Cap17 - Tomada de Decisões Complexas

- ✓ **Processos de Decisão de Markov**
  - ❖ ***Algoritmo de Iteração de Valor***
  - ❖ ***Algoritmo de Iteração de Política***

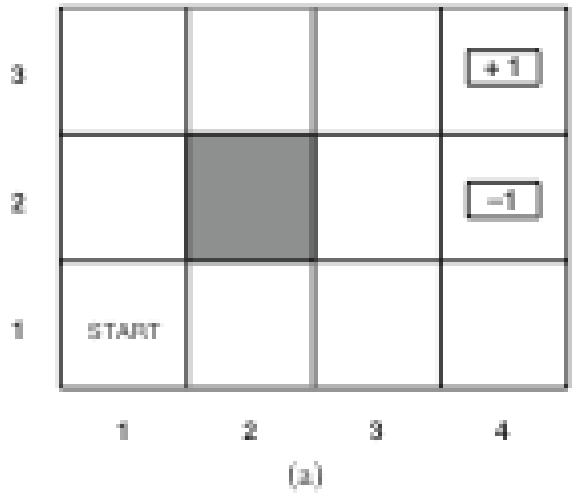
- ✓ **Processos de Decisão de Markov Parcialmente Observáveis**

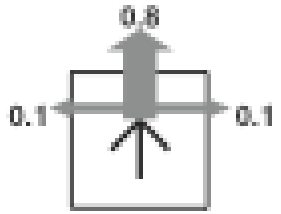
- ✓ **Teoria de Jogos**
  - ❖ **Projeto de Agentes**
  - ❖ **Projeto de Mecanismos**

Problema de decisão sequencial

- a utilidade do agente depende de uma **sequencia de decisões**
- **ambiente não determinístico**: o resultado da ação do agente é incerto.
- **modelo de transição**: especificação das probabilidades dos resultados de cada ação em cada estado possível
  - transições são Markovianas: probabilidade de alcançar S2 a partir de S1 depende somente de S1.

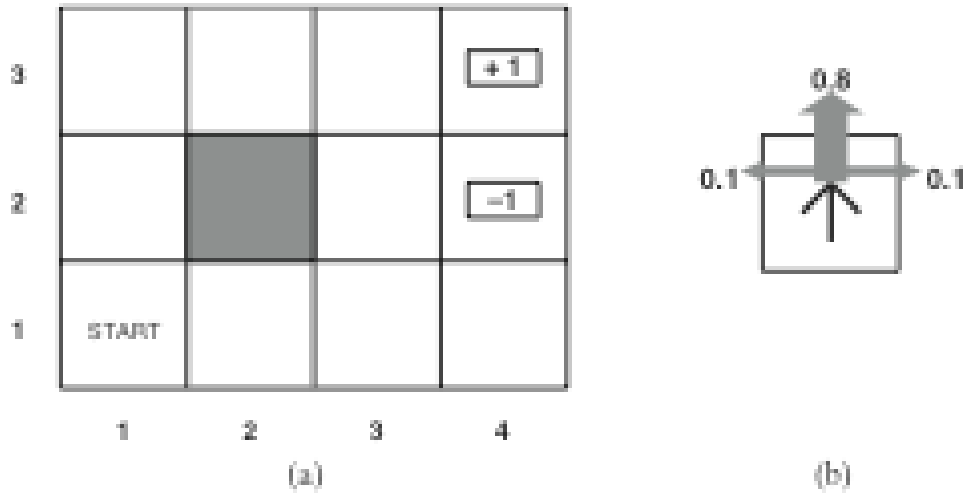
Definição de um Problema de Decisão Sequencial





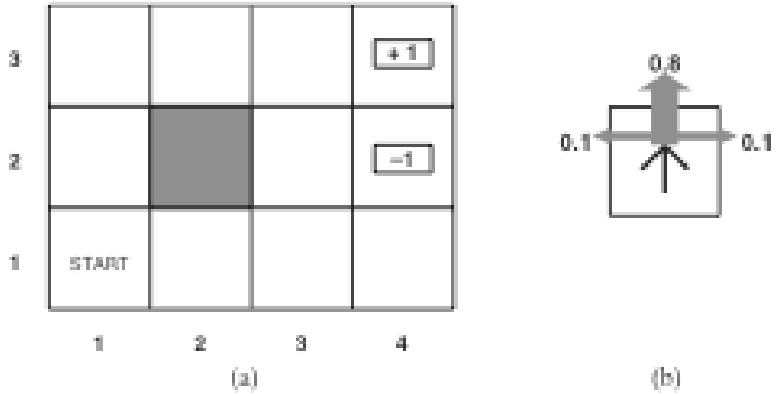
- Estado inicial: S0
- Modelo de transição: T(s, a, s'):
  - o resultado pretendido ocorre com prob. = 0.8,
  - com prob. 0.2 o agente se move em angulo reto cra direção pretendida
- Função de Recompensa: R(s):
  - estados terminais : +1 ou -1
  - outros estados: -0.04

