Obtenção dos compromissos Meta versus Custo em Processos Markovianos de Decisão

Isabella Kuo[1], **Valdinei Freire** [1]

[1]Escola de Artes, Ciências e Humanidades da Universidade de São Paulo
Rua Arlindo Béttio, 1000 - Ermelino Matarazzo,
São Paulo - SP, 03828-000

{15483114.ik, valdinei.freire}@usp.br

Resumo. *Processos Markovianos de Decisão modelam problemas de atingir uma meta com menor custo possível. Nesse trabalho estuda-se o compromisso entre probabilidade de chegar à meta e o custo médio incorrido.*

1. Processos Markovianos de Decisão

Processos Markovianos de Decisão (Markov Decision Process - MDP) modelam problemas nos quais um agente toma decisões sequenciais e os resultados das ações são probabilístico [Mausam and Kolobov 2012]. MDPs consideram um agente interagindo com um processo, e em todo tempo t: (*i*) o agente observa um estado s_t, (*ii*) o agente escolhe uma ação a_t, (*iii*) o agente paga um custo c_t; e (*iv*) o processo transita para um novo estado s_t. As transições e custos apresentam a propriedade de Markov, segundo a qual ambas só dependem do estado atual s_t e ação executada a_t. O objetivo do agente é escolher uma sequência de ações, de maneira que o sistema atue de forma ótima em função de um critério pré-estabelecido.

Nesse estudo, será utilizado uma formulação particular de MDP, o *Shortest Stochastic Path*, em que o agente deve buscar atingir um estado meta enquanto diminui o custo do caminho até o mesmo. Neste trabalho busca-se estabelecer os compromissos entre o custo e a probabilidade de chegar a meta, estabelecendo o ótimo de Pareto.

2. Shortest Stochastic Path

Este projeto apoia-se principalmente no problema de Caminho Estocástico mais Curto (*Shortest Stochastic Path* - SSP), que considera um estado meta e custos [Bertsekas and Tsitsiklis 1991]. Um estado meta é um estado terminal, isto é, quando o agente atinge a meta o processo se encerra. Um SSP é definido por uma tupla $\langle \mathcal{S}, \mathcal{A}, P, C, I, \mathcal{G} \rangle$ onde: $s \in \mathcal{S}$ são os estados possíveis; $a \in \mathcal{A}$ são as ações possíveis; $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ é a função de transição; $C : \mathcal{S} \times \mathcal{A} \to R^+$ é a função custo; $\mathcal{G}$ é o conjunto de estados meta; e $s_0 \in \mathcal{S}$ é o estado inicial.

A solução de um SSP consiste em políticas, que descrevem quais ações devem ser tomadas em cada situação. Uma política é própria se ela garante que a meta é alcançada com probabilidade 1.

Classicamente, considera-se problemas nos quais existem políticas próprias. Dada uma política própria, pode-se definir o custo acumulado esperado a partir de um estado $s \in \mathcal{S}$:

$$V^{\pi}(s) = \mathrm{E}\left[\sum_{t=0}^{\infty} c_t \middle| s_0 = s, \pi\right]$$

e define-se a política ótima $\pi^*(s) = \text{argmin}_\pi V^\pi(s)$.

Quando não existe política própria, ocorrem *dead-ends*, que são estados a partir dos quais não pode ser alcançado um estado meta. Na presença de *dead-ends* o custo acumulado esperado é infinito, portanto, deve-se considerar critérios diferentes do clássico.

3. Critério MCMP

Para avaliar políticas com dead-ends, é preciso considerar os seguintes conceitos:

- **Traço** T: uma sequência de estados que pode ser finita ou infinita,
- T^i: o i-ésimo estado do traço T,
- $\mathcal{T}_{\pi,s}$: o conjunto de todos traços obtidos a partir de s seguindo uma política π, e
- $\mathcal{T}^{\mathcal{G}}_{\pi,s} \subseteq \mathcal{T}_{\pi,s}$: o subconjunto de traços que alcança a meta.

Se um traço $T \in \mathcal{T}_{\pi,s}$ é finito, então ele alcança um estado meta, e define-se a probabilidade de um traço finito ocorrer seguindo uma política como:

$$P_{T,\pi} = \prod_{i=1}^{|T|} P(T^{i+1}|T^i, \pi(T^i)).$$

A probabilidade de uma política alcançar a meta é $P^{\mathcal{G}}_{\pi,s} = \sum_{T \in \mathcal{T}^{\mathcal{G}}_{\pi,s}} P_{T,\pi}$ e o custo acumulado de um traço finito é $C_{T,\pi} = \sum_{i=1}^{|T|} C(T^i, \pi(T^i))$.

