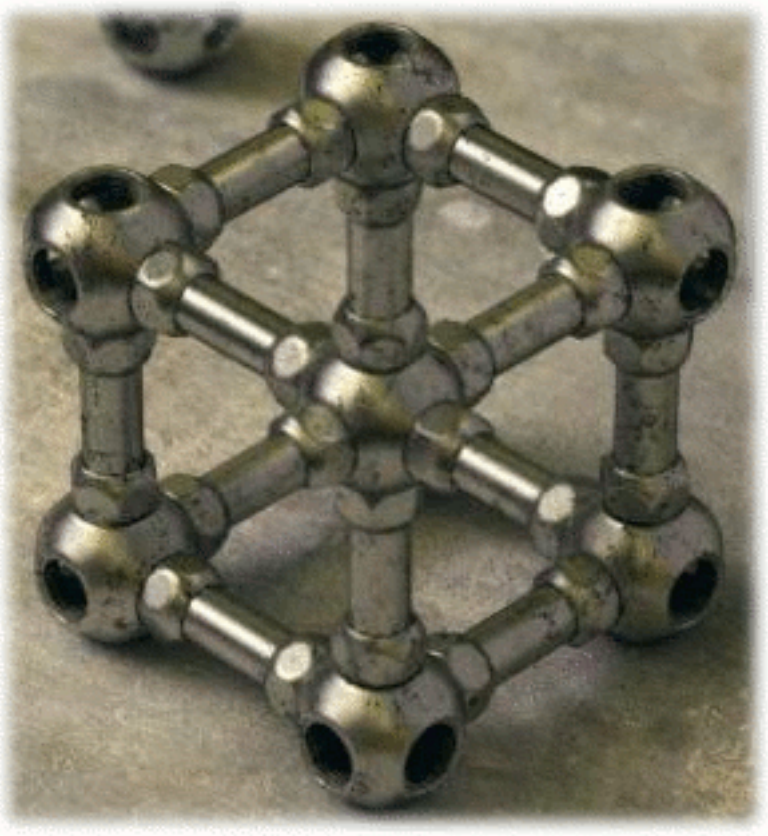

Jogos Markovianos Alternados com Probabilidades Imprecisas dadas por Conjuntos Credais [1]

Author(s):

Fábio de O. Franco
Leliane N. de Barros
Karina V. Delgado
Fábio G. Cozman

[1] This work was supported by Fapesp Project LogProb, grant 2008/03995-5, São Paulo, Brazil.

Jogos Markovianos Alternados com Probabilidades Imprecisas dadas por Conjuntos Credais

Fábio de O. Franco[1], Leliane N. de Barros[1], Karina V. Delgado[2], Fábio G. Cozman[3]

[1]Instituto de Matemática e Estatística – Universidade de São Paulo (USP)
Caixa Postal 66281 – 05508-090 – São Paulo – SP – Brazil

[2]Escola de Artes, Ciências e Humanidades – USP – São Paulo – SP – Brazil

[3]Escola Politécnica – Universidade de São Paulo – São Paulo – SP – Brazil

{ffranco,leliane}@ime.usp.br, {kvd,fgcozman}@usp.br

Abstract. *Alternate Markov Games (AMGs) are used to represent sequences of decisions with probabilistic effects and are particularly useful to model multi-agent interactions. However, in practice it is often impossible to obtain a precise value for all transition probabilities. A variant of AMGs has been recently proposed, where imprecision in probability values is represented by probability intervals. In this paper we present a more general variant of AMGs that allows for probabilistic imprecision to be expressed by credal sets (sets of probability distributions). We also define an algorithm that achieves the equilibrium value for the proposed model.*

Resumo. *Jogos Markovianos Alternados (AMGs) são usados para representar sequencias de decisões com efeitos probabilísticos e são particularmente úteis para modelar interações por múltiplos agentes. No entanto, na prática é frequentemente impossível obter um valor preciso para cada probabilidade de transição. Uma variante dos AMGs foi proposta recentemente, onde imprecisão em valores de probabilidades é representada por intervalos de probabilidade. Neste artigo apresentamos uma variante mais geral de AMGs que permite que a imprecisão seja expressa por conjuntos credais (conjuntos de distribuições de probabilidades). Também definimos um algoritmo que encontra o valor de equilíbrio para o modelo proposto.*

1. Introdução

A teoria dos jogos é amplamente utilizada em uma variedade de áreas, da economia à política, da ciência da computação ao entretenimento [Ferguson 2011]. Na pesquisa em técnicas para inteligência artificial, a teoria de jogos é uma ferramenta poderosa para analisar e sintetizar sistemas multi-agentes, em que a decisão de um agente pode afetar diretamente o comportamento de todos os outros agentes.

Normalmente, um jogo é caracterizado pelo número de jogadores que interagem entre si e com o ambiente, executando ações sob condições de incerteza e recebendo alguma recompensa (positiva ou negativa). Uma definição ampla e informal para jogos desse tipo é dada por [Russell and Norvig 2003]:

– um conjunto de jogadores que tomam decisões (ações);
– um conjunto de estados em que os jogadores podem se encontrar num determinado estágio do jogo;
– uma coleção de conjuntos de ações definidas, um conjunto para cada jogador;
– uma função de recompensa que dá a utilidade de possíveis combinações de ações.