Definição de um Problema de Decisão Sequencial



- Como o problema de decisão é sequencial, a função de utilidade dependerá de uma sequencia de estados (histórico de ambiente) em vez de depender de um único estado
- Em cada estado s, o agente recebe uma recompensa R(s), que pode ser positiva ou negativa
- A utilidade de um histórico de ambiente pode ser definida como a soma das recompensas definidas

Definição de um Problema de Decisão Sequencial



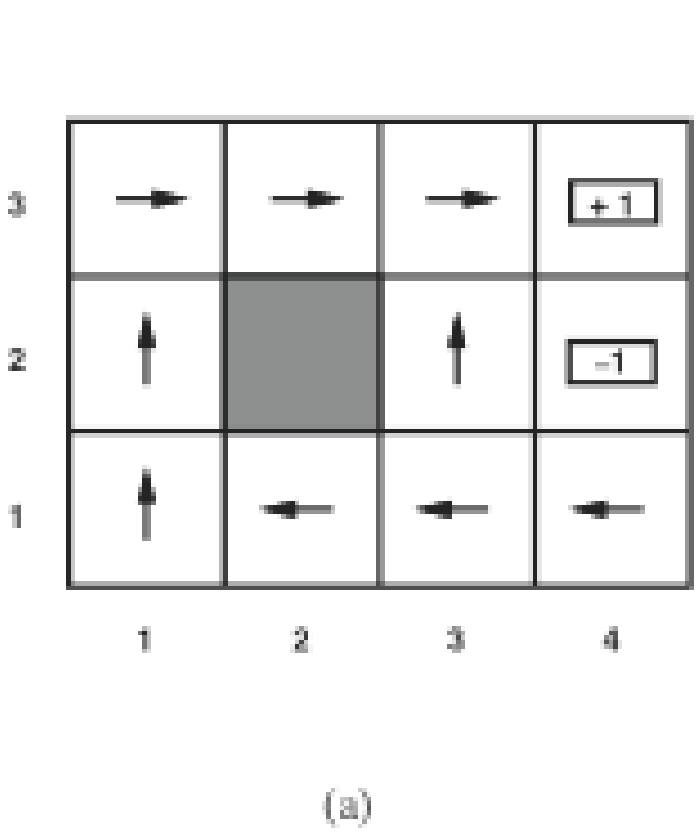
- Uma solução tem de especificar o que o agente deve fazer para qualquer estado que ele possa alcançar;
- Uma solução desse tipo é chamada **Política** (denotada por π(s))
- Se o agente tiver uma política completa, independente do resultado de suas ações, ele saberá o que fazer em seguida.
- Toda vez que uma dada política for executada a partir do estado inicial, a natureza estocástica do ambiente levará a um histórico diferente.
- Portanto a qualidade de uma política é medida pela utilidade esperada dos históricos de ambientes possíveis gerados por essa política

Política ótima

Política ótima π*(s): Política que maximiza a utilidade esperada, onde

- ✓ política é uma lista contendo, para cada estado s, a sua respectiva recomendação de ação e
- ✓ a função de utilidade mede a recompensa a longo prazo de uma política.

Políticas ótimas

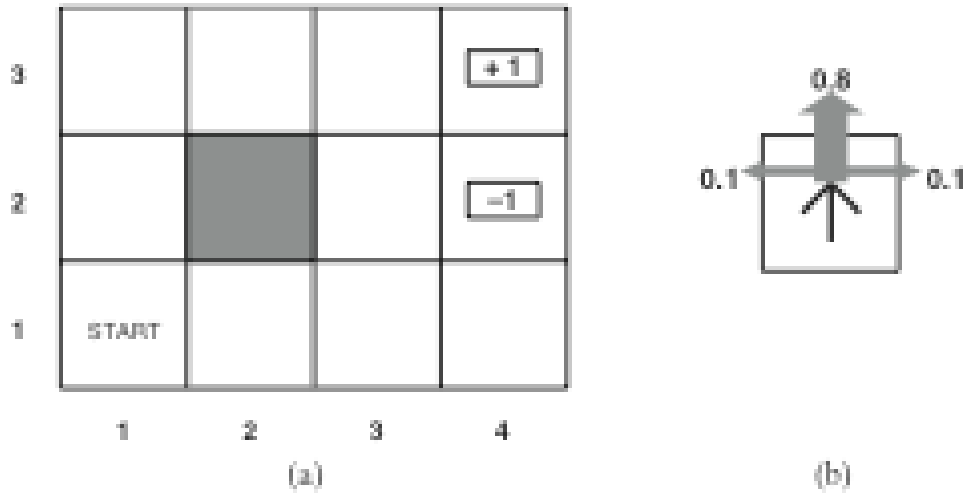


- O equilíbrio entre risco e recompensa muda dependendo do valor de R(s) para os estados não terminais
- A manutenção de um equilíbrio cuidadoso entre risco e recompensa é uma característica dos PDMs que não surge em problemas de busca determinística.

