Gry z jedną turą i procesy decyzyjne Markowa

Paweł Rychlikowski

Instytut Informatyki UWr

18 kwietnia 2023

Gry z jedną turą

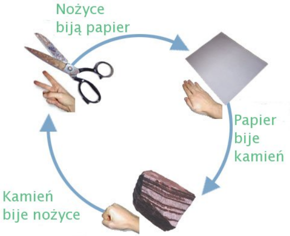
- Powiemy sobie trochę o grach z jedną turą
- Ale takich, w których gracze podejmują swoje decyzje jednocześnie

Gry z jedną turą

- Powiemy sobie trochę o grach z jedną turą
- Ale takich, w których gracze podejmują swoje decyzje jednocześnie

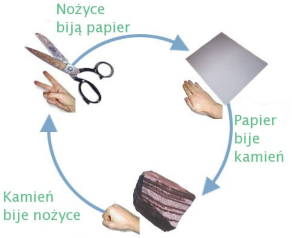
Rozważamy gry z sumą zerową.

Papier, nożyce, kamień



Źródło: Wikipedia

Papier, nożyce, kamień



Źródło: Wikipedia

Macierz wypłat

Grę definiuje **macierz wypłat**. Przykładowo poniżej dla P-N-K

Max/Min	Papier	Nożyce	Kamień
Papier	0	-1	+1
Nożyce	+1	0	-1
Kamień	-1	+1	0

Strategie

- Czysta strategia: zawsze akcja a

Strategie

- Czysta strategia: zawsze akcja a
- Mieszana strategia: rozkład prawdopobieństwa na akcjach

PNK – spostrzeżenia

- **Oczywisty fakt:** każdą strategię stałą można pokonać (też stałą strategią)

PNK – spostrzeżenia

- **Oczywisty fakt:** każdą strategię stałą można pokonać (też stałą strategią)
- **Fakt 1:** każdą strategię mieszaną można (prawie) pokonać za pomocą strategii stałej:

PNK – spostrzeżenia

- **Oczywisty fakt:** każdą strategię stałą można pokonać (też stałą strategią)
- **Fakt 1:** każdą strategię mieszaną można (prawie) pokonać za pomocą strategii stałej:
  Mój przeciwnik gra losowo, ale z przewagą kamienia – zatem ja daję **zawsze** papier

PNK – spostrzeżenia

- **Oczywisty fakt:** każdą strategię stałą można pokonać (też stałą strategią)
- **Fakt 1:** każdą strategię mieszaną można (prawie) pokonać za pomocą strategii stałej:
  Mój przeciwnik gra losowo, ale z przewagą kamienia – zatem ja daję **zawsze** papier
- **Fakt 2:** Optymalna strategia jest mieszana (w tej grze każde z $p = \frac{1}{3}$)

PNK – spostrzeżenia

- **Oczywisty fakt:** każdą strategię stałą można pokonać (też stałą strategią)
- **Fakt 1:** każdą strategię mieszaną można (prawie) pokonać za pomocą strategii stałej:
  Mój przeciwnik gra losowo, ale z przewagą kamienia – zatem ja daję **zawsze** papier
- **Fakt 2:** Optymalna strategia jest mieszana (w tej grze każde z $p = \frac{1}{3}$)
- **Fakt 3:** Znajomość optymalnej strategii mieszanej gracza A, nie daje żadnej przewagi graczowi B (i odwrotnie)

PNK – uwagi końcowe

- W prawdziwym P-N-K dochodzi kilka innych aspektów:

PNK – uwagi końcowe

- W prawdziwym P-N-K dochodzi kilka innych aspektów:
  - Grają ludzie, którzy nie potrafią realizować losowości,

PNK – uwagi końcowe

- W prawdziwym P-N-K dochodzi kilka innych aspektów:
  - Grają ludzie, którzy nie potrafią realizować losowości, Który człowiek (nie dysponując kostką do gry), przegrawszy 3 razy z rzędu jako papier pokaże papier?

