Processos de decisão de Markov com sensibilidade a risco com função de utilidade exponencial: Uma revisão sistemática da literatura

Alternative Title: Risk-sensitive Markov decision processes with exponential utility function: A systematic review of the literature

Elthon Manhas de Freitas
Universidade de São Paulo
elthon@usp.br

Karina Valdivia Delgado
Universidade de São Paulo
kvd@usp.br

Valdinei Freire da Silva
Universidade de São Paulo
valdinei.freire@usp.br

RESUMO

Os processos de decisão de Markov (em inglês Markov Decision Process - MDP) têm sido usados com muita eficiência para resolução de problemas de tomada de decisão sequencial. Existem problemas em que lidar com os riscos do ambiente para obter um resultado confiável é mais importante do que maximizar o retorno médio esperado. MDPs que lidam com esse tipo de problemas são chamados de processos de decisão de Markov sensíveis a risco (em inglês Risk-Sensitive Markov Decision Process - RSMDP). Esta revisão sistemática da literatura tem por objetivo identificar os resultados teóricos e os algoritmos propostos para resolver problemas de RSMDP que tenham função utilidade exponencial, avaliando as suas principais características, semelhanças, particularidades e diferenças de modo a permitir ao leitor o conhecimento desta ferramenta de tomada de decisão sequencial para problemas sensíveis ao risco.

Palavras-Chave

Processos de decisão de Markov, sensível a risco, averso a risco, planejamento probabilístico, utilidade exponencial

ABSTRACT

Markov Decision Process (MDP) has been used very efficiently to solve sequential decision-making problems. There are problems in which dealing with the risks of the environment to obtain a reliable result is more important than maximizing the expected average return. MDPs that deal with this type of problem are called risk-sensitive Markov decision processes (RSMDP). This systematic review of the literature aims to identify the theoretical results and proposed algorithms to solve RSMDP problems that have an exponential utility function, evaluating their main characteristics, similarities, particularities and differences in order to allow the reader the knowledge of this tool of decision making for risk sensitive problems.

SBSI 2017 June 5^{th} – 8^{th}, 2017, Lavras, Minas Gerais, Brazil

CCS Concepts

•Theory of computation → Markov decision processes; •Computing methodologies → Dynamic programming for Markov decision processes;

Keywords

Markov Decision Process, risk sensitive, risk averse, probabilistic planning, exponential utility

1. INTRODUÇÃO

Processos de decisão de Markov é o nome dado a uma série de modelos estocásticos robustos utilizados para tomada de decisão sequencial baseada apenas nas informações do estado atual do ambiente [35]. Este modelo é classificado como estocástico pois não há controle de todas as variáveis presentes no ambiente em que o agente tomador de decisão está inserido. Essa incerteza pode ocorrer por diversos fatores como imprecisão durante a execução de uma ação ou ainda pela existência de outros agentes que podem estar constantemente interferindo no sistema.

Existem muitos trabalhos realizados sobre a obra original de Andrei Andreyevich Markov [1]. A maioria destes trabalhos avalia o efeito probabilístico na execução de ações no sistema e tem como objetivo gerar uma política de decisões, a ser seguida por um agente executor, de modo a maximizar o retorno do sistema ou diminuir o seu custo [11]. Ao longo de sucessivas execuções, o sistema tende gerar um valor médio muito próximo ao valor esperado devido ao efeito probabilístico.

Há problemas da vida real que só podem ser executados apenas uma vez. Por exemplo, um veículo com navegação autônoma tem que considerar que cada trajeto é único e não irá se repetir, logo o processo não poderá simplesmente reiniciar em caso de falha. Realizar um transplante de coração é outro exemplo em que aumentar as chances de sucesso se faz tão importante que aumentos de custo são aceitos quase sem questionamento. Outros problemas tem uma duração tão longa que não podem ser executados várias vezes, como realizar uma viagem a Marte, investir para a aposentadoria de uma vida [27] ou implantar um grande projeto empresa-

rial. Estes são alguns exemplos em que mitigar, evitar e até eliminar os riscos do ambiente é muito mais importante do que maximizar o retorno esperado. Estes são problemas de máxima aversão ao risco e para este extremo existe uma otimização denominada *minmax*, que minimiza o risco a todo custo.

Exceto pelos casos extremos, nosso cotidiano é pautado por tolerância ao risco. Um piloto de corrida está disposto a forçar um pouco mais o carro na última volta para conseguir melhorar sua posição, assim como nós estamos dispostos a conhecer um lugar novo em busca de novas experiências. Devemos considerar também que pessoas diferentes tem níveis de aceitação diferentes de risco, até mesmo momentos distintos podem nos afetar e nos tornar mais propensos ou mais aversos ao risco.

Para lidar com esse tipo de problemas, há uma pequena parcela de trabalhos que avaliam a sensibilidade e a tolerância ao risco e de alguma forma consideram estes parâmetros em seus modelos, os chamados Processos de Decisão de Markov Sensíveis a Risco (em inglês Risk Sensitive Markov Decision Process - RSMDP) [17].

Esta revisão sistemática da literatura tem por objetivo identificar os resultados teóricos e os algoritmos propostos para resolver problemas de RSMDP, avaliando as suas principais características, semelhanças, particularidades e diferenças de modo a permitir ao leitor o conhecimento desta ferramenta de tomada de decisão sequencial para problemas sensíveis ao risco.