Um jogador deve selecionar e executar uma *política*: uma política *determinística* especifica uma ação para cada estado, enquanto que uma política *probabilística* especifica uma distribuição de probabilidades sobre ações para cada estado.

Uma *solução* para um jogo prescreve a melhor política para cada jogador. Frequentemente procura-se uma solução na qual as políticas estão em equilíbrio de Nash, isto é, nenhum jogador se beneficia em mudar sua própria política se todos os outros jogadores mantém intactas as suas próprias políticas de equilíbrio. Esse equilíbrio pode ser visto como um ótimo local no espaço de políticas.

De acordo com a definição acima, os jogadores estão todos em algum estado conhecido antes da execução de suas ações e, após executá-las, se movem para um outro estado. A transição entre os estados pode ser modelada por uma distribuição de probabilidades. Um jogo markoviano é um jogo em que a função de transição para o próximo estado é determinada apenas pelo estado atual e as ações selecionadas neste estado [Fudenberg and Tirole 1991].

Um tipo importante de jogo markoviano é o *jogo markoviano alternado* (*Alternate Markov Game* – AMG) [Littman 1996], no qual:

1. os jogadores alternam nas escolhas de suas ações (apenas um jogador realiza uma ação em cada estado), e
2. o jogador conhece as jogadas realizadas pelos outros jogadores neste jogo, ou seja, o jogo é de informação perfeita [Parthasarathy and Raghavan 1971].

Permitir que seja inserida incerteza na função de transição probabilística pode trazer mais realismo e flexibilidade a um AMG, bem como a capacidade de compactar jogos muito grandes (jogos envolvendo um número muito grande de estados). Podemos representar as transições entre os estados com intervalos de probabilidade ou conjuntos de distribuições de probabilidade em duas situações: quando houver crenças incompletas e ambíguas sobre as probabilidades de transição ou quando o jogo for muito grande e para resolvê-lo é preciso fazer a agregação de estados. Um modelo que permite representar as crenças com intervalos de probabilidades para lidar com essas duas limitações foi formulado em [Chang 2006]. Os resultados de Chang são voltados para a análise de sensibilidade e soluções aproximadas de jogos markovianos.

O objetivo deste artigo é generalizar o trabalho de Chang de modo que as transições sejam expressas através de conjuntos de distribuições de probabilidade, denominados *conjuntos credais* (*credal sets*) [Cozman 2000]. A idéia vem de trabalhos anteriores sobre Processos de Decisão de Markov com Probabilidades Imprecisas (*Markov Decision Process with Imprecise Probabilities* – MDPIPs) [Delgado et al. 2009]. Um MDPIP é simplesmente uma extensão de um Processo Markoviano de Decisão [Puterman 1994] em que as probabilidades de transição podem ser imprecisas, ou seja, ao invés de uma medida de probabilidade sobre o espaço

de estados, temos um conjunto de medidas de probabilidade [Delgado et al. 2011]. O objetivo é criar um modelo análogo para jogos markovianos alternados, que chamaremos de *jogos markovianos alternados com probabilidades imprecisas dadas por conjuntos credais* (*Alternate Markov Games with Imprecise Probabilities* – AMG-IP).

Este artigo está organizado da seguinte forma. Na Seção 2 fazemos uma breve revisão sobre jogos markovianos alternados de dois jogadores de soma zero com probabilidades precisas. Na Seção 3 discutimos modelos gerais de incerteza para jogos markovianos alternados. Na Seção 4, apresentamos um novo modelo de jogo markoviano alternado com probabilidades imprecisas dadas por conjuntos credais. Na Seção 5, apresentamos uma solução para o modelo proposto e descrevemos um algoritmo. Na Seção 6, discutimos os trabalhos relacionados. Na Seção 7, concluímos e sugerimos possíveis trabalhos futuros.

2. Jogos markovianos

Um jogo markoviano (ou jogo de Markov) [Owen 1982] é definido por um conjunto de k jogadores, um conjunto de estados S, uma coleção de conjuntos de ações $A_1, A_2, ..., A_k$, um para cada jogador, e definições para as funções de transições de estados e funções de recompensas. A transição de estado depende probabilisticamente do estado atual e da ação selecionada por cada jogador, isto é, $p(s'|s, a_1, a_2, ..., a_k)$ é a probabilidade do jogador ir para o estado s', dado que ele está em s e $a_1, a_2, ..., a_k$ são as ações escolhidas pelos jogadores $1, 2, ..., k$, respectivamente. É definido para cada jogador i uma função de recompensa $R_i(s, a_1, a_2, ..., a_k)$, que depende do estado e das ações tomadas por cada jogador. O objetivo de cada jogador é maximizar o valor esperado de suas recompensas descontadas, $E\{\sum_{j=0}^{\infty}\gamma^j r_{t+j}^i\}$, em que r_{t+j}^i é a recompensa recebida pelo jogador i em j passos para o futuro, sendo $\gamma \in]0, 1[$ um fator de desconto. O horizonte T é o número de estágios restantes do jogo. Assim, a recompensa obtida em t estágios para o futuro é descontada por um fator γ^t.