PNK – uwagi końcowe

- W prawdziwym P-N-K dochodzi kilka innych aspektów:
  - Grają ludzie, którzy nie potrafią realizować losowości, Który człowiek (nie dysponując kostką do gry), przegrawszy 3 razy z rzędu jako papier pokaże papier?
  - za to wysyłają swoimi ciałami różne informacje, które można analizować

PNK – uwagi końcowe

- W prawdziwym P-N-K dochodzi kilka innych aspektów:
  - Grają ludzie, którzy nie potrafią realizować losowości,
    Który człowiek (nie dysponując kostką do gry), przegrawszy 3 razy z rzędu jako papier pokaże papier?
  - za to wysyłają swoimi ciałami różne informacje, które można analizować
- Zatem ma sens organizowanie zawodów w PNK

PNK – uwagi końcowe

- W prawdziwym P-N-K dochodzi kilka innych aspektów:
  - Grają ludzie, którzy nie potrafią realizować losowości,
    Który człowiek (nie dysponując kostką do gry), przegrawszy 3 razy z rzędu jako papier pokaże papier?
  - za to wysyłają swoimi ciałami różne informacje, które można analizować
- Zatem ma sens organizowanie zawodów w PNK
- Sens miałyby również zawody ludzko-komputerowe, realizowane on-line (agent musiałby zgadnąć, czy gra z człowiekiem, czy z maszyną i czy opłaca się próbować zgadnąć model losowania używany przez człowieka)

Gra w zgadywanie (Morra 2)

Gra w zgadywanie (Morra 2)

- Mamy dwóch graczy:
  - A) Zgadywacz
  - B) Zmyłek

  którzy na sygnał pokazują 1 lub 2 palce.

Gra w zgadywanie (Morra 2)

- Mamy dwóch graczy:
  A) Zgadywacz
  B) Zmyłek

  którzy na sygnał pokazują 1 lub 2 palce.
- Jeżeli Zgadywacz nie zgadnie (pokazał coś innego niż Zmyłek), daje Zmyłkowi 3 dolary.

Gra w zgadywanie (Morra 2)

- Mamy dwóch graczy:
  1. A) Zgadywacz
  2. B) Zmyłek

  którzy na sygnał pokazują 1 lub 2 palce.
- Jeżeli Zgadywacz nie zgadnie (pokazał coś innego niż Zmyłek), daje Zmyłkowi 3 dolary.
- Jeżeli Zgadywacz zgadnie, to dostaje od Zmyłka:

Gra w zgadywanie (Morra 2)

- Mamy dwóch graczy:
  A) Zgadywacz
  B) Zmyłek

  którzy na sygnał pokazują 1 lub 2 palce.
- Jeżeli Zgadywacz nie zgadnie (pokazał coś innego niż Zmyłek), daje Zmyłkowi 3 dolary.
- Jeżeli Zgadywacz zgadnie, to dostaje od Zmyłka:
  - jak pokazali 1 palec, to 2 dolary
  - jak pokazali 2 palce, to 4 dolary

Gra w zgadywanie (Morra 2)

- Mamy dwóch graczy:
  A) Zgadywacz
  B) Zmyłek

  którzy na sygnał pokazują 1 lub 2 palce.
- Jeżeli Zgadywacz nie zgadnie (pokazał coś innego niż Zmyłek), daje Zmyłkowi 3 dolary.
- Jeżeli Zgadywacz zgadnie, to dostaje od Zmyłka:
  - jak pokazali 1 palec, to 2 dolary
  - jak pokazali 2 palce, to 4 dolary

Pytanie

Jak grać w tę grę? (prośba o podanie wstępnych intuicji)

Macierz wypłat

Definicja

Taką grę zadajemy za pomocą macierzy wypłat, w której $V_{a,b}$ jest wynikiem gry z punktu widzenia pierwszego gracza.

Macierz wypłat

Definicja

Taką grę zadajemy za pomocą macierzy wypłat, w której $V_{a,b}$ jest wynikiem gry z punktu widzenia pierwszego gracza.

Nasza gra:

Zg/Zm	1 palec	2 palce
1 palec	2	-3
2 palce	-3	4

Proste fakty

- Jak Zmyłek będzie grał cały czas to samo, to Zgadywacz wygra każdą turę (i odwrotnie)

Proste fakty

- Jak Zmyłek będzie grał cały czas to samo, to Zgadywacz wygra każdą turę (i odwrotnie)
- Muszą zatem stosować strategie mieszane, ale jakie?

Wartość gry

Definicja

Wartość gry dla dwóch strategii graczy jest równa:

$$V(\pi_A, \pi_B) = \sum_{a,b} \pi_A(a)\pi_B(b)V(a, b)$$

Przykładowo: Zgadywacz zawsze zgaduje 1, Zmyłek wybiera akcję losowo z prawdopodobieństwem 0.5.

