Processos de Decisão Markovianos Sensíveis a Risco: Uma abordagem de otimização mean-CVaR

Alternative Title: Risk Sensitive Markov Decision Processes: A mean-CVaR optimization approach

Denis Benevolo Pais
Escola de Artes, Ciências e Humanidades - USP
denis.pais@usp.br

Karina V. Delgado (Orientadora)
Escola de Artes, Ciências e Humanidades - USP
kvd@usp.br

RESUMO

Processos de decisão Markovianos (Markov Decision Process – MDP) são amplamente usados para resolver problemas de tomada de decisão sequencial. A função objetivo mais utilizada nesse tipo de problemas é minimizar o custo total esperado. Porém, esta abordagem não leva em consideração a variabilidade do custo (ou seja, flutuações em torno da média), o que pode afetar significativamente o seu desempenho geral. MDPs que lidam com esse tipo de problemas são chamados de MDPs sensíveis a risco. Um tipo de MDP sensível a risco é o CVaR MDP, que inclui a métrica CVaR comumente usada para medir risco financeiro. Neste trabalho propomos um algoritmo aproximado de iteração de valor para resolver um mean-CVAR MDP, um MDP sensível a risco que utiliza a média do custo total em conjunto com o critério CVaR.

Palavras-Chave

Processo de de Decisão Markoviano, Processo de Decisão Markoviano Sensível ao Risco, CVaR

ABSTRACT

Markov Decision Process (MDP) are widely used to solve sequential decision-making problems. The most commonly used objective function in this type of problem is to minimize the total expected cost. However, this approach does not take into account cost variability (i.e., fluctuations around the mean), which can significantly affect its overall performance. MDPs that deal with this type of problem are called risk-sensitive MDPs. One type of risk-sensitive MDP is the CVaR MDP, which includes the CVaR metric commonly used to measure financial risk. In this work, we propose an approximate value iteration algorithm to solve the mean-CVAR MDP, a risk-sensitive MDP that uses the average total cost in conjunction with the CVaR criterion.

SBSI 2018 June 4^{th} – 8^{th}, 2018, Caxias do Sul, Rio Grande do Sul, Brazil

CCS Concepts

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

Keywords

Markov Decision Process, Risk Sensitive Markov Decision Process, CVaR

1. INTRODUÇÃO

Um processo de decisão de Markov (MDP - Markov Decision Process) é uma forma de modelar processos nos quais as transições entre estados são probabilísticas, é possível observar em que estado o processo está e é possível interferir no processo periodicamente (em épocas de decisão) executando ações [4]. Cada ação tem um custo, que depende do estado em que o processo se encontra. São chamados "de Markov" (ou "Markovianos") porque os processos modelados obedecem a propriedade de Markov: o efeito de uma ação em um estado depende apenas da ação e do estado atual do sistema (e não de como o processo chegou a tal estado); e são chamados de processos "de decisão" porque modelam a possibilidade de um agente (ou tomador de decisões) interferir periodicamente no sistema executando ações.

A resolução de MDPs tipicamente envolve a minimização de uma determinada função objetivo de desempenho neutro em termos de risco, que é o custo total descontado esperado. Esta abordagem, embora muito popular, natural e atraente de um ponto de vista computacional, não leva em consideração a variabilidade do custo (ou seja, flutuações em torno da média), nem sua sensibilidade aos erros de modelagem, que podem afetar significativamente o desempenho geral [1].

Em muitas situações da vida real precisamos garantir com certo grau de certeza que obteremos um determinado resultado. Por exemplo, em um sistema de navegação autônomo, um agente tentará minimizar o comprimento esperado do seu caminho, e para isso ele pode provavelmente viajar perto de obstáculos na esperança de minimizar a distância. Entretanto esse mesmo agente também viajará perto de outros agentes e até mesmo de seres humanos, e uma falha ou desvio do caminho planejado pode resultar em uma colisão, ou um grave acidente causando uma perda irreversível (por exemplo, a morte de uma pessoa). Conseguimos lidar com esses problemas adicionando uma medida de risco, tornando o MDP sensível a risco.