O restante deste texto está organizado da seguinte forma. A Seção 2 apresenta os conceitos fundamentais de MDP e RSMDP. A Seção 3 apresenta o método utilizado nesta RS. A Seção 4 apresenta uma discussão dos resultados obtidos. Por fim, a Seção 5 apresenta as conclusões.

2. CONCEITOS FUNDAMENTAIS

Um processo de decisão de Markov (MDP) [35] é uma tupla $M = (S, A, p, r)$, em que: S é um conjunto finito de estados observáveis; A é um conjunto finito de ações; $p(s'|s, a)$ é a função probabilística de transição que descreve os efeitos da execução de uma ação $a \in A$ em um estado $s \in S$ resultando em um estado $s' \in S$; e $r(s, a)$ é a função recompensa (ou custo) por executar uma ação $a \in A$ em um estado $s \in S$.

O agente executa as ações em passos discretos no tempo. A cada ação executada, o estado do sistema é alterado segundo a função de transição p e produz uma recompensa conforme a função r. O objetivo do agente é acumular recompensas positivas e evitar recompensas negativas (custo) durante um horizonte.

O horizonte é o número de passos que o agente tem para agir, podendo ser: finito, infinito ou indeterminado. O horizonte finito estipula uma quantidade finita de passos T, enquanto no horizonte infinito considera-se que o processo nunca acaba. Finalmente, no horizonte indeterminado considera-se a possibilidade de o processo acabar ao atingir algum estado final específico.

Uma política π é um mapeamento de um espaço de estados para um espaço de ações ($\pi : S \rightarrow A$) que representa quais ações devem ser executadas em cada estado considerando um critério de otimização. A solução para um MDP tradicional é uma política ótima (π^*), considerada neutra ao risco, que maximize a soma total das recompensas esperadas (horizonte finito ou indeterminado) ou que maximiza a recompensa média esperada (horizonte infinito) [35]. Os algoritmos tradicionais para encontrar uma política ótima são: Iteração de Valor (IV) e Iteração de Política (IP).

Variações na formulação de um MDP dizem respeito à natureza dos espaços de estados e ações. Além de finito, o espaço de estado pode ser denumerável, equivalente ao conjunto dos naturais, ou contínuo com dimensão finita, usualmente um conjunto de Borel. O espaço de ações também pode ser contínuo com dimensão finita, usualmente um espaço compacto.

2.1 Tratamento de risco em MDP

Devido à natureza probabilística do MDP, o risco é inerente. Ao longo dos anos, muitos trabalhos tem apresentado diversos critérios de otimização para lidar com esse risco. Em [14] os autores classificam esses critérios em quatro grupos:

- *Critério do pior caso.* A política considerada ótima é aquela que maximiza o retorno associado ao cenário do pior caso, mesmo que o caso seja altamente improvável [14]. Este grupo é conhecido pela máxima aversão às perdas, e o algoritmo mais utilizado é o algoritmo *minmax* [16, 32].

- *Critério sensível ao risco.* Tenta balancear risco e retorno e é caracterizado por possuir um parâmetro que permite controlar a sensibilidade do risco. Nestas abordagens o critério de otimização é transformado: (i) em uma função utilidade exponencial [17] (detalhado na Seção 2.2), (ii) em uma combinação linear de risco, representado pela variância, e retorno, representado pela recompensa esperada [37], ou (iii) na probabilidade de entrar em um estado de erro [15].

- *Critério com restrições.* O MDP possui restrições que a política deve seguir. Nestes casos, a soma total das recompensas esperadas deve ser maximizada, desde que não viole as restrições do MDP [18, 28]. Essas restrições representam alguma ideia de risco, por exemplo, um estado de erro não é visitado com probabilidade menor que α.

- *Outros critérios.* Outras abordagens baseadas em engenharia financeira tem sido utilizadas para gerenciar o risco do MDP, entre elas a utilização do *r-squared*, *value-at-risk* (*VaR*) [19, 24, 25] e a densidade do retorno [30, 31].

2.2 MDP sensível a risco com função de utilidade exponencial

Com base na Teoria da Utilidade Esperada, em [17] é proposta a função utilidade exponencial para modelar atitude ao risco do agente. A função utilidade apresenta algumas propriedades interessantes: (i) considera-se um parâmetro arbitrário β que modela a atitude ao risco do agente; (ii) assim como na função identidade, um acréscimo de uma recompensa constante em toda história não altera a ordem relativa entre as políticas; e (iii) é possível construir uma equação de otimalidade, que é o caminho para definição dos algoritmos de IV e IP.

No caso de um horizonte finito ou indeterminado considera-

se a seguinte avaliação para uma política π:

$$-\beta^{-1} \log \mathrm{E}_{\pi} \left[\exp \left(-\beta \sum_{t=0}^{\infty} r_t \right) \right];$$

e no caso de horizonte infinito considera-se:

$$\lim_{N \to \infty} -\frac{1}{N} \beta^{-1} \log \mathrm{E}_{\pi} \left[\exp \left(-\beta \sum_{t=0}^{N-1} r_t \right) \right];$$

onde E_{π} representa a esperança quando a política π é executada, exp() é a função exponencial e o agente é averso ao risco se $\beta > 0$, propenso ao risco se $\beta < 0$ e neutro ao risco se $\beta \to 0$.

