat{F}_N(x) \right|.$$

AECDF 与经验分布（ECDF，Empirical CDF，以 $\hat{F}$ 表示）间差异的上界可计算如下

$$\delta' = \left\| \hat{F}_N - \hat{F} \right\| \leqslant \sum_{t=H+1}^{\infty} \gamma^{t-1} C .$$

当H相对γ足够大时，该差异可忽略不计。ECDF 与 CDF 间的差异由 DKW 不等式（Dvoretzky-Kiefer-Wolfowitz inequality）保证：

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \left| \hat{F}(x) - F(x) \right| > \varepsilon \right) \leqslant 2\mathrm{e}^{-2N\varepsilon^2}, \quad \varepsilon > 0.$$

命题 1（近似分布误差）

$$\mathbb{P}\left(\sup_{x \in \mathbb{R}} \left| \tilde{F}(x) - F(x) \right| > \varepsilon + \delta + \delta' \right) \approx \mathbb{P}\left(\sup_{x \in \mathbb{R}} \left| \tilde{F}(x) - F(x) \right| > \varepsilon + \delta \right) \leqslant 2\mathrm{e}^{-2N\varepsilon^2}, \quad \varepsilon > 0.$$

当 ACDF 与 AECDF 相似度较高时，该近似分布的误差概率界效果较好。

3.2 仿真结果

设初始分布$\mu(x) = 1$（即初始状态为x），γ=0.95，M=20，N=100，H=500，此时$\delta' \leqslant 1.454\,9 \times 10^{-10}$。通过应用 Monte Carlo 仿真模拟，可获取N个分位数的均值与样本标准差，进而绘制带有误差区域的 AECDF。分别计算报酬函数简化与 SAT 作用后的 MRP 的期望与方差，并在假设 1 下绘制两者的 ACDF。三条分布曲线如图 2 所示。由图可见，在假设 1 下，SAT 所得收益的 ACDF 与 AECDF 相似度较高（$\delta \approx 0.016\,3$），而报酬函数简化所得收益的 ACDF 与 AECDF 相似度很低。

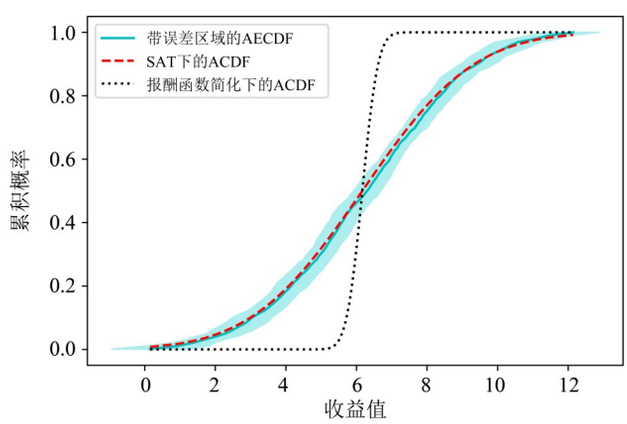

图 2 近似经验分布（AECDF）与假设 1 下的两个近似分布（ACDF）对比，两者的方差分别在报酬函数简化与 SAT 作用下通过定理 1 进行估计

Fig. 2 A comparison between the approximated empirical CDF and the two approximated CDFs, whose variances are calculated by Theorem 1 with the aid of the SAT and the reward simplification, respectively

SAT 与报酬函数简化作用下 MRP 的三种风险测度与仿真结果对比于表 1。通过对比可见，SAT 下的方差和 CVaR 与仿真结果较为接近，而报酬函数简化下的结果则相差甚远。在不同风险敏感参数下对比指数效用，可见相对报酬函数简化下的结果，SAT 所得结果与仿真结果更为接近。随着风险敏感参数的增大，SAT 对指数效用的估计精度也逐渐降低，这是因为式(1)中的误差项随着风险敏感参数的增大而增大。

4 结论与展望

风险敏感 MDP 是一类广泛且重要的随机动态决策问题，由于不同风险测度的特性各有不同，且风险敏感的应用场景较多，目前研究活跃且成果丰富。然而理论方法与实际问题间常有差异，若不能妥善处