Definição 1 (**Jogo markoviano alternado**). *Um jogo markoviano alternado de dois jogadores de soma zero é definido pela tupla* $\langle S, A_1, A_2, P, R, \gamma\rangle$ *de tal forma que: o conjunto de estados* S *é decomposto em dois conjuntos,* Y *(estados do jogador 1) e* Z *(estados do jogador 2), em que* $Y \cap Z = \emptyset$ *e* $Y \cup Z = S$*. A função de transição de estado é definida por* $P : S \times A_1 \times A_2 \to PD(S)$ *em que* $PD(S)$ *representa o conjunto de distribuições de probabilidades sobre* S*. Chamamos de* $A_i(s)$ *o conjunto finito de ações executáveis em* s *para o jogador* i*. Para cada jogador* i*, existe uma ação* $noop \in A_i$ *que pode ser executada em qualquer estado do jogo e cujo efeito é nulo, ou seja, se todos os agentes executarem suas ações* $noop$ *no estado* s*, o jogo permanecerá no mesmo estado. Os jogadores fazem suas jogadas alternadamente dado que* $P(y'|y, a_1, a_2) = 0$ *para todos os* $y, y' \in Y$*,* $a_1 \in A_1(y)$*,* $a_2 \in A_2(y)$ *e* $P(z'|z, a_1, a_2) = 0$ *para todos os* $z, z' \in Z$*,* $a_1 \in A_1(z), a_2 \in A_2(z)$*. Em cada estado* $s \in (Y \cup Z)$ *somente um jogador tem uma ou mais ações executáveis em* s*. A função de recompensa é dada por* $R : S \times A_1 \times A_2 \to \mathbb{R}$ *que representa as recompensas instantâneas do jogador. Como se trata de um jogo de soma zero, essa mesma função recompensa é positiva para o jogador 1 e negativa para o jogador 2. Assim, chamaremos o jogador 1 de maximizador e o jogador 2 de minimizador (ou de adversário).* $\gamma \in]0, 1[$ *é o fator de desconto, como descrito anteriormente.* □

Chamamos de política estacionária uma política em que a ação especificada para cada estado independe do estágio do jogo. Definimos $\pi : S \to A_1(S)$ como a política estacionária para o maximizador e por Π o conjunto de todas as políticas estacionárias do maximizador. Da mesma forma, Φ é o conjunto de todas as políticas estacionárias para o minimizador.

O valor de um jogo que segue as políticas π e ϕ, isto é, as políticas do maximizador e o minimizador, respectivamente, a partir de um estado inicial $s \in S$, é dado por:

$$V(\pi, \phi)(s) = R(s, \pi(s), \phi(s)) + \gamma \sum_{s' \in S} p(s'|s, \pi(s), \phi(s)) V(\pi, \phi)(s'). \tag{1}$$

Ou seja, o valor de um estado s quando os jogadores 1 e 2 estão seguindo as políticas π e ϕ, respectivamente, é a soma da recompensa atual e o valor esperado dos próximos estados.

A solução de um jogo é frequentemente vista como um equilíbrio (ou *saddle point*), isto é, as políticas π e ϕ são escolhidas de forma que os jogadores não podem obter uma recompensa maior mudando suas ações.

Definição 2 (**Políticas de equilíbrio**). *Um par da políticas $\pi^* \in \Pi$ e $\phi^* \in \Phi$ é um par de políticas (ótimas) de equilíbrio se não existe uma política $\phi \in \Phi$ tal que [Osborne and Rubinstein 1994]:*

$$V(\pi^*, \phi)(s) < V(\pi^*, \phi^*)(s), \quad s \in S, \tag{2}$$

e não há uma política $\pi \in \Pi$ tal que

$$V(\pi^*, \phi^*)(s) < V(\pi, \phi^*)(s), \quad s \in S. \quad \square \tag{3}$$

Seja $Q(s, a_1, a_2)$ o valor da recompensa esperada para cada estado $s \in S$, quando as ações $a_1 \in A_1$ e $a_2 \in A_2$ são selecionadas, dada por [Littman 1996]:

$$Q(s, a_1, a_2) = R(s, a_1, a_2) + \gamma \sum_{s' \in S} p(s'|s, a_1, a_2) V^*(s'), \tag{4}$$

Assim, a solução de um AMG de dois jogadores de soma zero é encontrar um par de políticas ótimas de equilíbrio π^* and ϕ^* que produzem o valor *maximin* em cada estado, ou seja:

$$V(\pi^*, \phi^*)(s) = V^*(s) = \max_{a_1 \in A_1} \min_{a_2 \in A_2} (Q(s, a_1, a_2))$$

$$= \min_{a_2 \in A_2} \max_{a_1 \in A_1} \left(R(s, a_1, a_2) + \gamma \sum_{s' \in S} p(s'|s, a_1, a_2) V^*(s') \right). \tag{5}$$

Note que as políticas ótimas, π^* e ϕ^*, são definidas calculando-se as funções $\arg\max$ e $\arg\min$, respectivamente, ou seja:

$$(\pi^*, \phi^*)(s) = \arg\max_{a_1 \in A_1} \arg\min_{a_2 \in A_2} \left(R(s, a_1, a_2) + \gamma \sum_{s' \in S} p(s'|s, a_1, a_2) V^*(s') \right). \quad (6)$$

Veja ainda que na Equação (5) a ordem dos cálculos de *min* e *max* não altera o valor do jogo. Esta propriedade foi provada por [Shapley 1953], que também provou que a solução *minimax* ou *maximin* converge, isto é, a Equação (5) encontra o valor de equilíbrio de um jogo AMG de dois jogadores de soma zero.