Wartość gry

Definicja

Wartość gry dla dwóch strategii graczy jest równa:

$$V(\pi_A, \pi_B) = \sum_{a,b} \pi_A(a)\pi_B(b)V(a, b)$$

Przykładowo: Zgadywacz zawsze zgaduje 1, Zmyłek wybiera akcję losowo z prawdopodobieństwem 0.5.
Wynik: $-\frac{1}{2}$ (tak samo często zyskuje 2 jak traci 3 dolary)

Strategia mieszana vs czysta

Uwaga

Jeżeli gracz A zapowie, że będzie grał strategią mieszaną (i ją poda), wówczas gracz B może grać strategią czystą (i osiągnie optymalny wynik).

Strategia mieszana vs czysta

Uwaga

Jeżeli gracz A zapowie, że będzie grał strategią mieszaną (i ją poda), wówczas gracz B może grać strategią czystą (i osiągnie optymalny wynik).

Dlaczego?

Strategia mieszana vs czysta

Uwaga

Jeżeli gracz A zapowie, że będzie grał strategią mieszaną (i ją poda), wówczas gracz B może grać strategią czystą (i osiągnie optymalny wynik).

Dlaczego?

Odpowiedź

- Możemy dla każdej akcji policzyć wartość oczekiwaną wypłaty

Uwaga

Jeżeli gracz A zapowie, że będzie grał strategią mieszaną (i ją poda), wówczas gracz B może grać strategią czystą (i osiągnie optymalny wynik).

Dlaczego?

Odpowiedź

- Możemy dla każdej akcji policzyć wartość oczekiwaną wypłaty
- i wybrać (dowolną) najlepszą akcję

Strategia mieszana vs czysta

Uwaga

Jeżeli gracz A zapowie, że będzie grał strategią mieszaną (i ją poda), wówczas gracz B może grać strategią czystą (i osiągnie optymalny wynik).

Dlaczego?

Odpowiedź

- Możemy dla każdej akcji policzyć wartość oczekiwaną wypłaty
- i wybrać (dowolną) najlepszą akcję
- (Jeżeli takich akcji jest więcej, wówczas można też dowolnie losować między nimi)

Gra w zgadywanie (Morra 2). Przypomnienie

Gra w zgadywanie (Morra 2). Przypomnienie

- Mamy dwóch graczy:
  - A) Zgadywacz
  - B) Zmyłek

  którzy na sygnał pokazują 1 lub 2 palce.
- Jeżeli Zgadywacz nie zgadnie (pokazał coś innego niż Zmyłek), daje Zmyłkowi 3 dolary.
- Jeżeli Zgadywacz zgadnie, to dostaje od Zmyłka:
  - jak pokazali 1 palec, to 2 dolary
  - jak pokazali 2 palce, to 4 dolary

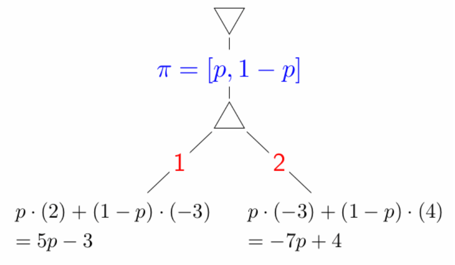
Znalezienie optymalnej strategii

Zaczyna gracz B – Zmyłek.