3. MÉTODO

Uma revisão sistemática é considerada um estudo secundário que têm nos estudos primários, sobre determinado tema, sua fonte de dados. A RS é um trabalho reprodutível e que visa minimizar possíveis direções indesejadas uma vez que não depende somente da experiência do especialista envolvido. Em [20] são apresentadas as três etapas principais para o desenvolvimento de uma RS e que são usadas neste trabalho: (i) Planejamento: em que o protocolo da revisão é gerado; (ii) condução: em que é feita uma seleção dos estudos considerando o protocolo desenvolvido; e (iii) relatório: em que é realizada a análise dos resultados.

3.1 Planejamento

3.1.1 Questões de pesquisa

As questões de pesquisa foram elaboradas de forma a atingir o objetivo desta RS. As questões de pesquisa são:

Q-1 Quais os resultados teóricos e/ou a abordagem proposta para resolução do RSMDP?

A questão Q-1 visa identificar a tendência das abordagens utilizadas como solução para RSMDP.

Q-2 Quais as características do RSMDP?

- Q-2.1 O RSMDP é de tempo contínuo ou discreto?
- Q-2.2 Qual a classificação do horizonte de tempo (finito/infinito/indeterminado)?
- Q-2.3 O RSMDP é observável ou parcialmente observável?
- Q-2.4 A abordagem proposta trata aversão ou propensão ao risco? Ou ambos?
- Q-2.5 Qual o critério de avaliação do RSMDP? Soma das recompensas esperadas ou média das recompensas?
- Q-2.6 Qual é o tipo do espaço de estados?
- Q-2.7 Qual é o tipo do espaço de ações?
- Q-2.8 Qual o tipo de recompensa ou custo utilizado?

O objetivo da Q-2 é classificar os problemas sensíveis aos riscos que podem ser resolvidos pelas abordagens das soluções. Para isso, Q-2 foi decomposta em questões mais específicas, uma para cada característica estrutural do problema.

Em Q-2.1 é considerado o tempo do RSMDP. A contagem de tempo mais tradicional em RSMDP é o tempo discreto. Neste tipo de contagem, tanto a avaliação da política quanto cada passo realizado pelo agente do RSMDP acontece em uma unidade de tempo discreta, ou seja, $t \in \mathbb{N}$. A contagem de tempo contínua avalia os valores da política de modo contínuo ao longo do tempo, com infinitas divisões de tempo entre cada iteração, ou seja, $t \in \mathbb{R}$.

Com Q-2.2 pretendemos classificar os RSMDPs pelo horizonte que pode ser finito, infinito ou indeterminado como visto na Seção 2.

Q-2.3 permite a classificação dos RSMDPs pela observação. Em um RSMDP parcialmente observável o agente não conhece com exatidão o estado do sistema, sendo necessário trabalhar com estados de crença. Caso contrário o RSMDP é totalmente observável.

Q-2.4 permite classificar os RSMPDs de acordo com a classe de risco que é tratada.

Em Q-2.5 o critério de desempenho da função valor do RSMDP será classificado. Existem funções valor baseadas na soma das recompensas esperadas, como visto na Seção 2 e funções valor baseadas na média das recompensas.

Q-2.6 e Q-2.7 permitem classificar os RSMDPs de acordo com o tipo do espaço de estados que pode ser finito, denumerável ou contínuo. O espaço é denumerável se tem a mesma cardinalidade que $\mathbb{Z}^+$, portanto se é denumerável então é infinito. Quando o espaço é contínuo, usualmente impõem-se alguma restrição topológica, por exemplo, espaço Borel e compacto.

Em Q-2.8 será classificado o RSMDP pela recompensa ou custos utilizados.

Q-3 Quais as características dos experimentos realizados?

- Q-3.1 Foram realizados testes ou o artigo apresenta apenas estudos teóricos?
- Q-3.2 Qual a quantidade de estados dos problemas resolvidos?
- Q-3.3 Qual o domínio dos testes?
- Q-3.4 O domínio usado é artificial ou real?

O objetivo de Q-3 é identificar como os experimentos foram realizados, se a solução consegue resolver RSMDPs grandes e se foram aplicados em dados reais ou artificiais.

3.1.2 Seleção de fontes e string de busca

Foi utilizado o motor de busca Scopus [1]. Este motor de busca foi escolhido pois indexa outras fontes de dados entre elas IEEE Xplore[2], Science Direct[3], ACM Digital Library[4] e Engineering Village[5]. Uma string de pesquisa foi desenvolvida (Tabela 1) considerando as palavras da string nos campos título, resumo e palavras chave dos artigos.

3.1.3 Critérios de Inclusão e Exclusão

Para seleção dos estudos foram criados critérios de inclusão (I) e exclusão (E). Esses critérios tem como objetivo garantir um nível mínimo de qualidade dos estudos incluídos. Os estudos devem ser primários e não devem fugir do objetivo desta RS. Os critérios estão descritos na Tabela 2.

[1]https://www.scopus.com
[2]http://ieeexplore.ieee.org
[3]http://www.sciencedirect.com
[4]http://dl.acm.org
[5]http://www.engineeringvillage.com

Table 1: String de busca.

(TITLE-ABS-KEY ("Markov* decision process*") OR
TITLE-ABS-KEY ("MDP") OR
TITLE-ABS-KEY ("Probabilistic Planning")) AND
(TITLE-ABS-KEY ("Risk Sensitiv*") OR
TITLE-ABS-KEY ("Risk Averse") OR
TITLE-ABS-KEY ("Certainty equivalent") OR
TITLE-ABS-KEY ("exponential utility") OR
TITLE-ABS-KEY ("risk measure"))

Table 2: Critérios de inclusão e exclusão.