Escolhas para a medida de desempenho

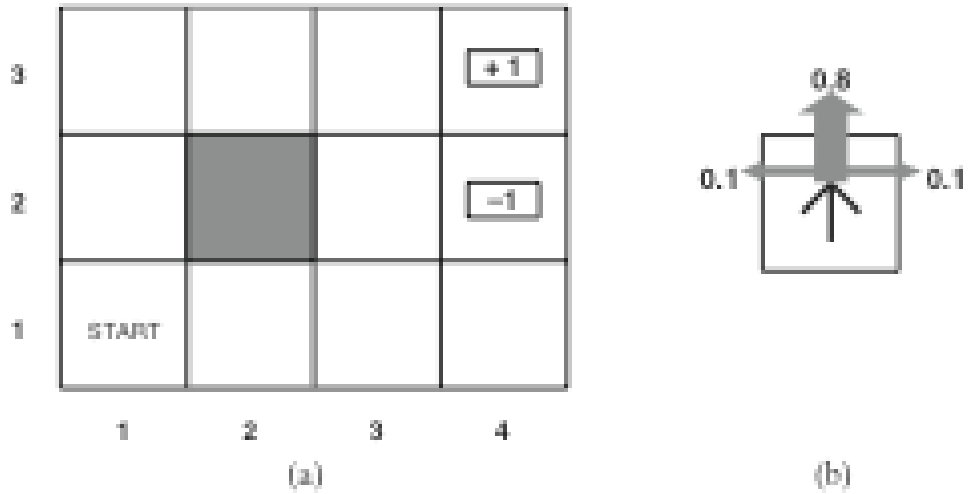
- Existe um horizonte finito ou infinito para a tomada de decisão:
  - horizonte finito: existe um tempo fixo N depois do qual dada importa: GAME OVER
  - horizonte infinito: existe tempo para seguir uma rota segura

Existe um horizonte finito ou infinito



- ex. Agente em (3,1) e N=3.
  - para ter chance de alcançar o estado +1, o agente deve ir diretamente para ele (i.e. ir para cima mesmo arriscando cair em -1)
- **com horizonte finito, a ação ótima em um dado estado pode mudar com o passar do tempo: política não estacionária**

Existe um horizonte finito ou infinito



- se N=100 há tempo de encontrar uma trajetória mais segura

- **com horizonte infinito, a ação ótima depende apenas do estado atual: política ótima estacionária**

Como calcular a utilidade de sequencias de estados

- Podemos visualizar isso como um problema de teoria da utilidade multi-atributo
  - cada estado si é um atributo da sequencia de estados [s0,...,sn]
- precisamos fazer alguma suposição de independência de preferências
  - as preferências do agente entre sequencia de estados são estacionárias

- preferências estacionárias:
  - se [s0,s1,s2,...] e [s0,s1’,s2’,...] estas sequencias devem ser ordenadas por preferencia do mesmo modo que as sequencias [s1,s2,...] e [s1’,s2’,...]
- *Se preferir um futuro ao outro que comece amanha, entao voce deve preferir este futuro mesmo que comece hoje!*

- Sob esse carater estacionário, existem apenas duas maneiras de atribuir utilidades a sequencias:
  - recompensas aditivas:
    - $U([s0,..., sn]) = R(S0) + R(s1) + ...$
  - recompensas descontadas:
    - $U([s0,..., sn]) = R(S0) + gR(s1) + g^2R(s2)+....$
      - fator de desconto g é um número entre 0 e 1 e *descreve a preferencia de um agente por recompensas atuais sobre recompensas futuras.*

- o uso de recompensas descontadas apresenta o menor número de dificuldades na avaliação da seq de estados
  - a utilidade de uma sequencia infinita é finita

- uma política não gera uma seq. de estado, mas in intervalo inteiro de seq. de estado possíveis, cada uma com uma probabilidade diferente de acordo com o modelo de transição.
- o valor de uma política pode entao ser definido como a ***soma esperada*** de recompensas descontadas obtidas
  - expectativa sobre todas as possíveis seq. de estado dado que a política seja executada:
  - $\pi^* = \text{argmax } E[\Sigma\ g^t\ R(s_t) \mid \pi]$

Fundamentação do Algoritmo Iteração de Valor

Como calcular a utilidade de cada estado e como definir a política ótima?

- calcular a utilidade de cada estado e depois empregar as utilidades para selecionar uma ação ótima para cada estado.
- utilidade de um estado é a utilidade esperada das sequencias de estados que poderiam segui-lo.
- como as sequencias dependem da política executada, definimos $U^{\pi}(s)$ com relação a uma política específica

$$U^{\pi}(s) = E[R(s_0) + \gamma R(s_1) + \gamma^2 R(s_2) + \ldots \mid \pi, s_{0=s}]$$

- a utilidade verdadeira de um estado é a soma esperada de recompensas descontadas se o agente executa uma política otima

Fundamentação do Algoritmo Iteração de Valor

- A função de utilidade permite ao agente selecionar ações usando o princípio da utilidade máxima esperada:

$$\pi^*(s) = \text{argmax}_a \ \Sigma_{s'} T(s,a,s')U(s')$$

A utilidade de um estado é a recompensa imediata correspondente a este estado, somada à utilidade descontada esperada do próximo estado, supondo que o agente escolha a ação ótima:

$$U(s) = R(s) + \gamma \ \text{max}_\alpha \Sigma_{s'} T(s, \alpha, s') \ U(s')$$

Esta é a chamada Equação de Bellman

- As utilidades dos estados, como função de sequencias de estados subsequentes são soluções do conjunto de eq. De Bellman para todos os estados s.