表 1　三种风险测度在SAT、报酬函数简化与仿真模拟中的结果对比

Table 1　The comparison among the three risk measures with the SAT, the reward simplification and the simulation

	期望	方差	CVaR（α=0. 9）	指数效用		
				$\beta=1$	$\beta=0.1$	$\beta=0.01$
SAT	6. 168 1	6. 144 9	10. 518 5	9. 240 5	6. 475 3	6. 198 8
报酬函数简化	6. 168 1	0. 122 9	6. 783 3	6. 229 5	6. 174 0	6. 168 7
仿真	6. 108 4	6. 117 6	10. 189 8	8. 678 2	6. 528 0	6. 259 2

理此类差异，则将错误评估风险程度，以致决策失败。本文针对无限阶段风险敏感MDP理论方法与实际问题在报酬函数上的差异，研究了SAT方法，并通过仿真实验，对比了SAT与报酬函数简化对三类常用的风险测度的影响。数值结果显示，通过SAT所得到的数值与仿真结果较为接近，而报酬函数简化将大幅改变风险测度值。SAT的本质在于通过扩充状态空间，保留了完整的$\{R_t\}$信息，进而在简化报酬函数的同时保持风险测度不变。该方法为理论研究提供了带有不同报酬函数的MRP间的等价性，并为相关从业人员提供了一种直接将理论方法应用于复杂实际问题的解决方案。

SAT在策略评价情景中的应用较为直观，而在决策优化情景中的应用则较为复杂。将SAT直接应用于MDP进而优化决策时，由于扩充了状态空间，该MDP的策略空间也被扩充，故需对扩充策略空间加以约束，以保证与原策略空间的一一对应。SAT的另一个问题是状态空间规模的扩充导致问题维度组合式增大。考虑到定义在扩充状态空间上的转移概率与原MDP的转移概率信息量相同，如何降低扩充问题的维度是值得研究的问题。Ma和Yu针对扩充状态的相似性，给出了状态归并(state lumping)的条件，满足该条件的状态可归并为一个状态，且不影响风险测度[55]。处理该问题的另一种思路是从报酬值的差异程度出发，当两个扩充状态由同一原始状态扩充而来，且两者报酬值差异不大时，可近似为一个状态，这种近似会导致风险测度的改变，而这种差异的上界应为报酬值差异的函数。

参考文献：

[1] PUTERMAN M L. Markov decision processes：Discrete stochastic dynamic programming [M]. New Jersey：John Wiley & Sons，2014.

[2] 刘克，曹平. 马尔可夫决策过程理论与应用 [M]. 北京：科学出版社，2015.

[3] DULLERUD G E，PAGANINI F. A course in robust control theory：A convex approach [M]. New York：Springer Science & Business Media，2013.

[4] RUSZCZYŃSKI A. Risk-averse dynamic programming for Markov decision processes [J]. Math Program，2010，125(2)：235−261.

[5] BÄUERLE N，GLAUNER A. Markov decision processes with recursive risk measures [J]. European J Oper Res，2022，296(3)：953−966.

[6] SOBEL M J. The variance of discounted Markov decision processes [J]. J Appl Probab，1982，19(4)：794−802.

[7] MANNOR S，TSITSIKLIS J. Mean-variance optimization in Markov decision processes [C]// In Proceedings of the 28th International Conference on Machine Learning，2011：177−184.

[8] TAMAR A，DI CASTRO D，MANNOR S. Policy gradients with variance related risk criteria [C]// In Proceedings of the 29th International Conference on Machine Learning，2012：1651−1658.

[9] XIE T，LIU B，XU Y，et al. A block coordinate ascent algorithm for mean-variance optimization [C]// In Proceedings of the 32nd International Conference on Neural Information Processing Systems，2018：1073−1083.

[10] HERNÁNDEZ-LERMA O，VEGA-AMAYA O，CARRASCO G. Sample-path optimality and variance-minimization of average cost Markov control processes [J]. SIAM J Control Optim，1999，38(1)：79−93.