(I-1)	Estudos contendo simultaneamente as palavras-chaves Markov Decision Process e Risk Sensitive, ou qualquer um de seus sinônimos, no título, no resumo ou nas palavras chaves.
(I-2)	Estudos publicados em workshop, conferência ou periódico.
(E-1)	Estudos não dedicados exclusivamente a processos de decisão de Markov com tratamento de sensibilidade a risco.
(E-2)	Trabalhos que não estejam em língua inglesa.
(E-3)	Trabalhos em áreas de pesquisa que não sejam matemática, engenharia ou computação.
(E-4)	Trabalhos que não apresentem resultados teóricos ou algoritmos.
(E-5)	Trabalhos que não tenham sido avaliados por pares.
(E-6)	Trabalhos que tratem apenas problemas do tipo bandit, ou seja, problemas com um único estado.
(E-7)	Estudos que tratem apenas problemas cuja probabilidade de transição é desconhecida e precisa ser aprendida, ou seja, trabalhos com aprendizado por reforço.
(E-8)	Trabalhos cujo foco principal seja a aplicação de RSMDP.
(E-9)	Estudos secundários ou terciários.
(E-10)	Trabalhos anteriores a 2000.
(E-11)	Estudos não disponíveis integralmente na Web, ou não acessíveis de forma gratuita via a instituição dos autores desta RS.
(E-12)	Estudos incompletos ou que não descrevem de forma detalhada os resultados teóricos e/ou algoritmo utilizado como solução.
(E-13)	Utiliza critério de função utilidade que não seja exponencial (p.e. pior caso, combinação linear risco e retorno, critério com restrições, outros critérios como VaR).

3.2 Condução

A pesquisa realizada no motor de busca Scopus realizada em 4 de fevereiro de 2017 retornou 108 trabalhos. Desse total, 23 foram excluídos pelos critérios E-3, E-5, E-9 e E-10 através das ferramentas de filtros do motor de busca, restando 85 estudos para serem analisados. Para auxiliar na condução desta RS foi utilizado o Software StArt *(State of the Art through Systematic Review)*[6] que identificou 2 artigos duplicados. Os demais critérios foram aplicados manualmente em cada um dos 83 estudos restantes, foi analisado o título, resumo e palavras chaves, e no caso de dúvida o estudo foi analisado por completo. No fim dessa etapa foram selecionados 19 estudos para o desenvolvimento da RS.

4. RESULTADOS E DISCUSSÃO

Nesta seção são analisados os 19 estudos selecionados nesta RS. Na Tabela 3 mostramos o identificador (ID) considerando o tipo de veículo (P–Periódico e C–Conferência), o título e o ano de publicação desses estudos. Veja que em 2016 foram publicados três trabalhos, sendo que dois deles consideram tempo contínuo.

[6]Disponível em: http://lapes.dc.ufscar.br/tools/start_tool

4.1 Contribuições: Q-1

A Tabela 4 apresenta de forma resumida as contribuições dos trabalhos selecionados. A maioria dos trabalhos apresenta contribuição teórica sobre o modelo, conforme indicado na segunda coluna. Outros trabalhos demonstram a convergência ou fazem uma estimativa do erro do algoritmo proposto. Nestes casos temos indicado qual o nome do algoritmo. A quarta coluna da tabela tem como objetivo indicar se o trabalho selecionado apresenta experimentos, informação importante para novos estudos na área para saber com quem comparar resultados computacionais ou para reproduzir os resultados obtidos.

A seguir apresentamos as contribuições de cada um dos trabalhos selecionados utilizando o identificador do trabalho apresentado na Tabela 3. Para facilitar a descrição dividimos os trabalhos em três grandes grupos: RSMDP de tempo discreto, RSMDP de tempo contínuo e uma subclasse de RSMDPs. O RSMDP de tempo discreto ainda é sub-divido em subgrupos dependendo do horizonte e do critério de avaliação usado.

4.1.1 RSMDP de tempo discreto

Em P-16 são considerados tanto problemas de horizonte finito (soma total de recompensa) como de horizonte infinito (média de recompensas), com desconto e sem desconto, e cuja função utilidade é crescente. Esse problemas têm como caso especial os RSMDPs. Os autores mostram que essa classe de problemas pode ser resolvida por MDPs clássicos com o espaço de estados estendido e fornecem as condições que garantem a existência de políticas ótimas. Além disso, para problemas de horizonte infinito, os autores apresentam um algoritmo de iteração de valor e demonstram a convergência do método de iteração de política. Os outros trabalhos são organizados a seguir segundo seu horizonte.

Horizonte indeterminado e Recompensa Total.

Em P-1, considerando recompensa não negativa, é demonstrado que quando o conjunto de estados é finito e o conjunto de ações é contínuo, existe uma política estacionária *quasi* ótima independente da atitude ao risco. Além disso, é demonstrado que quando o conjunto de estados é denumerável e o conjunto de ações é contínuo, existe uma política estacionária *quasi* ótima para a atitude aversa ao risco.