Fundamentação do Algoritmo Iteração de Valor

3	0,812	0,868	0,918	1
2	0,762	-	0,66	-1
1	0,705	0,655	0,611	0,388
	1	2	3	4

Utilidade dos estados assumindo γ = 1 e R(s) = -0,04

U(1,1) = -0,04 + γ max{ 0,8U(1,2) + 0,1U(2,1) + 0,1U(1,1), Acima

0,9U(1,1) + 0,1U(1,2), Esquerda

0,9U(1,1) + 0,1U(2,1), Abaixo

0,8U(2,1) + 0,1U(1,2) + 0,1U(1,1) } Direita

Substituindo os números das utilidades do mundo de grades temos que *Acima* é a melhor ação

Algoritmo de iteração de valor

- Se houver n estados possíveis, teremos n equações de Bellman (uma para cada estado) com n incognitas (as funções de utilidade)
- Para encontrar a política ótima temos que resolver esse sistema de equações NÃO lineares

Algoritmo de iteração de valor

- Abordagem iterativa:
  - Valores iniciais arbitrários para as utilidades
  - Calcula-se o lado direito da eq. e inserimos no lado esquerdo
    - Atualizando a utilidade de cada estado a partir da utilidade de seus vizinhos
  - Repetimos esse processo até chegar ao equilibrio
  - A cada iteração aplicamos a atualização de Bellman:
    - $U_{i+1}(s) \leftarrow R(s) + \gamma \max_{\alpha} \Sigma_{s'} T(s, \alpha, s') U(s')$
    - Aplicando essa atualização com frequencia infinita garantimos que o equilibrio será alcançado
    - Os valores finais serão soluções para as eq. De Bellman

Descrição do Algoritmo Iteração de Valor

Função ITERAÇÂO-DE-VALOR (pdm, ϵ) retorna uma função de utilidade

Entradas: pdm, um PDM com: estados S, modelo de transição T, função de recompensa R, desconto γ

ϵ, o erro máximo permitido na utilidade de cada estado

Variáveis locais: U, U', vetores de utilidades para os estados em S, inicialmente zero

δ: a mudança máxima na utilidade de um estado em qualquer iteração

Repita

U ← U'; δ ← 0;

para cada estado s em S faça

$U'[s] \leftarrow R[s] + \gamma \max_{\alpha} \Sigma_{s'} T(s, \alpha, s')U[s']$

se $|U'[s] - U[s]| > \delta$ então $\delta \leftarrow |U'[s] - U[s]|$

até $\delta < \epsilon (1 - \gamma) / \gamma$

retornar U

Fundamentação do Algoritmo Iteração de Política

Como definir a política ótima simplificando o cálculo da função de utilidade?

A função MAX presente no algoritmo de iteração de valor é não linear ➔ Propriedades algébricas desejáveis não podem ser empregadas no algoritmo

Recursivamente,

Calcular, para um certa política, o valor da utilidade (sem usar o operador MAX) e,

Então, usar essa utilidade para aperfeiçoar a política selecionada,

Até que a política se estabilize.

$U_i(1,1) = 0,8U_i(1,2) + 0,1U_i(2,1) + 0,1U_i(1,1)$ Acima

$U_i(1,2) = 0,8U_i(1,3) + 0,2U_i(1,2)$ Acima

Descrição do Algoritmo Iteração de Política

Função ITERAÇÂO-DE-POLÍTICA (pdm) retorna uma política

Entradas: pdm, um PDM com: estados S, modelo de transição T

Variáveis locais: U, U', vetores de utilidades para os estados em S, inicialmente zero

π: um vetor de política indexado pelo estado, inicialmente aleatório

Repita

U ← AVALIAÇÃO-DE-POLÍTICA(π, U, PDM);

inalterado? ← VERDADEIRO

para cada estado s em S faça

se $\max_{\alpha} \Sigma_{s'} T(s, \alpha, s')U[s'] > \Sigma_{s'} T(s, \pi[s], s')U(s')$ então

$\pi[s] \leftarrow \text{argmax}_{\alpha} \Sigma_{s'} T(s, \alpha, s')U[s']$

inaltetado? ← Falso

até inalterado?

retornar P

Processos de Markov Parcialmente Observáveis - PDMPO

Incerteza no estado atual do agente, implica em que:

(i) executar **π(s)** é impraticável e

(ii) **π*(s)** e **U(s)** "não dependem apenas de s, mas também do quanto o agente sabe quando está em s"

Modelo de transição – **T (s, α, s')** e Função de Recompensa – **R(s)**: Já discutidas

Modelo de Observação – **O(s, o)**: probabilidade do agente observar o no estado s

Estado de crença – **b(s):** probabilidade atribuída ao estado real pelo estado de crença b

$b'(s') = \kappa\, O(s', o) \Sigma_s\, T(s, \alpha, s')\, b(s)$ – novo estado de crença resultante da observação de **o** após a execução de **α** a partir do estado **s**