Note que com a formulação da Equação (5) temos uma única função valor de equilíbrio, embora possa haver diferentes pares de políticas que satisfaçam a Equação (6). Como veremos nas próximas seções, no caso do AMG-IP, a função valor de equilíbrio não é única [Chang 2006].

3. Incerteza sobre a função de transição de estado

Como dissemos anteriormente, em geral, é difícil ou impossível obter de forma precisa as probabilidades de transição de estados para jogos markovianos. Existem várias maneiras de abordar tais situações, entre elas:

- Associar um intervalo de probabilidades para a transição do estado s para o estado s' dado por $[\underline{p}(.|s, a_1, a_2), \overline{p}(.|s, a_1, a_2)]$. A Figura 1a ilustra um conjunto de distribuições de probabilidades. Note que uma distribuição de probabilidade é definida escolhendo-se um valor de probabilidade dentro de cada intervalo, tal que a soma da distribuição de probabilidade seja igual a 1.
- Podemos impor restrições sobre os valores das probabilidades de transição a partir do estado s para estado s'. Estas restrições podem ser definidas como intervalos, mas também podem ser mais gerais como na Figura 1b que ilustra um exemplo de transições probabilísticas parametrizadas por p, q e r. Uma distribuição de probabilidades é escolhida de tal forma que satisfaça as restrições sobre p, q e r.

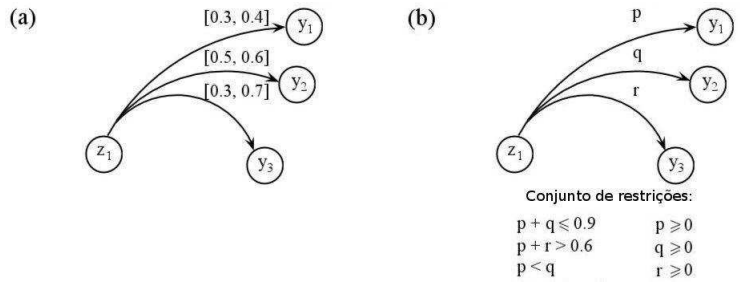

Figura 1. Duas maneiras de representar a imprecisão nas probabilidades de transição em um AMG.

Um *conjunto credal* é definido como um conjunto de distribuições de probabilidades. Por exemplo, a Figura 1a tem como conjunto de probabilidades todas as distribuições limitadas pelos intervalos que definem um conjunto credal. Na Figura 1b, o conjunto de distribuições de probabilidades que satisfazem o conjunto de restrições também define um conjunto credal.

Um conjunto credal que define as distribuições condicionais sobre o próximo estado s', dado um estado s e ações a_1 e a_2 é referido como um conjunto credal de transição (*transition credal set*) e denotado por $K(s'|s, a_1, a_2)$. Assumimos que todos os conjuntos credais são fechados, conexos e convexos [Walley 1991].

Exemplo 1. *Suponha que dado* $S = \{s_0, s_1, s_2\}$ *e as escolhas de ações* $a_1 \in A_1$ *e* $a_2 \in A_2$*, o conjunto de parâmetros de probabilidade* $p(s'|s, a_1, a_2) \in K(s'|s, a_1, a_2)$*, nomeados por* p_0*,* p_1 *e* p_2*, em que*

$$p_0 = p(s_0|s_0, a_1, a_2),$$
$$p_1 = p(s_1|s_0, a_1, a_2),$$
$$p_2 = p(s_2|s_0, a_1, a_2),$$

são definidos pelo seguinte conjunto de restrições:

$$C = \{p_0 \leq 2/3, \quad p_2 \leq 2/3, \quad 2p_0 \geq p_1\}.$$

A região bi-dimensional de todas as medidas de probabilidade que satisfazem C *é mostrada na região cinza da Figura 2, que corresponde ao conjunto credal para o exemplo dado.*

$\Diamond$

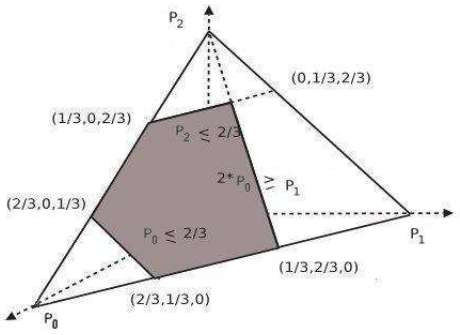

Figura 2. Um exemplo de um conjunto credal definido pelos valores dos parâmetros de probabilidade p_0, p_1 e p_2 que satisfazem o conjunto de restrições C (região cinza).