Em P-3, considerando apenas aversão ao risco, custo estritamente positivo e assumindo a existência de uma política estacionária λ-factível, os autores demonstram a existência de uma função de custo ótima a qual pode ser alcançada por uma política ótima estacionária. Além disso, em P-3 é demonstrada a convergência do algoritmo de iteração de valor e de política para essa classe de problemas.

Em C-7 e C-8, problemas de planejamento probabilístico sensível ao risco com funções de utilidade não linear são analisados, sendo o RSMDP um caso especial. Nesses estudos são apresentadas as condições que garantem que a utilidade esperada ótima existe e é finita considerando recompensa positiva e/ou negativa.

[7]Apenas tem experimentos com um problema com restrições.

Table 3: Estudos selecionados nesta RS

ID	Ref.	Título	Ano
P-1	[8]	Nearly optimal policies in risk-sensitive positive dynamic programming on discrete spaces	2000
C-2	[6]	Markov decision processes with risk-sensitive criteria: Dynamic programming operators and discounted stochastic games	2001
P-3	[34]	On terminating Markov decision processes with a risk-averse objective function	2001
P-4	[3]	Risk-sensitive optimal control for Markov decision processes with monotone cost	2002
P-5	[5]	Solution to the risk-sensitive average cost optimality equation in a class of Markov decision processes with finite state space	2003
P-6	[9]	The value iteration algorithm in risk-sensitive average Markov decision chains with finite state space	2003
C-7	[22]	Existence and finiteness conditions for risk-sensitive planning: First results	2004
C-8	[23]	Existence and finiteness conditions for risk-sensitive planning: Results and conjectures	2005
P-9	[12]	A turnpike theorem for a risk-sensitive Markov decision process with stopping	2006
P-10	[10]	The discounted method and equivalence of average criteria for risk-sensitive Markov decision processes on Borel spaces	2010
C-11	[13]	A unifying framework for linearly solvable control	2011
P-12	[7]	Discounted approximations for risk-sensitive average criteria in: Markov decision chains with finite state space	2011
C-13	[27]	Risk aversion in Markov decision processes via near-optimal Chernoff bounds	2012
C-14	[33]	Robustness and risk-sensitivity in Markov decision processes	2012
C-15	[26]	Shortest stochastic path with risk sensitive evaluation	2013
P-16	[4]	More risk-sensitive Markov decision processes	2014
P-17	[38]	Continuous-time Markov decision processes with risk-sensitive finite-horizon cost criterion	2016
P-18	[39]	Continuous-time Markov decision processes under the risk-sensitive average cost criterion	2016
C-19	[21]	Finite horizon risk sensitive MDP and linear programming	2016

Table 4: Contribuições dos estudos selecionados nesta RS

ID	Contribuição teórica sobre o modelo	Convergência ou estimação de erro do Algoritmo	Experimentos
P-1	Sim	Não	Não
C-2	Sim (short paper)	Não	Não
P-3	Sim	Sim: IV e IP	Sim
P-4	Sim	Sim: IV e IP	Não
P-5	Sim	Não	Não
P-6	Não	Sim: IV	Não
C-7	Sim	Não	Não
C-8	Sim	Não	Não
P-9	Sim	Não	Não
P-10	Sim	Não	Não
C-11	Sim	Não	Sim
P-12	Sim	Não	Não
C-13	Sim	Sim: *Near optimal Chernoff bound algorithm*	Sim
C-14	Sim	Não	Não
C-15	Sim	Não	Sim
P-16	Sim	Sim para horizonte infinito: IV e IP	Sim
P-17	Sim	Sim: Método de Iteração	Sim
P-18	Sim	Não	Não
C-19	Não	Sim: Programação Linear	Não [7]

Em P-9 são analisados RSMDPs com tempo discreto, conjuntos de estados e ações finitos. Considera-se tanto recompensa negativa quanto recompensa positiva, assim como propensão e aversão ao risco. Os autores mostram como encontrar uma política estacionária e condições para verificar se essa política é ótima.

Em C-15 são mostradas as diferenças e semelhanças entre MDP e RSMDP. Além disso, é explicada a relação entre o fator de desconto e a propensão ao risco em problemas com custo constante.

Horizonte infinito e Recompensa Média.

C-2 é um estudo curto em que são estudados problemas com o espaço de estados denumerável. Nesse estudo é demonstrado que existe uma solução para a equação de otimalidade que considera a média do custo, apenas quando o parâmetro de risco é suficientemente pequeno.

Em P-4 as condições que garantem a existência de uma política ótima são estabelecidas para RSMDPs aversos ao risco, com o espaço de estados denumerável e ações finitas. Além disso, os autores apresentam os algoritmos de iteração de valor e iteração de política com a suposição de que todas as políticas geram cadeias irredutíveis.

Problemas com o espaço de estados finito são estudados em P-5 e P-6. Em P-5 é considerado que o problema tem uma estrutura de comunicação, i.e., que para qualquer par de estados x e y, existe uma política tal que a probabilidade de alcançar y a partir de x é positiva. Para esse tipo de problemas, os autores demonstram a existência de soluções para equação de otimalidade sempre que o parâmetro de risco não exceda determinado valor positivo. Supondo que existe uma solução para a equação de otimalidade, em P-6 é demonstrado que o algoritmo de iteração de valor pode ser usado para aproximar o custo médio com um erro limitado e para encontrar uma política ótima estacionária em um número finito de passos.