Solução – PDMPO

$$\Theta(b, \alpha, b') = \Sigma_o [P(b' | o, \alpha, b) \Sigma_{s'} [O(s', o) \Sigma_s T(s, \alpha, s') b(s)]]$$

- modelo de transição: *probabilidade de alcançar* ***b'*** *a partir de* ***b****, dado a ação* ***α***

$$\rho(b) = \Sigma_s b(s)R(s)$$

- função de recompensa para estados de crença

PDMPO = PDM no espaço completamente observável de *estados de crença*

✓ $\pi^*(b) = \pi^*(s)$ (resultado que pode ser comprovado)

Algoritmo de iteração de valor pode ser empregado (mas precisa ser adaptado, pois o espaço de estado de b é contínuo)

Em PDMPO a ação ótima depende só do estado de crença atual do agente.

Decisões com vários agentes: Teoria de Jogos

- Se a fonte de incerteza são os outros agentes?
- Se as decisões desses agentes forem influenciadas por nossas decisões?
- Teoria de jogos: usada em situações de tomada de decisão quando existem mais de um agente. Aplicações:
  - Relatórios de falência; leilão de espéctros de frequência sem fios; desenvolvimento e cotação de produtos de defesa, etc...

Duas maneiras de se utilizar teoria de jogos

- Projeto de agentes
  - Analisar decisões e calcular a utilidade esperada para cada decisão
- Projeto de mecanismos
  - Definir regras no ambiente de forma que o bem coletivo de todos os agentes seja maximizado quando cada agente adotar a solução que maximiza sua própria utilidade

Single move games

- Todos os jogadores atuam simultaneamente (i.e. nenhum jogador tem conhecimento sobre as escolhas do outros);
- Componentes:
  - Jogadores: quem toma as decisões
  - Ações: o que os jogadores podem escolher
  - *Payoff function* (matriz de recompensa): fornece a utilidade de cada jogador para cada combinação das ações de todos os jogadores:

	O: one	O: two
E: one	E=+2, O = -2	E=-3, O = +3
E: two	E=-3, O = +3	E=+4, O = -4

Single move games

- Cada jogador deve adotar e executar uma **estratégia** (i.e. uma **política**)
  - Est. pura: é uma política determinística
  - Est. mista: é uma política estocástica que seleciona uma ação **a** com probabilidade **p** e ação **b** com prob. **1-p**:
    - i.e.: **[p:a; (1-p):b]**
    - E.g. **[0.5: one; 0.5: two]**

- Perfil de estratégia: uma associação de estratégia com cada jogador
  - Dado um PE, a saída do jogo é um valor numérico para cada jogador.
- Solução: um perfil de estratégia em que cada jogador adota uma estratégia ***racional***
  - A questão mais importante de teoria de jogos é definir o que "racional significa" quando cada agente utiliza parte do perfil para determinar um resultado

Exemplo: dilema do prisioneiro

- Video:
- http://www.youtube.com/watch?v=p3Uos2fzIJ0&feature=player_embedded
  - Qual é a payoff function (matriz de recompensa)?
  - Há alguma solução de equilíbrio?
    - (em que nenhum jogador pode se beneficiar se mudar de estratégia)
  - Quais são os tipos de estratégia possíveis?

Teoria de Jogos – Dilema do Prisioneiro

	Alice	
Bob	Testemunha	Não Testemunha
Testemunha	A = -5, B = -5	A = -10, B = 0
Não Testemunha	A = 0, B = -10	A = -1, B = -1

Para Alice, testemunhar é a melhor opção independentemente da escolha de Bob

- "*suponha que Bob testemunhe, então -5 se Alice testemunhar ou -10 caso contrário; se Bob não testemunhar, então 0 se Alice testemunhar ou -1 caso contrário*"

Bob pensa da mesma maneira, o que leva ao equilíbrio (Testemunhar, Testemunhar)

Testemunhar é uma estratégia dominante para o jogo

Teoria de Jogos – Dilema do Prisioneiro

	Alice	
Bob	Testemunha	Não Testemunha
Testemunha	A = -5, B = -5	A = -10, B = 0
Não Testemunha	A = 0, B = -10	A = -1, B = -1

•Uma estratégia s para um jogador p **domina fortemente** a estratégia s' se o resultado correspondente a s é melhor para p que o resultado correspondente a s', considerando-se todas as escolhas de estratégias pelos outros jogadores.