4. AMG-IP

Num AMG de dois jogadores de soma zero com a imprecisão nas probabilidades dadas por conjuntos credais, o maximizador e o minimizador selecionam e executam suas ações com o conhecimento do estado $s \in S$ em que eles se encontram. O jogo faz uma transição com uma escolha de probabilidades $p(s'|s, a_1, a_2) \in K(s'|s, a_1, a_2)$. Interpretamos essa escolha como uma escolha da Natureza. Como resultado o jogador maximizador recebe uma recompensa de $R(s, a_1, a_2)$ que depende somente do estado em que o jogo está e das escolhas de ações dos jogadores.

Definição 3 (AMG-IP). *Um jogo markoviano alternado de dois jogadores de soma zero com probabilidades imprecisas dadas por conjuntos credais (Alternate Markov Game with Imprecise Probabilities – AMG-IP) é dado pela tupla* $\langle S, A_1, A_2, K, R, \gamma \rangle$*, em que:* S *é um conjunto discreto e finito de estados completamente observáveis do*

jogo. S é composto de dois subconjuntos disjuntos Y e Z; A_1 é o conjunto finito de ações disponíveis para o jogador 1; A_2 é o conjunto finito de ações disponíveis para o jogador 2; $K(s'|s, a_1, a_2)$ define um conjunto de medidas de probabilidade de transição válidas, isto é, um conjunto credal de transição. $K(s'|s, a_1, a_2)$ pode ser implicitamente representado por um conjunto de probabilidades de transição consistente com um conjunto de restrições lineares; $R(s, a_1, a_2) \in \mathbb{R}$ é a função recompensa e; $\gamma \in]0, 1[$ é o fator de desconto.

A Equação (1) pode ser usada para calcular $V(\pi, \phi)$ num AMG-IP considerando que $p(s'|s, \pi(s), \phi(s)) \in K(s'|s, \pi(s), \phi(s))$, e assumindo que K é convexo, conexo e fechado.

Há vários critérios que podem ser usados para definir o valor ótimo de um jogo AMG-IP. Isto é, podemos usar diferentes critérios para escolher quais políticas os jogadores devem usar diante das escolhas da Natureza. Podemos usar, por exemplo, o critério sob a suposição que a Natureza seleciona $p(s'|s, a_1, a_2)$ adversariamente, isto é, minimizando a função valor. Assim, a Equação (5) deve ser modificada da seguinte forma:

$$V^*(s) = \max_{a_1 \in A_1} \min_{a_2 \in A_2} \left(R(s, a_1, a_2) + \gamma \min_{p \in K} \sum_{s' \in S} p(s'|s, a_1, a_2) V^*(s') \right), \tag{7}$$

ou seja, dadas as escolhas $a_1 \in A_1$ e $a_2 \in A_2$ dos dois jogadores, $p \in K$ é selecionado pela Natureza de forma a minimizar a recompensa esperada nos estágios futuros.

Outro possível critério é assumir que a Natureza selecionará $p \in K$ para maximizar a função valor. Assim, a Equação (5) é substituída por:

$$V^*(s) = \max_{a_1 \in A_1} \min_{a_2 \in A_2} \left(R(s, a_1, a_2) + \gamma \max_{p \in K} \sum_{s' \in S} p(s'|s, a_1, a_2) V^*(s') \right). \tag{8}$$

Note que (7) define o limite mínimo para $V^*(s)$ (a melhor escolha no pior caso), que chamaremos de $\underline{V}^*(s)$, enquanto que (8) define o limite superior para $V^*(s)$ (a melhor escolha no melhor caso), que chamaremos de $\overline{V}^*(s)$.

O melhor compromisso entre esses dois critérios de escolha da Natureza é garantir a melhor escolha no pior caso. Uma vez que existe um conjunto Δ de pares de políticas (π^*, ϕ^*) que satisfaz a Equação (7), podemos agora escolher o melhor par de políticas do conjunto Δ usando a Equação (8). Ou seja, sem o risco de diminuir a recompensa esperada, é possível assumir uma Natureza que coopera com o maximizador e escolhe a melhor política em Δ.

Portanto, para encontrar uma solução segundo o critério descrito acima, calculamos primeiro:

$$\underline{V}^*(s) = \max_{a_1 \in A_1} \min_{a_2 \in A_2} \left(R(s, a_1, a_2) + \gamma \min_{p \in K} \sum_{s' \in S} p(s'|s, a_1, a_2) \underline{V}^*(s') \right), \tag{9}$$

e em seguida, utilizamos as ações que alcançam $\underline{V}^*(s)$ para finalmente calcular o valor ótimo (V^*) assumindo a Natureza cooperativa, ou seja:

$$V^*(s) = \max_{a_1 \in A_1[\underline{V}^*](s)} \min_{a_2 \in A_2[\underline{V}^*](s)} \left(R(s, a_1, a_2) + \gamma \max_{p \in K} \sum_{s' \in S} p(s'|s, a_1, a_2) V^*(s') \right), \quad (10)$$

em que $A_1[\underline{V}^*](s)$ e $A_2[\underline{V}^*](s)$ são os conjuntos de ações do jogador 1 e 2, respectivamente, que alcançam o valor ótimo quando a Natureza é adversária do jogador 1.