Wybiera strategię mieszaną z parametrem p

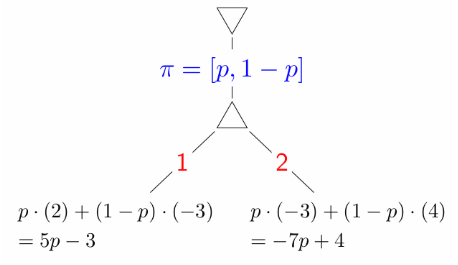
Znalezienie optymalnej strategii

Zaczyna gracz B – Zmyłek.
Wybiera strategię mieszaną z parametrem p



Znalezienie optymalnej strategii

Zaczyna gracz B – Zmyłek.
Wybiera strategię mieszaną z parametrem p

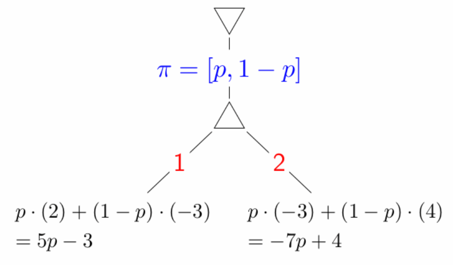


Wartość takiej gry to

$$\min_{p \in [0,1]} (\max(5p - 3, -7p + 4))$$

Znalezienie optymalnej strategii

Zaczyna gracz B – Zmyłek.
Wybiera strategię mieszaną z parametrem p

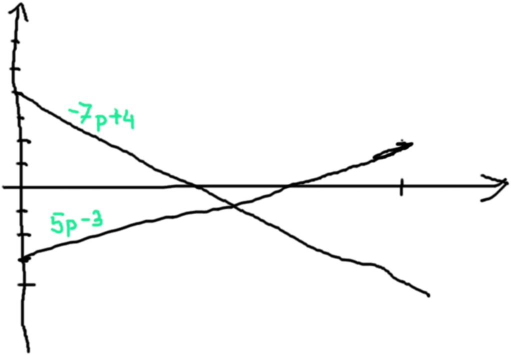


Wartość takiej gry to

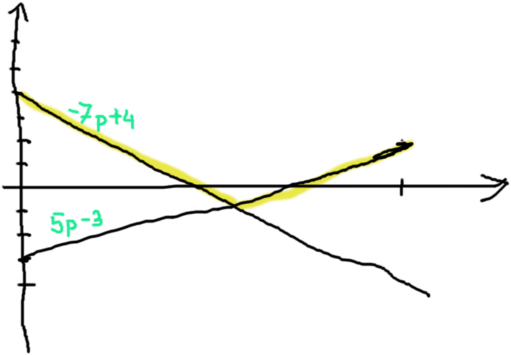
$$\min_{p \in [0,1]} \left(\max(5p - 3, -7p + 4) \right)$$

Zauważmy, dla jakich p wygrywa lewe, dla jakich prawe i co z tego wynika.

Optymalna strategia. Wykresy



Optymalna strategia. Wykresy



Znalezienie optymalnej strategii (2)

- W powyższej grze, Zmyłek osiągnie najlepszy wynik, gdy przyjmie $p = \frac{7}{12}$, wynik ten to $-\frac{1}{12}$

Znalezienie optymalnej strategii (2)

- W powyższej grze, Zmyłek osiągnie najlepszy wynik, gdy przyjmie $p = \frac{7}{12}$, wynik ten to $-\frac{1}{12}$
- Ok, on zaczynał, miał trudniej – a gdyby zaczynał Zgadywacz? I podał swoją strategię mieszaną?

Znalezienie optymalnej strategii (2)

- W powyższej grze, Zmyłek osiągnie najlepszy wynik, gdy przyjmie $p = \frac{7}{12}$, wynik ten to $-\frac{1}{12}$
- Ok, on zaczynał, miał trudniej – a gdyby zaczynał Zgadywacz? I podał swoją strategię mieszaną?

Wynik gry

Wynik jest dokładnie taki sam, czyli $-\frac{1}{12}$!

Twierdzenie von Neumana

Twierdzenie, von Neuman, 1928

Dla każdej jednoczesnej gry dwuosobowej o sumie zerowej ze skończoną liczbą akcji mamy:

$$\max_{\pi_A} \min_{\pi_B} V(\pi_A, \pi_B) = \min_{\pi_B} \max_{\pi_A} V(\pi_A, \pi_B)$$

dla dowolnych mieszanych polityk π_A, π_B.

Twierdzenie von Neumana

Twierdzenie, von Neuman, 1928

Dla każdej jednoczesnej gry dwuosobowej o sumie zerowej ze skończoną liczbą akcji mamy:

$$\max_{\pi_A} \min_{\pi_B} V(\pi_A, \pi_B) = \min_{\pi_B} \max_{\pi_A} V(\pi_A, \pi_B)$$

dla dowolnych mieszanych polityk π_A, π_B.

- Można ujawnić swoją politykę optymalną!
- **Dowód**: pomijamy, programowanie liniowe, przedmiot J.B.
- Algorytm: programowanie liniowe

Gry wieloturowe

- Można o grze wieloturowej myśleć jako o grze jednoturowej
- Gracze na sygnał kładą przed sobą opis strategii (program)

Gry wieloturowe

- Można o grze wieloturowej myśleć jako o grze jednoturowej
- Gracze na sygnał kładą przed sobą opis strategii (program)

Uwaga

Optymalną strategią jest MiniMax (ExpectMiniMax w grach losowych). Ale wiedząc o strategii gracza różnej od optymalnej możemy oczywiście ugrać więcej.