• s fracamente domina s' se s é melhor do que s' em pelo menos um perfil de estratégia e não é pior em nenhum outro

•Uma estratégia dominante é aquela que domina todas as outras

- Jogador racional: escolhe a estratégia **dominante**
- Um resultado é **ótimo de pareto** se não há nenhum outro resultado que os jogadores prefiram

- *Alice e Bob são racionais, então ambos vão preferir a solução ótima de pareto: "prestar testemunho*"
- Quando dois jogadores possuem uma estratégia dominante, a combinação delas é chamada: **equilíbrio de estratégia dominante**

Equilíbrio de Nash

- Um perfil de estratégia forma um **equilíbrio** se nenhum jogador se beneficiar de troca de estratégia
  - Mínimo local no espaço de políticas
- "*todo jogo possui pelo menos um equilíbrio*" (Nash)
- (testemunhar, testemunhar) é uma solução de equilíbrio...
  - Mas não é pareto ótima

Dilema

- O resultado de equilíbrio é pior que a escolha (recusar, recusar).
  - i.e. (testemunhar, testemunhar) é solução **dominada de pareto** por (-1,-1) de (recusar, recusar)

Dilema

- Há como chegar em (-1,-1)?
  - Certamente, mas é dificil de imaginar um agente racional *arriscar* a pena de 10 anos!
  - Este é o *poder de atração* do equilíbrio

Escapando do Dilema

- Mudar o jogo:
  - Jogo de repetição
  - Assumir crenças morais nos agentes
    - Cooperação e integridade
    - i.e., assumir uma função de utilidade diferente, com uma nova *payoff function*, criando um novo jogo.

Teoria de Jogos – Jogos de *Soma Zero (a soma dos ganhos é zero)*

		Jogador O	
		O:one	O:two
Jogador E	E:one	E= +2 O= -2	E = -3, O=+3
	E:two	E = -3, O =+3	E =+4, O =-4

➔

Resultado Par	E = 1, O = -1
Resultado Ímpar	E = -1, O = 1

Cada jogador se beneficia de um resultado ➔ Com estratégia pura, não há equilíbrio.

Porém, por meio de estratégia mista, o método maximin encontra o ponto de equilíbrio.

– estratégia mista: escolher a ação ***a*** com probabilidade ***p*** ou a ação ***b*** com prob. ***1-p:***

- ***[p:a, (1-p):b]***
- No par ou impar de dois dedos:
  - ***[0.5:one, 0.5:two]***

Teoria de Jogos – Equilíbrio de Nash

	Acme	
Best	bluray	CD
bluray	A = 9, B = 9	A = -4, B = -1
CD	A = -3, B = -1	A = 5, B = 5

- Não há estratégias dominantes
- Mas há dois equilíbrios de Nash: (bluray,bluray) e (CD, CD)
  - pois se um jogador mudar sua estratégia unilateralmente para uma estratégia diferente, ficará em pior situação
- Solução:

Utilizar a solução de equilíbrio ótima de Pareto: (bluray,bluray)

Cada jogo possui uma solução ótima de Pareto, mas pode ter várias ou não ter pontos de equilíbrio (e.g. Se (bluray,bluray) = (5,5))

Teoria de Jogos – Jogos de Soma Zero

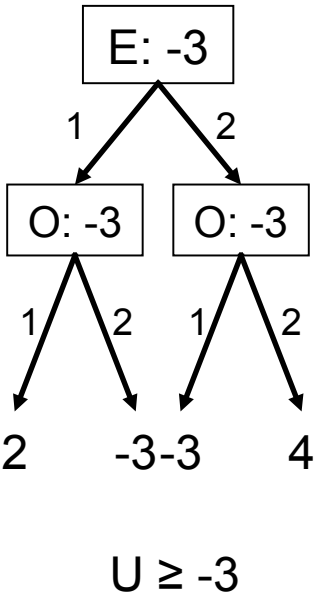
		Jogador O	
		O:one	O:two
Jogador E	E:one	E= +2 O= -2	E = -3, O=+3
	E:two	E = -3, O =+3	E =+4, O =-4

➔

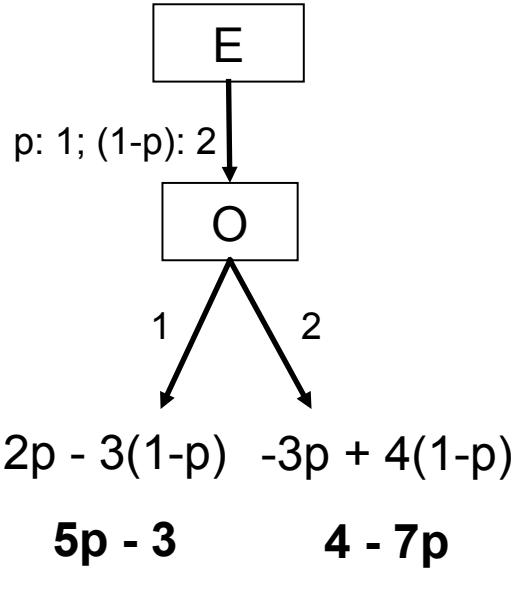
Resultado Par	E = 1, O = -1
Resultado Ímpar	E = -1, O = 1

Cada jogador se beneficia de um resultado ➔ Com estratégia pura, não há equilíbrio.

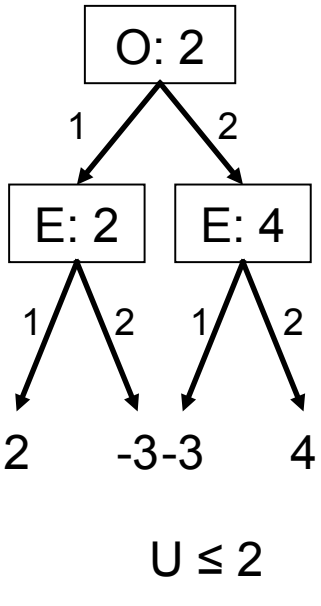
Porém, por meio de estratégia mista, o método maximin encontra o ponto de equilíbrio.