Em [Chang 2006] foi provado que esse critério de compromisso entre escolhas no pior caso (Natureza min) e escolhas no melhor caso (Natureza max) define a função valor de equilíbrio para um AMG com probabilidades imprecisas dadas por intervalos, ou seja, $p \in [\underline{p}(.|s, a_1, a_2), \overline{p}(.|s, a_1, a_2)]$. É possível provar que isso também vale quando $p \in K(.|s, a_1, a_2)$, ou seja, para imprecisões nas probabilidades definidas por conjuntos credais se assumirmos que eles sejam fechados, conexos e convexos.

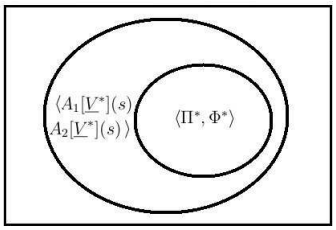

Figura 3. Pares de políticas de equilíbrio em um AMG-IP.

A Figura 3 ilustra o critério usado para encontrar os pares de políticas de equilíbrio num jogo AMG-IP. Os conjuntos de pares de políticas ótimas $\langle \Pi^*, \Phi^* \rangle$ são subconjuntos das políticas formadas pelos conjuntos $A_1[\underline{V}^*](s)$ e $A_2[\underline{V}^*](s)$.

Políticas de equilíbrio para AMG-IPs Como foi dito na seção anterior, dado um AMG-IP e considerando os conjuntos de distribuições de probabilidades $K(.|s, a_1, a_2)$, com $s \in S, a_1 \in A_1, a_2 \in A_2$, fechados, conexos e convexos, o limite inferior $\underline{V}^*(s)$ da função valor pode ser calculado pela Equação 9.

Estendemos os resultados de Chang (2006) que considera intervalos reais da forma $[\underline{p}(.|s, a_1, a_2), \overline{p}(.|s, a_1, a_2)]$ gerando um conjunto credal fechado, conexo e convexo de conjuntos de probabilidades, fazendo essa mesma suposição sobre $K(.|s, a_1, a_2)$ gerado a partir de um conjunto de restrições lineares. Com isso, garantimos encontrar a função valor ótima $\underline{V}^*(s)$ para o caso da Natureza adversária. Como Chang (2006) provou, a condição *maximin* nas escolhas de ações da Equação 9 pode ser invertida para *minimax*, garantindo assim que as políticas ótimas geradas por essa estratégia (que chamaremos de políticas conservadoras) sejam políticas de equilíbrio, independente das escolhas da Natureza.

Finalmente, o par de políticas π^* e ϕ^* que alcançam V^* (Equação 10) também estão em equilíbrio uma vez que elas pertencem ao conjunto de políticas conservadoras, que por sua vez, são política de equilíbrio A escolha de π^* e ϕ^* leva em

consideração a possibilidade da Natureza agir de maneira cooperativa. Isso só pode aumentar a utilidade para ambos os jogadores, mas nunca diminui-la. Portanto, π^* e ϕ^* também são políticas de equilíbrio.

5. Iteração de valor para AMG-IP

Podemos implementar as Equações (9) e (10) usando programação dinâmica, por exemplo, generalizando o algoritmo de Iteração de valor. O algoritmo VALUE-ITERATION-AMG-IP (Algoritmo 1) devolve a função valor ótima para um AMG-IP e tem como parâmetros de entrada o AMG-IP e o número N de iterações desejado. Para isso, VALUE-ITERATION-AMG-IP faz chamadas a dois outros algoritmos, SOLVEONE (Algoritmo 2) e SOLVETWO (Algoritmo 3).

O algoritmo SOLVEONE implementa a Equação (9) devolvendo os conjuntos $A_1[\underline{V}^*](s)$ e $A_2[\underline{V}^*](s)$. De forma semelhante, o algoritmo SOLVETWO implementa a Equação (10) e devolve V^*. Estes dois algoritmos são semelhantes ao algoritmo de iteração por valor de AMGs [Littman 1996], porém o cálculo da recompensa total descontada em cada estado $s \in S$ é realizado por REGRESS (Algoritmo 4).

A seqüencia de execução do algoritmo VALUE-ITERATION-AMG-IP é definida da seguinte forma:

1. O algoritmo VALUE-ITERATION-AMG-IP chama o algoritmo SOLVEONE, passando os mesmos parâmetros recebidos na entrada.
2. O algoritmo SOLVEONE, em cada iteração, encontra para todos os estados do jogo os pares de políticas $\pi^*(s)$ e $\phi^*(s)$ que alcançam $\underline{V}^*$. Para isso o algoritmo SOLVEONE chama o algoritmo REGRESS afim de calcular:
$$Q(s, a_1, a_2) = R(s, a_1, a_2) + \gamma \min_{p \in K} \sum_{s' \in S} p(s'|s, a_1, a_2)\underline{V}^*(s').$$
Note que REGRESS precisa chamar um otimizador para poder calcular Q. Após ocorrerem as N iterações, SOLVEONE devolve os conjuntos $A_1[\underline{V}^*]$ e $A_2[\underline{V}^*]$ para VALUE-ITERATION-AMG-IP.
3. O algoritmo SOLVETWO é chamado para calcular $V^*(s)$ passando como parâmetros o AMG-IP e o número N de iterações, porém, ao invés de A_1 e A_2 são passados os novos conjuntos de ações $A_1[\underline{V}^*]$ e $A_2[\underline{V}^*]$.
4. O algoritmo SOLVETWO, em cada iteração, encontra para todos os estados os pares de políticas $\pi(s)^*$ e $\phi(s)^*$ que alcançam V^*. Para isso o algoritmo SOLVETWO chama o algoritmo REGRESS afim de calcular:
$$Q(s, a_1, a_2) = R(s, a_1, a_2) + \gamma \max_{p \in K} \sum_{s' \in S} p(s'|s, a_1, a_2)V^*(s')$$