Diversas medidas para mensuração de risco financeiro têm sido constantemente estudadas e aplicadas em diversos setores. Tais medidas são comumente empregadas em modelos de otimização estocástica aplicados para problemas do mercado financeiro e também da engenharia no geral. Entre essas medidas estão: variância , Value-at-Risk (VAR) e Conditional-Value-at-Risk (CVaR).

CVaR é considerada a principal e mais promissora métrica de risco. O CVaR é a perda esperada durante o intervalo de tempo, condicionada ao fato de se estar no ponto (100 - c%) da cauda esquerda da distribuição. Em palavras simples, CVaR é um métrica de risco que responde a pergunta: "se as coisas pioraram , quanto se pode esperar perder?".

CVaR destaca-se pelas seguintes características: (i) possui propriedades computacionais como eficiência numérica e estabilidade dos cálculos [6]; (ii) possui capacidade para proteger um tomador de decisão dos resultados que mais o prejudiquem [1]; (iii) CVaR é uma medida de risco coerente, isto é, ela atende quatro propriedades matemáticas importante de acordo com Rockefllar e Uryasee [5]; e (iv) CVaR pode ser representada por uma fórmula de minimização de baixa complexidade e esta fórmula pode ser facilmente incorporada em problemas de otimização, minimizando o risco do problema ou modelando-a como uma restrição.

Motivados pelas vantagens mencionadas anteriormente, diversos trabalhos sobre MDPs sensíveis a risco que utilizam esse critério foram propostos, entre eles [1], [3] e [7]. Esse novo MDP é chamado de CVaR MDP.

Neste trabalho propomos um algoritmo aproximado de iteração de valor para resolver um MDP sensível a risco que utilize a média do custo total em conjunto com o critério CVaR, esse problema é chamado de mean-CVAR MDP. A hipótese do trabalho é que se mean-CVaR MDP for utilizado para modelar os problemas, teremos um melhor compromisso entre custo e garantia de atingir a meta quando comparado com CVaR MDP, garantindo assim uma melhor modelagem de erro e um uso eficiente dos recursos.

2. APRESENTAÇÃO DO PROBLEMA

Nesta seção são apresentados os principais conceitos para mensuração de risco e os conceitos de MDPs e CVaR MDP.

2.1 Risco e CVaR

Considere um espaço de probabilidade S = (Ω, F, P), em que Ω é o espaço amostral, F são os eventos e P é chamada de medida de probabilidade.

Como mencionado existem muitas métricas de risco que são usadas para aplicações financeiras em especial destacamos a função Value at Risk (VaR), uma função muito popular utilizada constantemente para gestão de portfólio de ativos financeiros (chamados também de ações) e que utiliza técnicas estatísticas. VaR mede a pior perda esperada ao longo de determinado intervalo de tempo sob condições normais, e dentro de determinado nível de confiança, isto é, ela pode responder por exemplo a seguinte pergunta: "Qual é a perda mínima incorrida pela carteira nos a% piores cenários?

Podemos definir Value-at-risk (VaR) com nível de confiança $\alpha \in (0,1)$, como o quantil $1-\alpha$ de Z, i.e.,

$$VaR_\alpha(Z) = \min\{z | F(z) \geqslant \alpha\}, \quad (1)$$

em que Z, neste trabalho, é interpretado como custo.

Apesar da sua ampla utilização a função VaR tem certas limitações, entre as principais estão: (i) não é uma medida coerente de risco, (ii) é instável (existe alta flutuação sobre perturbações), (iii) é inapropriada quando Z não é distribuído normalmente, e (iv) não fornece a medida das perdas potenciais que excedem o valor do próprio VaR.