Em P-10 é considerado apenas aversão ao risco e problemas com os espaços de estados e ações contínuos. Os autores demonstram que existem soluções para a inequação de otimalidade. Além disso, é demonstrado que considerando algumas condições no custo, as funções de valor ótimo correspondentes ao limite superior e inferior do critério de custo médio coincidem em um determinado subconjunto do espaço de estados. Finalmente, os autores mostram que equações baseadas em fator de desconto podem ser utilizadas para aproximar arbitrariamente equações com base em avaliação de média.

Em P-12 são estudados problemas com o espaço de estados finito e o conjunto de ações contínuo. Os autores demonstram que considerando as condições padrão de continuidade-compacidade, as aproximações descontadas convergem para a função valor ótima. Além disso, é demonstrado que o limite superior e inferior do critério de custo médio tem a mesma função valor ótima. Diferente de P-5, essa demons-

tração independe da estrutura de comunicação do problema e é válida para qualquer valor do parâmetro de risco.

Horizonte finito e Recompensa Total.
Em C-14 é demonstrado que maximizar a utilidade exponencial esperada em um RSMDP é equivalente a um MDP robusto que maximiza o critério do pior caso com uma penalidade para o desvio dos parâmetros incertos de seus valores nominais.

Em C-19 são propostos dois modelos de Programação Linear, um primal e um dual para resolver RSMDPs. A formulação em programas lineares permite inserir restrições que lidam com risco mais facilmente.

4.1.2 RSMDP de tempo contínuo

RSMDPs com tempo contínuo são estudados em P-17 e P-18. Em P-17 são considerados problemas de horizonte finito com o espaço de estados denumerável e espaço de ações contínuo. Nesse trabalho são dadas as condições que garantem a existência de uma única solução para a equação de otimalidade e a existência de uma política Markoviana determinística ótima. Além disso, é apresentado um método iterativo para calcular o valor e uma política ótima com garantias de erro. Uma vez que o espaço de estados é denumerável é construída uma sequência de modelos de controle com um número finito de estados para que esse método proposto seja numericamente tratável. Em P-18 é usada a média como critério de avaliação para problemas de horizonte infinito, com o espaço de estados finito e com o espaço de ações contínuo. Nesse artigo também são estabelecidas as condições para garantir a existência de uma solução para a equação de otimilidade e a existência de uma política estacionária determinística ótima.

4.1.3 Função Exponencial e outros modelos

Em C-11 os MDPs linearmente solucionáveis (em inglês Linearly Solvable Markov Decision Processes – LMDPs) de tempo discreto e de horizonte finito, infinito e indeterminado são estendidos para tratar risco. LMDP é uma classe de problemas de controle que são mais tratáveis que MDP em que é permitido que o controlador escolha diretamente a densidade de transição.

Em C-13 é considerado um problema com espaço de estados e ações finito, de horizonte finito e o critério de avaliação de soma esperada. Nesse trabalho é proposto um novo critério de otimização para tratar o risco chamada de função *Chernoff*, que envolve a função exponencial como parte da medida. Além disso, é apresentado um algoritmo para resolver esse problema.

4.2 Características do RSMDP: Q-2

Ao avaliar como as diversas abordagens tratam o risco, identificamos quais características do RSMDP são tratadas por cada solução. O levantamento realizado está resumido na Tabela 5.

Contagem de tempo.
Quase a totalidade dos trabalhos explorados trabalham com a contagem de tempo discreta. A exceção ocorre nos trabalhos P-17 e P-18 em que o tempo do MDP é contínuo.

Horizonte de tempo.
Geralmente as soluções propostas trabalham com apenas um tipo de horizonte de tempo, entretanto o artigo P-16 trata tanto horizonte finito quanto infinito. O trabalho C-11 merece destaque por tratar os três tipos de horizonte de tempo através de um arcabouço unificado, porém só trata um tipo específico de RSMDP, o LMDP com risco.

Observação do RSMDP.
Todos os trabalhos selecionados tratam apenas de RSMDP totalmente observáveis. Como nenhum trabalho selecionado trata RSMDPs parcialmente observáveis, vemos esta como um importante lacuna a ser explorada.

Tipo de risco.
A maioria dos trabalhos possibilita que problemas aversos e propensos ao risco sejam tratados com suas abordagens. Entre os estudos selecionados que não exploraram a propensão a risco estão P-3, P-4, P-10 e C-14. Desta forma, a propensão nestes trabalhos ainda pode ser explorada em projetos futuros.

Critério de avaliação.
A avaliação da política ocorre pela média ou soma das recompensas ou custos. Os trabalhos C-11 e P-16 oferecem propostas para trabalhar com esses dois critérios.

Espaço de estados e ações.
Tanto o espaço de estados dos problemas quanto o espaço de ações concentram-se em conjuntos finitos. Há um pequeno número de abordagens que consideram espaço denumerável ou contínuo, tanto para estados quanto para ações. Espaço compacto é utilizado apenas no conjunto de ações dos trabalhos P-3 e P-12.

4.3 Características dos experimentos realizados: Q-3

Como observamos na Tabela 4, apenas seis dos trabalhos selecionados apresentam experimentos. Algumas características destes experimentos estão resumidos na Tabela 6.

A quantidade de estados dos experimentos merece destaque. Vemos que mesmo problemas com maior número de estados puderam ser processados nos experimentos de C-13 e P-17. Em C-13, além de um problema artificial, Moldovan e Abbeel tentaram processar um problema real que possuía 124791 estados. Devido às limitações computacionais, em seus experimentos os autores reduziram o número de estados para 500, trabalhando assim com um sub-conjunto de estados do domínio real.