Algoritmo 1: VALUE-ITERATION-AMG-IP(AMG-IP,*maxIter*)

input : AMGP-IP (given by $\langle S, A_1, A_2, R, K, \gamma \rangle$),
maxIter (maximum number of iterations)
output: V^* (t-state-to-go value function)
begin
$\langle A_1[\underline{V}^*], A_2[\underline{V}^*] \rangle$ = SOLVEONE($\langle S, A_1, A_2, R, K, \gamma \rangle$,*maxIter*);
V^* = SOLVETWO($\langle S, A_1[\underline{V}^*], A_2[\underline{V}^*], R, K, \gamma \rangle$,*maxIter*);
return V^* ;

Algoritmo 2: SolveOne(AMG-IP,$maxIter$)

input : AMGP-IP (given by $\langle S, A_1, A_2, R, K, \gamma \rangle$),
$maxIter$ (maximum number of iterations)
output: $\langle A_1[\underline{V}^*], A_2[\underline{V}^*] \rangle$ (players set actions that achieves $\underline{V}^*$)

```
begin
    V^0 = 0;
    t = 0;
    //build new actions set for the players
    A_1[V*] = ∅;
    A_2[V*] = ∅;
    //construct t-stage-to-go value functions V^t until termination condition is met
    while i < maxIter do
        t = t + 1;
        foreach s ∈ S do
            V^t(s) = 0;
            //update V^t with Q^t
            foreach a_1 ∈ A_1(s) do
                V^t_min(s) = ∞;
                foreach a_2 ∈ A_2(s) do
                    Q^t = Regress(V^{t-1},s,S,a_1,a_2,R,K,γ,true);
                    φ(s) = argmin(V^t_min(s),Q^t);
                    V^t_min(s) = min(V^t_min(s),Q^t);
                π(s) = argmax(V^t(s),V^t_min(s));
                V^t(s) = max(V^t(s),V^t_min(s));
            Add π(s) in A_1[V*](s);
            Add φ(s) in A_2[V*](s);
    return ⟨A_1[V*], A_2[V*]⟩;
```

Algoritmo 3: SolveTwo(AMG-IP,$maxIter$)

input : AMGP-IP (given by $\langle S, A_1[\underline{V}^*], A_2[\underline{V}^*], R, K, \gamma \rangle$),
$maxIter$ (maximum number of iterations)
output: V^t (t-state-to-go value function)

```
begin
    V^0 = 0;
    t = 0;
    //construct t-stage-to-go value functions V^t until termination condition is met
    while i < maxIter do
        t = t + 1;
        foreach s ∈ S do
            V^t(s) = 0;
            //update V^t with Q^t
            foreach a_1 ∈ A_1[V*](s) do
                V^t_min(s) = ∞;
                foreach a_2 ∈ A_2[V*](s) do
                    Q^t = Regress(V^{t-1},s,S,a_1,a_2,R,K,γ,false);
                    φ(s) = argmin(V^t_min(s),Q^t);
                    V^t_min(s) = min(V^t_min(s),Q^t);
                π(s) = argmax(V^t(s),V^t_min(s));
                V^t(s) = max(V^t(s),V^t_min(s));
    return V^t ;
```

Algoritmo 4: REGRESS(V,s,S,a_1,a_2,R,K,γ,isMin)

input : V (value function vector), s (currently state), S (states set), a_1 (action choose by Player 1 in state s), a_2 (action choose by Player 2 in state s), R (reward for the players), K (credal transition set), γ (discount factor), isMin (flag)
output: Q (discounted total reward)

```
begin
    if isMin = true then
        //Choose p ∈ K such as
        Q = R(s,a_1,a_2) + γ min_{p∈K} Σ_{s'∈S} p(s'|s,a_1,a_2)V(s');
    else
        //Choose p ∈ K such as
        Q = R(s,a_1,a_2) + γ max_{p∈K} Σ_{s'∈S} p(s'|s,a_1,a_2)V(s');
    return Q ;
```

Após ocorrerem as N iterações, SolveTwo devolve V^* para Value-Iteration-AMG-IP.