[11] GUO X，SONG X. Mean-variance criteria for finite continuous-time Markov decision processes [J]. IEEE Trans Automat Control，2009，54(9)：2151−2157.

[12] SOBEL M J. Mean-variance tradeoffs in an undiscounted MDP [J]. Oper Res，1994，42(1)：175−183.

[13] CHUNG K J. Mean-variance tradeoffs in an undiscounted MDP：The unichain case [J]. Oper Res，1994，42(1)：184−188.

[14] PRASHANTH L A，GHAVAMZADEH M. Actor-critic algorithms for risk-sensitive MDPs [C]//In Proceedings of the Neural Information Processing Systems，2013：252−260.

[15] GOSAVI A. Variance-penalized Markov decision processes：Dynamic programming and reinforcement learning techniques [J]. Int J Gen Syst，2014，43(6)：649−669.

[16] XIA L. Optimization of Markov decision processes under the variance criterion [J]. Automatica，2016，73：269−278.

[17] MA S，MA X，XIA L. A unified algorithm framework for mean-variance optimization in discounted Markov decision processes [OL]. arXiv:2201.05737，2022.

[18] MARKOWITZ H. Portfolio selection [J]. J Finance，1952，7(1)：77−91.

[19] KOUVELIS P，PANG Z，DING Q. Integrated commodity inventory management and financial hedging：A dynamic mean−variance analysis [J]. Prod Oper Manag，2018，27(6)：1052−1073.

[20] DONG M，LI Y，SONG D，et al. Uncertainty and global sensitivity analysis of levelized cost of energy in wind power generation [J]. Energy Convers Manag，2021，229：113781.

[21] YAO Z，XU T，JIANG Y，et al. Linear stability analysis of heterogeneous traffic flow considering degradations of connected automated vehicles and reaction time [J]. Phys A，2021，561：125218.

[22] HARRISON C A，QIN S J. Minimum variance performance map for constrained model predictive control [J]. J Process Control，2009，19(7)：1199−1204.

[23] MORGENSTERN O，von NEUMANN J. Theory of games and economic behavior [M]. New Jersey：Princeton University Press，1953.

[24] ALLAIS M. Le comportement de l'homme rationnel devant le risque：Critique des postulates et axiomes de l'ecole Americaine [J]. Econometrica，1953，21：503−546.

[25] BERNOULLI D. Exposition of a new theory on the measurement of risk [J]. Econometrica，1954，22：23−36.

[26] HOWARD R A，MATHESON J E. Risk-sensitive Markov decision processes [J]. Manag Sci，1972，18(7)：356−369.

[27] CHUNG K J，SOBEL M J. Discounted MDP's：Distribution functions and exponential utility maximization [J]. SIAM J Control Optim，1987，25(1)：49−62.

[28] BÄUERLE N，RIEDER U. More risk-sensitive Markov decision processes [J]. Math Oper Res，2014，39(1)：105−120.

[29] ZHANG Y. Continuous-time Markov decision processes with exponential utility [J]. SIAM J Control Optim，2017，55(4)：2636−2660.

[30] SPEYER J. An adaptive terminal guidance scheme based on an exponential cost criterion with application to homing missile guidance [J]. IEEE Trans Automat Control，1976，21(3)：371−375.

[31] AVILA-GODOY G，FERNÁNDEZ-GAUCHERAND E. Controlled Markov chains with exponential risk-sensitive criteria：Modularity，structured policies and applications [C]// In Proceedings of the IEEE Conference on Decision and Control，1998：778−783.

[32] BREZAS P，SMITH M C. Linear quadratic optimal and risk-sensitive control for vehicle active suspensions [J]. IEEE Trans Control Syst Technol，2013，22(2)：543−556.

[33] FILAR J A，KRASS D，ROSS K W，et al. Percentile performance criteria for limiting average Markov decision processes [J]. IEEE Trans Automat Control，1995，40：2−10.

[34] WU C，LIN Y. Minimizing risk models in Markov decision processes with policies depending on target values [J]. J Math Anal Appl，1999，231(1)：47−67.

[35] ROCKAFELLAR R，URYASEV S. Optimization of conditional value-at-risk [J]. J Risk，2000，2：21−42.