5. DISCUSSÃO E CONCLUSÃO

Com base nos trabalhos pesquisados, é possível notar como o tratamento de risco ganhou espaço na tomada de decisão sequencial. A utilização de função utilidade exponencial tem se mostrado muito eficiente e versátil para tratar risco, possibilitando sua utilização em diversos tipos de problemas teóricos, seja ele de aversão ou de propensão a risco.

Uma limitação notável dos estudos selecionados é a possibilidade de lidar com um número grande de estados no RSMDP. Como a função utilidade é exponencial, a resolução de problemas com muitos estados tende a consumir muito tempo e a gerar erros de precisão. Nos trabalhos selecionados, o C-13 resolveu o problema com o maior número de estados - 500 - enquanto que MDPs convencionais

Table 5: Características do MDP

ID	Tempo		Horizonte			Observação		Tipo de Risco		Critério de Avaliação		Espaço de Estados	Espaço de Ações	Tipo de recompensa
	Contínuo	Discreto	Finito	Infinito	Indeterminado	Parcial	Total	Averso	Propenso	Soma	Média			
P-1		X			X		X	X	X	X		Denumerável	Contínuo	Não-negativo
C-2		X		X			X	X	X		X	Denumerável	Contínuo	Não-positivo
P-3		X			X		X	X		X		Finito	Contínuo	Negativo
P-4		X		X			X	X			X	Denumerável	Finito	Negativo
P-5		X		X			X	X	X		X	Finito	Contínuo	Negativo
P-6		X		X			X	X	X		X	Finito	Contínuo	Negativo
C-7		X			X		X	X	X	X		Finito	Finito	Real
C-8		X			X		X	X	X	X		Finito	Finito	Real
P-9		X			X		X	X	X	X		Finito	Finito	Real
P-10		X		X			X	X			X	Contínuo	Contínuo	Não-Positivo
C-11		X	X	X	X		X	X	X	X	X	Finito	Finito	Negativo
P-12		X		X			X	X	X		X	Finito	Contínuo	Não-Positivo
C-13		X	X				X	X	X	X		Finito	Finito	Não-Positivo
C-14		X	X				X	X		X		Finito	Finito	Não-Positivo
C-15		X			X		X	X	X	X		Finito	Finito	Não-Positivo
P-16		X	X	X			X	X	X	X	X	Contínuo	Contínuo	Negativo
P-17	X		X				X	X	X	X		Denumerável	Contínuo	Não-Positivo
P-18	X			X			X	X	X		X	Finito	Contínuo	Não-Positivo
C-19		X	X				X	X	X	X		Finito	Finito	Não-Positivo

Table 6: Estudos selecionados nesta RS organizados pelas características dos experimentos realizados

ID	Quantidade de estados	Domínio	Real(R) ou Artificial(A)
P-3	4 - 16	Political Commitee	A
C-11	91	Terrain	A
C-13	120 e 500	Grid / Air travel plan	A/R
C-15	12	Drivers Licence	A
P-16	–	Casino Game	A
P-17	50, 75 e 100	Birth and Death	A

conseguem resolver problemas com milhões de estados. Trabalhos recentes [2, 36] que utilizam programação paralela e processamento distribuído se mostraram capazes de aumentar consideravelmente o tamanho dos problemas solucionados. Estas mesmas técnicas poderiam ser aplicadas nos RSMDPs.

Outro fator que chama a atenção é a ausência de trabalhos de RSMDP que tratem problemas parcialmente observáveis, o que teve muita expansão no universo de processos de tomada de decisão sequencial [29]. Os MDPs parcialmente observáveis são muito mais próximos às aplicações reais e soluções que integrem gerenciamento de risco e observabilidade parcial devem ser demandadas em um futuro próximo.

Referências

[1] G. P. Basharin, A. N. Langville, and V. A. Naumov. The life and work of a.a. Markov. *Linear Algebra and its Applications*, 386:3–26, 2004.

[2] D. P. Bertsekas and H. Yu. Distributed asynchronous policy iteration in dynamic programming. In *Conference on Communication, Control, and Computing*, pages 1368–1375. IEEE, 2010.

[3] V. Borkar and S. Meyn. Risk-sensitive optimal control for Markov decision processes with monotone cost. *Mathematics of Operations Research*, 27(1):192–209, 2002.

[4] N. Bäuerle and U. Rieder. More risk-sensitive Markov decision processes. *Mathematics of Operations Research*, 39(1):105–120, 2014.

[5] R. Cavazos-Cadena. Solution to the risk-sensitive average cost optimality equation in a class of Markov decision processes with finite state space. *Mathematical Methods of Operations Research*, 57(2):263–285, 2003.

[6] R. Cavazos-Cadena and E. Fernández-Gaucherand. Markov decision processes with risk-sensitive criteria: Dynamic programming operators and discounted stochastic games. In *Proceedings of the IEEE Conference on Decision and Control*, volume 3, pages 2110–2112, 2001.

[7] R. Cavazos-Cadena and D. Hernández-Hernández. Discounted approximations for risk-sensitive average criteria in: Markov decision chains with finite state space. *Mathematics of Operations Research*, 36(1):133–146, 2011.