5. Finalmente, Value-Iteration-AMG-IP retorna o valor de equilíbrio V^* do jogo.

6. Trabalhos relacionados

Jogos markovianos alternados de dois jogadores de soma zero são amplamente estudados na literatura de teoria dos jogos, sendo que os resultados fundamentais da área foram obtidos a partir de trabalhos em MDPs (*Markov Decision Processes*) [Puterman 1994]. [Kurano et al. 1998] generalizou o modelo MDP em que a probabilidade de transição varia a cada estágio e demonstrou que seu modelo converge para uma solução ótima. O modelo proposto por Kurano está relacionado a um MDP-IP (*Markov Decision Process with Imprecise Probabilities*) [Satia and Lave Jr. 1973, White III and Eldeib 1994, Delgado et al. 2011] e a um BMDP (*Bounded-parameter Markov Decision Processes*) [Givan et al. 2000]. [Chang 2006] estendeu o trabalho de Kurano para resolver AMGs de dois jogadores de soma zero com imprecisão nas probabilidades dadas por intervalos e provou que é possível encontrar políticas de equilíbrio para esse jogo. A nossa extensão foi inspirada no modelo MDP-IP [Delgado et al. 2011]. Sendo este um modelo mais geral que os modelos MDP e BMDP, ao estendermos o MDP-IP para dois jogadores, definimos um tipo mais geral de AMG, que chamamos de AMG-IP. Um trabalho que também inclui imprecisão em um jogo de dois jogadores é [Quaeghebeur and de Cooman 2009] em que o principal objetivo é fazer com que o jogador 1 aprenda a estratégia do jogador 2 (adversário) usando o modelo de Dirichlet impreciso para representar e atualizar as crenças do jogador 1. Esse trabalho trata apenas jogos simultâneos com políticas probabilísticas, enquanto que estamos interessados em jogos alternados com políticas determinísticas.

7. Conclusão

A teoria dos jogos interpreta qualquer ambiente multi-agente como um jogo, desde que o impacto de cada agente sobre os outros seja significativo, independentemente dos agentes serem cooperativos ou competitivos [Russell and Norvig 2003]. Assim, este trabalho está relacionado a área de sistemas multi-agentes, visto que, propõe uma variante de Jogos Markovianos Alternados com dois jogadores de soma zero (AMGs) que permite tratar imprecisão em probabilidades de transição expressa por conjuntos credais.

Para esse novo modelo, definimos uma solução que estende a solução proposta por Chang [Chang 2006] e mostramos como encontrar políticas de equilíbrio diante da imprecisão sobre a função de transição probabilística. Além disso, definimos um algoritmo, Value-Iteration-AMG-IP, que implementa a solução proposta.

Como trabalho futuro, pretendemos construir uma solução fatorada para o modelo AMG-IP (assim como foi construído no trabalho de [Delgado 2010] para o modelo MDP-IP), que permita resolver de maneira eficiente AMG-IPs com um grande número de estados.

Agradecimentos

Este trabalho recebeu apoio financeiro do CNPq e FAPESP (projeto temático 2008/03995-5).

Referências

Chang, H. S. (2006). Perfect information two-person zero-sum Markov games with imprecise transition probabilities. In *Mathematical Methods of Operations Research*, pages (64)335–351. Springer-Verlag.

Cozman, F. G. (2000). Credal networks. In *AI Journal*, pages 120(2):199–233.

Delgado, K. V. (2010). *Processos de decisão Markovianos fatorados com probabilidades imprecisas.* PhD thesis, IME-USP, Brasil.

Delgado, K. V., de Barros, L. N., Cozman, F. G., and Shirota, R. (2009). Representing and solving factored Markov decision processes with imprecise probabilities. In *ISIPTA*, pages 169–178. Durham, United Kingdom.

Delgado, K. V., Sanner, S., and de Barros, L. N. (2011). Efficient solutions to factored MDPs with imprecise transition probabilities. In *AI Journal*. Accepted 03 January 2011.

Ferguson, T. S. (Accesso em 18 fevereiro de 2011.). *Game Theory.* http://www.math.ucla.edu/t̃om/Game_Theory/ Contents.html.

Fudenberg, D. and Tirole, J. (1991). *Game Theory.* Cambridge, MA.

Givan, R., Leach, S., and Dean, T. (2000). Bounded-parameter Markov decision processes. In *AI Journal*, pages (39)71–109.

Kurano, M., Song, J., Hosaka, M., and Huang, Y. (1998). Controlled Markov set-chains with discounting. In *J. Appl. Prob.*, pages (35)293–302.

Littman, M. L. (1996). *Algorithms for Sequential Decision Making.* PhD thesis, Department of Computer Science - Brown University.

Osborne, M. J. and Rubinstein, A. (1994). *A Course in Game Theory.* Cambridge, MA.

Owen, G. (1982). *Game Theory: Second Edition.* Orlando, Florida.

Parthasarathy, T. and Raghavan, T. E. S. (1971). *Some topics in two-person games.* New York.

Puterman, M. L. (1994). *Markov Decision Processes.* New York.

Quaeghebeur, E. and de Cooman, G. (2009). Learning in games using the imprecise Dirichlet model. In Inc., E. S., editor, *Int. J. Approx. Reasoning*, pages (50)243–256. New York, NY, USA.

Russell, S. J. and Norvig, P. (2003). *Artificial Intelligence - A modern Approach.* Prentice-Hall, New Jersey.

Satia, J. K. and Lave Jr., R. E. (1973). Markovian decision processes with uncertain transition probabilities. In *Operations Research*, pages (21)728–740.

Shapley, L. S. (1953). Stochastic games. In *Proceedings of the National Academy of Sciences*, pages 39 (10):1095–1100.

Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities.* London.

White III, C. C. and Eldeib, H. K. (1994). Markov decision processes with imprecise transition probabilities. In *Operations Research*, pages (42)739–749.