Uma alternativa que contorna as limitações da função VaR é a função conditional-value-at-risk (CVaR). Essa medida indica de forma mais adequada o potencial de perdas que ultrapassam o intervalo de confiança, definido ao se calcular a média das perdas que excedem o valor do VaR. Além disso, CVaR não precisa de um distribuição normalizada para o custo Z e por último apresenta uma maior estabilidade pois a flutuação sobre perturbações é menor. CVaR pode ser definida, com nível de confiança $\alpha \in (0,1)$ como demonstrado em [5] da seguinte forma:

$$CVaR_\alpha(Z) = \min_{w \in \mathbb{R}}\{w + \frac{1}{1-\alpha}\mathbb{E}[(Z-w)^+]\}, \quad (2)$$

em que $(x)^+ = \max(x, 0)$, representa a parte positiva de x; Z representa o custo em um determinado período de tempo e w representa a variável de decisão que, no ponto ótimo, atinge o valor do VaR.

A Equação 2 pode ser resolvida por meio de otimização linear e sua modelagem em computação tem vantagens que explicam a sua adoção, entre elas (i) o não uso de variáveis binárias para computar os resultados; e (ii) a não computação de integrais que permite uma implementação menos complexa dessa medida de risco. A crescente adoção do CVaR em modelos de otimização e em conjunto com MDPs se deve além de sua potencialidade, ao fato de Rockefllar e Uryasee [5] tê-la formulado-a matematicamente de modo a permitir sua otimização através de técnicas de programação linear. Tal feito alavancou a aplicação dessa medida de risco para resolução de problemas que envolvem decisões sob condições de incerteza.

2.2 MDP

Um processo de decisão de Markov (MDP) [4] é uma tupla M = (X,A,P,C,γ, x_0), em que: X é um conjunto finito de estados observáveis; A é um conjunto finito de ações; $P(\cdot|x,a)$ é a função probabilística de transição que descreve os efeitos da execução de uma ação $a \in A$ em um estado $x \in X$ resultando em um estado $x' \in X$; C(x,a) é a função custo de executar uma ação $a \in A$ em um estado $x \in X$; γ é o fator de desconto; e x_0 é o estado inicial.

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, sendo que a execução de uma ação em um estado tem um custo. A tomada de decisão é realizada durante um horizonte. O horizonte é o número de passos que o agente tem para agir, podendo ser: finito, infinito ou indeterminado.

Uma política π é um mapeamento de estados em ações, 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 minimiza a esperança descontada da soma total dos custos. Os algoritmos tradicionais para encontrar uma política ótima são: iteração de valor e iteração de política.

2.3 CVaR MDP

Em MDPs sensíveis a risco [1], a expectativa de risco neutro é substituída com alguma medida de risco como: vari-

ância, Value-at-Risk (VaR) ou condicional (CVaR). Neste trabalho estamos interessados no CVaR MDP. Uma formulação usando programação dinâmica para o problema CVaR MDP foi definido em [1].

Diferente da função valor de MDPs, que depende apenas do estado, na formulação em [1], a função valor V depende do estado e do nível de confiança. Assim, o espaço de estados X foi estendido com o nível de confiança Y =(0,1] e a função valor V(x, y) para o estados aumentado (x, y) foi definida como:

$$V(x,y) = \min_{\mu \in \Pi_H} CVaR_y(\lim_{T\to\infty} C_{0,T}|x_0 = x, \mu) \quad (3)$$

Em que Π_H é um conjunto de políticas e $C_{0,T} = \sum_{t=0}^{T} \gamma^t Z_t$, isto é, o custo total descontado até o tempo T.