3.1. Definição

Para lidar com traços finitos, o Critério MCMP (*Minimum Cost given Maximum Probability*) [Trevizan et al. 2017] permite uma ação de desistência, ou seja, quando ele se depara com um *dead-end*, ele pode desistir e pagar o custo gasto até aquele momento. Para garantir que a probabilidade de chegar a meta é maximizada, a ação de desistência é utilizada apenas quando um *dead-end* é encontrado.

Considere a seguinte função auxiliar que torna todos os traços de uma política finitos:

$$\psi(sT) = \begin{cases} s & \text{se } |T| = 0 \text{ ou } P^{\mathcal{G}}_{\pi,s} = 0 \\ s\psi(T) & \text{caso contrário} \end{cases}.$$

Considere $p_{max} = \max_\pi P^{\mathcal{G}}_{\pi,s_0}$, isto é, a probabilidade máxima de atingir a meta a partir do estado inicial s_0, o Critério MCMP é definido da seguinte forma:

$$\pi^* = \underset{\left\{\pi | P^{\mathcal{G}}_{\pi,s_0} = p_{max}\right\}}{\text{argmin}} \sum_{T \in \mathcal{T}_{\pi,s}} C_{\psi(T),\pi} P_{T,\pi}.$$

3.2. Solução via Programação Linear

O critério MCMP pode ser resolvido utilizando dois problemas de programação linear, LP1 e LP2 [Trevizan et al. 2017]. LP2 encontra a probabilidade máxima de se atingir a meta. LP1 encontra a política π com menor custo médio e com $P^{\mathcal{G}}_{\pi,s_0} = p_{max}$.

Ambos problemas consideram variáveis $x_{s,a}$ para todo $s \in \mathcal{S}$ e $a \in \mathcal{A}$, que indicam a frequência de ocorrência acumulada média para cada par estado-ação. A dinâmica do MDP é restringem as soluções ao especificar um modelo de fluxo de entrada $in(s)$ e

saída $out(s)$ para cada estado s, com as particularidades para o estado inicial e estados metas. Ambos problemas consideram as seguintes definições para todo $s \in \mathcal{S}$:

$$in(s) = \sum_{s' \in \mathcal{S}, a \in \mathcal{A}} x_{s',a} \mathcal{P}(s|s', a) \quad \text{e} \quad out(s) = \sum_{a \in \mathcal{A}_{(s)}} x_{s,a}.$$

O LP2 calcula o valor de p^{max}, utilizando as entradas para os estados metas como objetivo, desconsiderando completamente os custos envolvidos e é descrito abaixo:

$$\begin{array}{rll} \max_x & \sum_{s_g \in \mathcal{G}} in(s_g) & \\ \text{sujeito a} & x_{s,a} \geq 0 & \forall s \in \mathcal{S}, a \in \mathcal{A}(s) \\ & out(s) - in(s) = 0 & \forall s \in \mathcal{S} \setminus (\mathcal{G} \cup s_0) \\ & out(s_0) - in(s_0) = 1 & \end{array}$$

O LP1 calcula o custo mínimo restrito às dinâmicas do ambiente e também às políticas que garantem a probabilidade de chegar à meta e é descrito abaixo:

$$\begin{array}{rll} \min_{x_{s,a}} & \sum_{s \in \mathcal{S}, a \in \mathcal{A}} x_{s,a} C(s,a) & \\ \text{sujeito a} & x_{s,a} \geq 0 & \forall s \in \mathcal{S}, a \in \mathcal{A}(s) \\ & out(s) - in(s) \leq 0 & \forall s \in \mathcal{S} \setminus (\mathcal{G} \cup s_0) \\ & out(s_0) - in(s_0) \leq 1 & \\ (C1) & \sum_{s_g \in \mathcal{G}} in(s_g) = p_{max} & \end{array}$$

4. Compromisso entre Probabilidade para a Meta e Custo Médio

O critério MCMP estabelece um sistema de preferências que prioriza a chance de chegar à meta, independente do custo associado. No entanto, como já discutido por [Freire and Delgado 2017], uma pequena redução na probabilidade de chegar a meta pode trazer uma redução considerável no custo médio.

Um compromisso entre probabilidade de chegar à meta e custo médio pode ser obtido escolhendo uma probabilidade de chegar à meta arbitrária p e encontrando o menor custo correspondente. Esse custo pode ser obtido com o LP1, alterando a restrição $(C1)$ para $\sum_{s_g \in \mathcal{G}} in(s_g) = p$. Aqui obtemos os compromissos por amostragem com p variando entre 0 e p_{max} com intervalo de 0,02 para $p \leq 0,9$ e intervalo de 0,01 para $p > 0,9$.