[8] R. Cavazos-Cadena and R. Montes-de Oca. Nearly optimal policies in risk-sensitive positive dynamic programming on discrete spaces. *Mathematical Methods of Operations Research*, 52(1):133–167, 2000.

[9] R. Cavazos-Cadena and R. Montes-De-Oca. The value iteration algorithm in risk-sensitive average Markov decision chains with finite state space. *Mathematics of Operations Research*, 28(4):752–776, 2003.

[10] R. Cavazos-Cadena and F. Salem-Silva. The discounted method and equivalence of average criteria for risk-sensitive Markov decision processes on Borel spaces.

Applied Mathematics and Optimization, 61(2):167–190, 2010.

[11] E. Delage and S. Mannor. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

[12] E. Denardo and U. Rothblum. A turnpike theorem for a risk-sensitive Markov decision process with stopping. *SIAM Journal on Control and Optimization*, 45(2):414–421, 2006.

[13] K. Dvijotham and E. Todorov. A unifying framework for linearly solvable control. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011*, pages 179–186, 2011.

[14] J. Garcıa and F. Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[15] P. Geibel and F. Wysotzki. Risk-sensitive reinforcement learning applied to control under constraints. *J. Artif. Intell. Res.(JAIR)*, 24:81–108, 2005.

[16] M. Heger. Risk and reinforcement learning: Concepts and dynamic programming. Technical report, Universitat Bremen, Germany, 1994.

[17] R. A. Howard and J. E. Matheson. Risk-sensitive markov decision processes. *Management science*, 18(7):356–369, 1972.

[18] Y. Kadota, M. Kurano, and M. Yasuda. Discounted markov decision processes with utility constraints. *Computers & Mathematics with Applications*, 51(2):279 – 284, 2006.

[19] H. Kashima. Risk-sensitive learning via minimization of empirical conditional value-at-risk. *IEICE TRANSACTIONS on Information and Systems*, 90(12):2043–2052, 2007.

[20] B. A. Kitchenham. Systematic review in software engineering: Where we are and where we should be going. In *Proceedings of the 2Nd International Workshop on Evidential Assessment of Software Technologies*, pages 1–2. ACM, 2012.

[21] A. Kumar, V. Kavitha, and N. Hemachandra. Finite horizon risk sensitive MDP and linear programming. In *Proceedings of the IEEE Conference on Decision and Control*, volume 2016-February, pages 7826–7831, 2016.

[22] Y. Liu and S. Koenig. Existence and finiteness conditions for risk-sensitive planning: First results. In *19th National Conference on Artificial Intelligence. AAAI Workshop*, volume WS-04-08, pages 49–54, 2004.

[23] Y. Liu and S. Koenig. Existence and finiteness conditions for risk-sensitive planning: Results and conjectures. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence, UAI 2005*, pages 354–363, 2005.

[24] D. G. Luenberger et al. Investment science. *OUP Catalogue*, 1997.

[25] H. Mausser and D. Rosen. Beyond VaR: From measuring risk to managing risk. In *Proceedings of the IEEE/IAFE 1999 Conference on Computational Intelligence for Financial Engineering (CIFEr)*, pages 163–178. IEEE, 1999.

[26] R. Minami and V. F. da Silva. Shortest stochastic path with risk sensitive evaluation. In *11th Mexican International Conference on Artificial Intelligence, MICAI*, pages 371–382. Springer Berlin Heidelberg, 2013.

[27] T. M. Moldovan and P. Abbeel. Risk aversion in Markov decision processes via near optimal Chernoff bounds. In *Advances in Neural Information Processing Systems, NIPS 2012*, pages 3131–3139, 2012.

[28] T. M. Moldovan and P. Abbeel. Safe exploration in Markov decision processes. In *Proceedings of the 29th International Conference on Machine Learning, ICML*, 2012.

[29] G. E. Monahan. State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms. *Management Science*, 28(1):1–16, 1982.

[30] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 799–806, 2010.

[31] T. Morimura, M. Sugiyama, H. Kashima, H. Hachiya, and T. Tanaka. Parametric return density estimation for reinforcement learning. *arXiv preprint arXiv:1203.3497*, 2012.

[32] A. Nilim and L. El Ghaoui. Robust control of Markov decision processes with uncertain transition matrices. *Oper. Res.*, 53(5):780–798, Sept. 2005.

[33] T. Osogami. Robustness and risk-sensitivity in Markov decision processes. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, NIPS'12, pages 233–241, 2012.

[34] S. Patek. On terminating Markov decision processes with a risk-averse objective function. *Automatica*, 37(9):1379–1386, 2001.

[35] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[36] W. A. S. Reis, K. V. Delgado, and L. N. de Barros. Distributed and asynchronous policy iteration for bounded parameter Markov decision processes. *XIII Encontro Nacional de Inteligência Artificial e Computacional - ENIAC*, 2016.

[37] M. Sato. TD algorithm for the variance of return and mean-variance reinforcement learning. *J. Japanese Society of Artificial Intelligence*, 16(3):353–362, 2002.

[38] Q. Wei. Continuous-time Markov decision processes with risk-sensitive finite-horizon cost criterion. *Math. Methods of Operations Research*, 84(3):461–487, 2016.

[39] Q. Wei and X. Chen. Continuous-time Markov decision processes under the risk-sensitive average cost criterion. *Operations Research Letters*, 44(4):457–462, 2016.