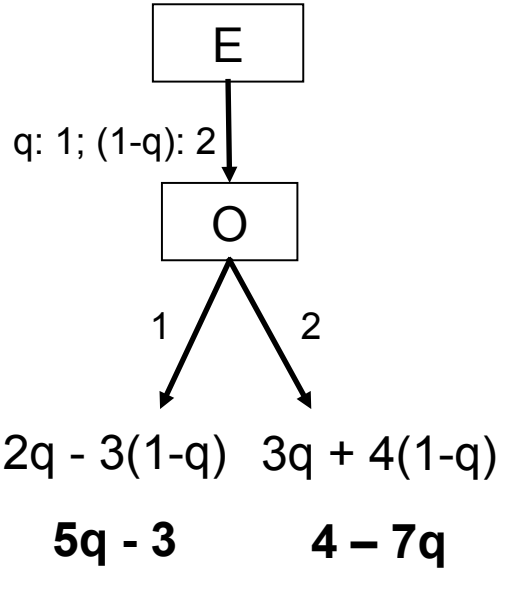
U ≥ -3



5p - 3 **4 - 7p**



U ≤ 2



5q - 3 **4 – 7q**

Teoria de Jogos – Jogos de Soma Zero

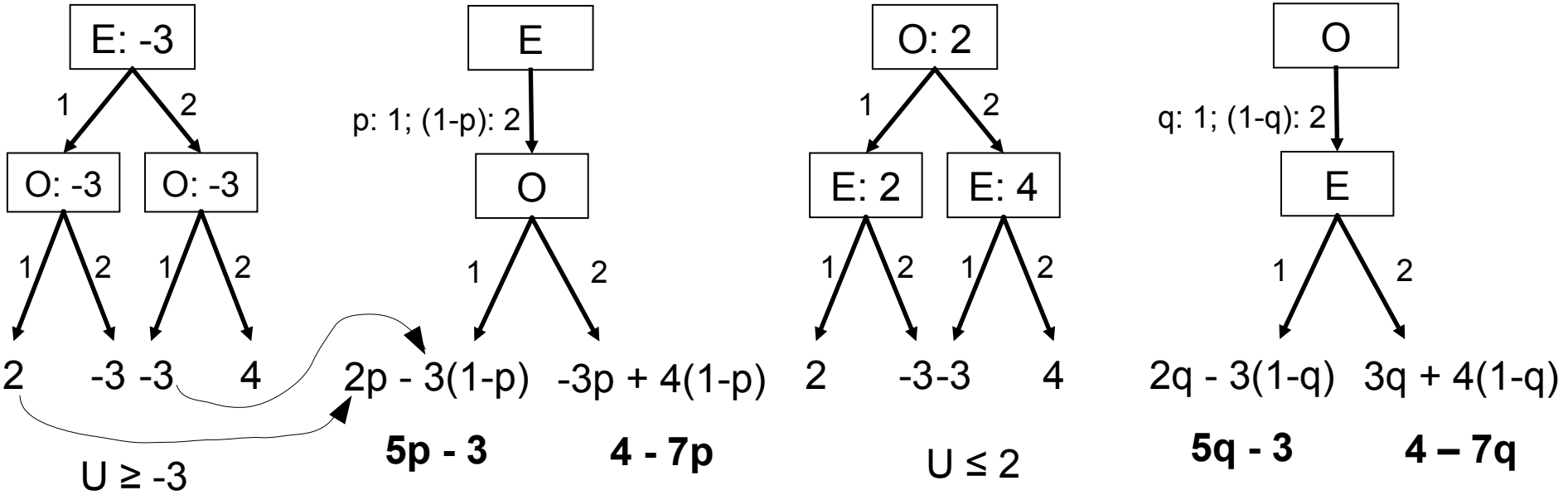
		Jogador O	
		O:one	O:two
Jogador E	E:one	E= +2 O= -2	E = -3, O=+3
	E:two	E = -3, O =+3	E =+4, O =-4