Na computação de problemas de programação dinâmica é muito conveniente usar operadores que são definidos no espaço da função valor. Em [1] foi descrito o teorema da decomposição do CVaR, o qual conduziu a criação do operador de Bellman para CVaR $T : X \times Y \to X \times Y$:

$$T[V](x,y) = \min_{a\in\mathbb{A}}[C(x,a)+ \gamma \max_{\xi \in U_{CVAR}(y,P(.|x,a))} \sum_{x'\in X} \xi(x')V(x', y\xi(x'))P(x'|x,a)] \quad (4)$$

em que U_{CVAR} é chamado de envelope de risco [1] e é representado da seguinte forma::

$$U_{CVaR}(y,P) = (\xi : \xi(\omega) \in [0, 1/y], \int_{\omega\in\Omega} \xi(\omega)P(\omega)d\omega = 1) \quad (5)$$

Na Equação 4 é escolhida a melhor ação ($\min_{a\in\mathbb{A}}$) e é feita uma maximização contínua da expressão considerando o envelope de risco. Essa equação fornece duas propriedades fundamentais para a tratabilidade dos problemas: (i) contração e (ii) concavidade em y.

O operador de Bellman (Equação 4) apresenta duas dificuldades:

- Nos estados aumentados, Y é contínua.
- Aplicar T envolve realizar a maximização sobre ξ.

Em [1] foi proposto um algoritmo chamado *CVaR Value Iteration with Linear Interpolation* para lidar com essas dificuldades. O primeiro desafio é contornado com uso da interpolação linear o que permitiu a discretização do Y e também foi explorada a concavidade de yV(x,y) para delimitar o erro introduzido por essa técnica. O segundo desafio é contornado explorando a concavidade do problema de maximização para garantir que a otimização seja executada de forma eficaz.

3. PROPOSTA DE SOLUÇÃO

Este trabalho tem por objetivo principal propor um novo algoritmo de iteração de valor aproximado que resolva um MDP que minimize a soma da média dos custos descontados e o CVaR, i.e:

$$\min_{\mu\in\Pi_H} \lambda E(\lim_{T\to\infty} C_{0,T}|x_0,\mu) + (1-\lambda)CVaR_\alpha(\lim_{T\to\infty} C_{0,T}|x_0,\mu) \quad (6)$$

em que λ é um parâmetro que pondera cada um dos dois termos. Esse problema é chamado de mean-CVAR MDP. Pretende-se resolver instâncias de problemas relativamente grandes, ou seja que possuam um grande número de estados.

No nosso conhecimento existem poucos trabalhos que fazem o casamento do critério clássico para MDPs, isto é, a média dos custos descontados, com a métrica de risco CVaR, com isso esperamos expandir a literatura a respeito dessa abordagem. Nossa abordagem pretende estender o operador de Bellman de CVaR MDP para mean-CVaR MDP e usar interpolação.

4. AVALIAÇÃO DA SOLUÇÃO

Será feita a comparação do desempenho do algoritmo estado da arte de iteração de valor utilizando a função CVaR proposto em [1] com o novo algoritmo proposto em três domínios de teste: o problema de travessia do rio, o mundo grid e um terceiro relacionado ao mercado financeiro que será definido posteriormente.

Políticas com diferentes níveis de confiança serão obtidas considerando $\alpha = 0.01$, $\alpha = 0.11$ e $\alpha = 1$. Os aspectos de maior interesse a serem avaliados são: (i) o tempo de convergência do algoritmo, i.e., o tempo gasto para achar a politica; e (ii) a quantidade de falhas encontradas durante a simulação da política.

Para o domínio da travessia, as políticas obtidas com diferentes níveis de confiança serão avaliadas contabilizando a quantidade de falhas e sucessos em 500 simulações.

Similar a [1], para o domínio do mundo grid, serão conduzidas simulações em instâncias em que são introduzidos pertubações na posição dos obstáculos. Cada posição do obstáculo será perturbada, com uma probabilidade de 50%, em uma direção aleatória para uma de suas células vizinhas. Isso pode modelar erros de medição ao definir o mapa. As políticas obtidas com diferentes níveis de confiança serão avaliadas contabilizando a quantidade de falhas e sucessos em 500 mapas com pertubações.

Os códigos em python usados para os testes serão disponibilizados em um repositório do github.