Para fazer a avaliação do compromisso em um problema concreto, utilizou-se um problema artificial, o problema de Travessia do Rio, como descrito em [Freire et al. 2019]. O problema de travessia do rio considera um agente que deve chegar em um estado meta a partir de um estado inicial (estado 7 e 2, respectivamente, na figura 1). Para isso ele precisa atravessar um rio, que possui uma ponte (estado 6) e uma cachoeira (estado 4). O caminho pela ponte é maior, porém garante uma probabilidade maior de chegar na meta. Se o agente escolhe passar pelo rio o caminho é menor, porém, por conta da força da correnteza, há a probabilidade de cair na cachoeira, que é um *dead-end*. Os experimentos foram realizados em uma instância do problema do rio com: 20 de altura e 3 de largura, probabilidade de cair no rio 0,1, e força da correnteza 0,5.

A figura 1 mostra o compromisso entre probabilidade de chegar à meta e custo médio. Para o critério MCMP, tem-se $p_{max} = 0.95238$ e $c_{max} = 16.835$. Note que, pequenas alterações na probabilidade de chegar à meta pode gerar grandes ganhos em termos de custo médio ($p = 0.95000$ implica $c = 9.1240$).

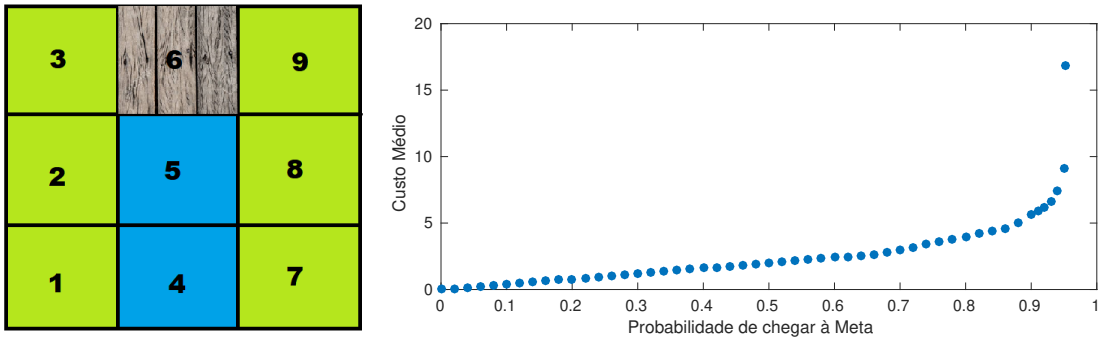

Figura 1. Esquerda: Problema do Rio de tamanho 3x3. Direita: Compromisso entre probabilidade da meta e custo médio.

5. Conclusão

Neste trabalho foram apresentados: (*i*) um método para levantar o compromisso entre probabilidade de alcançar a meta e custo médio; e (*ii*) um experimento em um problema artificial que deixa evidente a necessidade de estabelecer um compromisso entre esses dois objetivos. Embora mostrou-se o ótimo de Pareto para esse compromisso, o mesmo foi obtido por amostragem uniforme, os próximos passos consistem em estabelecer o ótimo de Pareto gerando apenas as políticas não-dominadas [Silva and Costa 2011] e criar um método para obtenção do melhor compromisso entre os dois objetivos para um usuário específico [Silva et al. 2006].

Referências

Bertsekas, D. P. and Tsitsiklis, J. N. (1991). An analysis of stochastic shortest path problems. *Mathematics of Operations Research*, 16(3):580–595.

Freire, V. and Delgado, K. V. (2017). Gubs: a utility-based semantic for goal-directed markov decision processes. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 741–749.

Freire, V., Delgado, K. V., and Reis, W. A. S. (2019). An exact algorithm to make a trade-off between cost and probability in ssps. In *Proceedings of the Twenty-Ninth International Conference on Automated Planning and Scheduling*, pages 146–154.

Mausam and Kolobov, A. (2012). Planning with markov decision processes: An ai perspective. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1).

Silva, V. F. d. and Costa, A. H. R. (2011). A geometric approach to find nondominated policies to imprecise reward mdps. In *Proceedings of the 2011 European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 439–454.

Silva, V. F. d., Costa, A. H. R., and Lima, P. (2006). Inverse reinforcement learning with evaluation. In *IEEE International Conference on Robotics and Automation (ICRA'06)*, pages 4246–4251, Orlando, FL. IEEE.

Trevizan, F., Teichteil-Königsbuch, F., and Thiébaux, S. (2017). Efficient solutions for stochastic shortest path problems with dead ends. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (UAI)*.