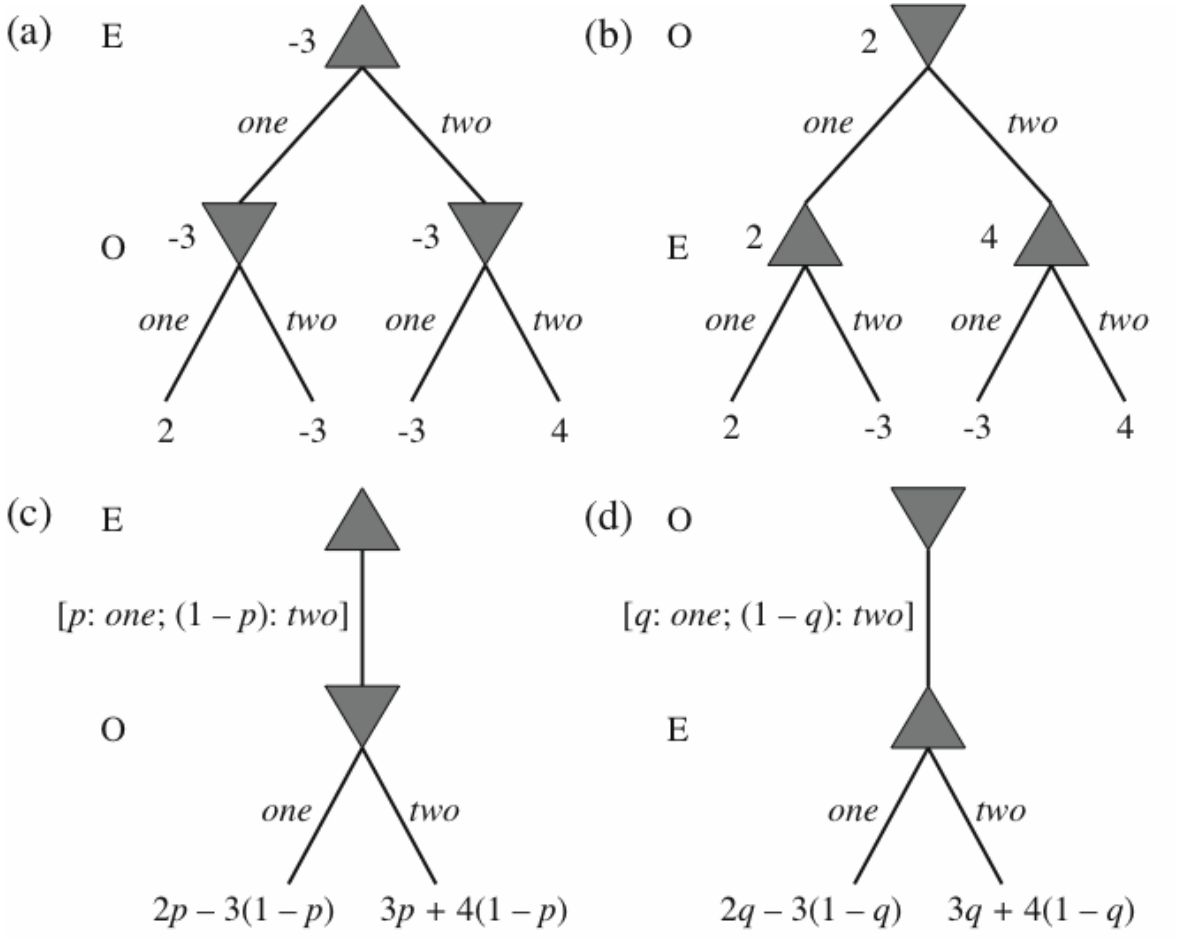
➔

Resultado Par	E = 1, O = -1
Resultado Ímpar	E = -1, O = 1

Cada jogador se beneficia de um resultado ➔ Com estratégia pura, não há equilíbrio.

Porém, por meio de estratégia mista, o método maximin encontra o ponto de equilíbrio.




(a)
E
-3
one
two
O
-3
-3
one
two
one
two
2
-3
-3
4
(b)
O
2
one
two
E
2
4
one
two
one
two
2
-3
-3
4
(c)
E
[p: one; (1 – p): two]
O
one
two
2p – 3(1 – p)
3p + 4(1 – p)
(d)
O
[q: one; (1 – q): two]
E
one
two
2q – 3(1 – q)
3q + 4(1 – q)
(e)
U
+4
+3
+2
+1
0
–1
–2
–3
two
one
p
1
(f)
U
+4
+3
+2
+1
0
–1
–2
–3
two
one
q
1

If E choses first, the situation is shown in fig. A. E chooses the strategy [p:one; (1-p):two] at the root, and then O chooses a pure strategy (and hence a move) given the value of p. If O chooses on, the expected payoff is -3p+4(1-p) = 4-7p. We can draw these two payoffs as straight lines on a graph, where p ranges from 0 to 1 on the x-axis, as shown in Fig. e. O, the minimizer, will always choose the lower of the two lines, as shown by the heavy lines in the figure. Therefore, the best that E can do at the root is to choose p to be at the intersection point:

$$5p-3 = 4 - 7p \quad \Rightarrow \quad p=7/12$$

The utility for E at this point is Ueo = -1/12

If O moves first, the situation is as shown in Fig. b. O chooses the strategy [q:one; (1-q):two] at the root, and then E chooses a move given the value of q. The payoffs are 2q-3(1-q) = 5q -3 and -3q+4(1-q) = 4-7q. Again, the best that O can do is to choose the intersection point:

$5q-3 = 4-7q \quad \Rightarrow \quad q= 7/12$

The utility of E at this point is Uoe = - 1/12

→ the true utility is attained by the mixed strategy [7/12:one, 5/12:two], this is called the *maximin equilibrium* of the game and is a Nash equilibrium

Result by von Neummann:

“every two-player game has a maximin equilibrium when you allow mixed strategies”