[36] ARTZNER P，DELBAEN F，EBER J M，et al. Coherent measures of risk [J]. Math Finance，1999，9(3)：203−228.

[37] BORKAR V，JAIN R. Risk-constrained Markov decision processes [J]. IEEE Trans of Automat Control，2014，59(9)：2574−2579.

[38] BÄUERLE N，OTT J. Markov decision processes with average-value-at-risk criteria [J]. Math Methods Oper Res，2011，74(3)：361−379.

[39] HASKELL W，JAIN R. A convex analytic approach to risk-aware Markov decision processes [J]. SIAM J Control Optim，2014，53(3)：1569−1598.

[40] PRASHANTH L A. Policy gradients for CVaR-constrained MDPs [C]// In Proceedings of the International Conference on Algorithmic Learning Theory，2014：155−169.

[41] CHOW Y，TAMAR A，MANNOR S，et al. Risk-sensitive and robust decision-making：A CVaR optimization approach [C]// In Proceedings of the Neural Information Processing Systems，2015：1522−1530.

［42］ ZHU B，WEN B，JI S，et al. Coordinating a dual-channel supply chain with conditional value-at-risk under uncertainties of yield and demand［J］. Comput Ind Eng，2020，139：106181.

［43］ ROUSTAI M，RAYATI M，SHEIKHI A，RANJBAR A. A scenario−based optimization of smart energy hub operation in a stochastic environment using conditional-value-at-risk［J］. Sustain Cities Soc，2018，39：309−316.

［44］ HOSSEINI S D，VERMA M. Conditional value-at-risk（CVaR）methodology to optimal train configuration and routing of rail hazmat shipments［J］. Transportation Res Part B，2018，110：79−103.

［45］ HE F，CHAUSSALET T，QU R. Controlling understaffing with conditional Value-at-Risk constraint for an integrated nurse scheduling problem under patient demand uncertainty［J］. Oper Res Perspect，2019，6：100119.

［46］ FILIPPI C，GUASTAROBA G，SPERANZA M G. Conditional value-at-risk beyond finance：A survey［J］. Int Trans Oper Res，2020，27(3)：1277−1319.

［47］ BORKAR V S. Q-learning for risk-sensitive control［J］. Math Oper Res，2002，27(2)：294−311.

［48］ SHEN Y，TOBIA M J，SOMMER T，et al. Risk-sensitive reinforcement learning［J］. Neural Comput，2014，26(7)：1298−1328.

［49］ GARCÍA J，FERNÁNDEZ F. A comprehensive survey on safe reinforcement learning［J］. J Mach Learn Res，2015，16(1)：1437−1480.

［50］ HUANG W，HASKELL W. B. Risk-aware Q-learning for Markov decision processes［C］// In Proceedings of the IEEE Conference on Decision and Control，2017：4928−4933.

［51］ CHOW Y，GHAVAMZADEH M，JANSON L，et al. Risk-constrained reinforcement learning with percentile risk criteria［J］. J Mach Learn Res，2017，18(1)：6070−6120.

［52］ NORTON M，KHOKHLOV V，URYASEV S. Calculating CVaR and bPOE for common probability distributions with application to portfolio optimization and density estimation［J］. Ann Oper Res，2021，299(1)：1281−1315.

［53］ MA S，YU J Y. State-augmentation transformations for risk-sensitive reinforcement learning［C］// In Proceedings of the AAAI Conference on Artificial Intelligence，2019：4512−4519.

［54］ MA S，YU J Y. Transition-based versus state-based reward functions for MDPs with value-at-risk［C］// In Proceedings of the Annual Allerton Conference on Communication，Control，and Computing，2017：974−981.

［55］ MA S，YU J Y. Variance-based risk estimations in Markov processes via transformation with state lumping［C］// In Proceedings of the IEEE International Conference on Systems，Man and Cybernetics，2019：958−963.

［56］ DURBIN J. Distribution theory for tests based on the sample distribution function［M］. Philadelphia：Society for Industrial and Applied Mathematics，1973.

（责任编辑　冯兆永）