4.1 Travessia do rio

O problema consiste na travessia do agente em um *grid* ($Nx \times Ny$) do canto inferior esquerdo para o respectivo canto inferior direito, que é o estado meta para o problema [2]. Nesse domínio existe apenas um agente com 4 ações de movimento (Norte, Sul, Leste, Oeste). A travessia só pode ser realizada (i) nadando no rio ou (ii) caminhando pela borda do rio até chegar a uma ponte. Neste domínio, o rio leva a uma cachoeira em que o agente pode cair ou até morrer.

4.2 Mundo grid

Neste problema os estados representam pontos de um mapa de terreno 2D [1]. Um agente (por exemplo, um veículo robótico) começa em uma região segura e seu objetivo é viajar para um determinado destino. Em cada passo do tempo, o agente pode se mover para qualquer um dos suas posições vizinhas. Porém, existe uma probabilidade p de se movimentar para um estado vizinho aleatório. O custo para se mover de um estado para o outro é 1 que está associado ao uso de combustível. Entre o ponto de partida e o destino, há uma série de obstáculos que o agente deve evitar. Bater em um obstáculo custa 1 e termina a missão. O objetivo é calcular um caminho seguro e que seja eficiente no consumo

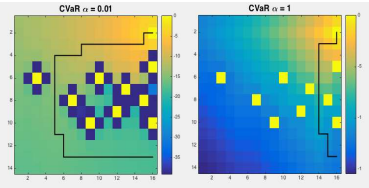

Figura 1: Política para uma instância do mundo grid com 224 estados com $\alpha = 0{,}01$ e $\alpha = 1$.

de combustível.

5. ATIVIDADES REALIZADAS

Foi realizada a revisão bibliográfica, implementados os algoritmos de interação de valor para MDPs clássicos a fim de conhecer a área de planejamento probabilístico e foi executado e avaliado o algoritmo de iteração de valor aproximado para resolver MDPs utilizando a função objetivo CVaR proposto em [1]. Esse algoritmo foi testado com uma instância pequena do problema do mundo grid contendo apenas 224 estados. A Figura 1 mostra a política com dois níveis de confiança diferentes, $\alpha = 0{,}01$ e $\alpha = 1$. Observe que com α baixo o agente prefere um caminho mais seguro (é mais avesso ao risco) enquanto com α alto é acontece o contrário, ele não se importa muito com os obstáculos, preferindo o caminho mais curto.

O algoritmo que gera problemas do domínio da travessia do rio já foi implementado e possue uma interface para facilitar a geração de instâncias de diferentes tamanhos.

6. CONCLUSÃO

A apesar dos avanços e dos resultados obtidos por alguns trabalhos que utilizam a função CVaR em MDPs, existe uma carência de trabalhos que mostrem resultados experimentais abrangentes avaliando esse novo critério. Além disso, poucos trabalhos conseguem resolver problemas com um número significativamente grande de estados e em tempo factível.

Com a proposta do novo critério que minimiza a soma da média dos custos descontados e o CVaR e do algoritmo para resolvê-lo, pretende-se contribuir com a área de planejamento probabilístico em inteligência artificial.

7. REFERÊNCIAS

[1] Y. Chow, A. Tamar, S. Mannor, and M. Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *Advances in Neural Information Systems*, 2015.

[2] V. Freire and K. V. Delgado. Gubs: A utility-based semantic for goal-directed Markov decision processes. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '17, pages 741–749, 2017.

[3] G. Iyengar. and A. K. C. Ma. Fast gradient descent method for mean-cvar optimization. *Annals of Operations Research*, 2009.

[4] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, New York, NY, 1994.

[5] R. Rockafeller and S. Uryasev. Conditional value-at-risk: Optimization approach. *Journal of risk*, 2000.

[6] R. Rockafeller and S. Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, 2002.

[7] Y.-L. C. Stefano Carpin. and M. Pavone. Risk aversion in finite Markov decision processes using total cost criteria and average value at risk. *Advances in Neural Information Systems*, 2016.