Co pomijamy

- Gry o sumie niezerowej, w których dochodzi możliwość kooperacji.

Co pomijamy

- Gry o sumie niezerowej, w których dochodzi możliwość kooperacji.
- Punkt równawagi Nasha (jest zawsze para strategii, że żaden gracz nie chce jej zmienić, wiedząc, że ten drugi nie zmienia).

Co pomijamy

- Gry o sumie niezerowej, w których dochodzi możliwość kooperacji.
- Punkt równawagi Nasha (jest zawsze para strategii, że żaden gracz nie chce jej zmienić, wiedząc, że ten drugi nie zmienia). **Również dla gier o sumie niezerowej!**

Co pomijamy

- Gry o sumie niezerowej, w których dochodzi możliwość kooperacji.
- Punkt równowagi Nasha (jest zawsze para strategii, że żaden gracz nie chce jej zmienić, wiedząc, że ten drugi nie zmienia). **Również dla gier o sumie niezerowej!**
- Agent musi zdecydować, czy ma być miły dla innego agenta (i budować reputację przy wielu rozgrywkach, słynny dylemat więźnia).

Procesy decyzyjne Markowa (MDP)

Procesy decyzyjne Markowa (MDP)

- Coś pomiędzy grami a zwykłym zadaniem przeszukiwania
- (zwłaszcza jeżeli przypomnimy sobie gry z węzłami losowymi)
- a jednocześnie krok w stronę uczenia ze wzmocnieniem

MDP a przeszukiwanie

Standardowe przeszukiwanie

Znamy mechanikę świata i wiemy, że akcja w stanie da nam konretny rezultat (inny stan).

MDP a przeszukiwanie

Standardowe przeszukiwanie

Znamy mechanikę świata i wiemy, że akcja w stanie da nam konretny rezultat (inny stan).

MDP

Znamy mechanikę świata i wiemy, że akcja w stanie da nam pewien rozkład prawdopodobieństwa na następnych stanach.

MDP a przeszukiwanie

Standardowe przeszukiwanie

Znamy mechanikę świata i wiemy, że akcja w stanie da nam konretny rezultat (inny stan).

MDP

Znamy mechanikę świata i wiemy, że akcja w stanie da nam pewien rozkład prawdopodobieństwa na następnych stanach.

Nie wiemy, co dokładnie się stanie, ale wiemy co **może** się stać i z jakim prawdopodobieństwem.

Własność Markowa

- Przyszłość zależy od ostatniego stanu.

Własność Markowa

- Przyszłość zależy od ostatniego stanu.
- Nie zależy od historii...

Własność Markowa

- Przyszłość zależy od ostatniego stanu.
- Nie zależy od historii...
- Chyba, że jej fragment (o długości N) uznamy za część stanu.

Własność Markowa

- Przyszłość zależy od ostatniego stanu.
- Nie zależy od historii...
- Chyba, że jej fragment (o długości N) uznamy za część stanu.

Ważna uwaga

Zakładamy **skończoną** liczbę stanów

Uwaga na wulkany (1)

- Dobrze omawia się MDP na prostych światach na prostokątnej kratce.
- I od takich modeli zaczniemy.

Uwaga na wulkany (1)

- Dobrze omawia się MDP na prostych światach na prostokątnej kratce.
- I od takich modeli zaczniemy.

Generalnie myślimy na początku o przestrzeni stanów na tyle małej, że nie będzie kłopotów z pamiętaniem różnych wartości dla **każdego stanu**.

Uwaga na wulkany (2)

Volcano crossing

		-50	20
$\boxed{}$		-50	
2			

CS221 / Autumn 2017 / Liang & Ermon

Mechanika świata wulkanów

		-50	20
		-50	
2			

- Możliwe 4 akcje (UDLR)

Mechanika świata wulkanów

		-50	20
		-50	
2			

- Możliwe 4 akcje (UDLR)
- W normalnym przypadku efekt oczywisty (próba wyjścia poza planszę oznacza pozostanie na polu)

Mechanika świata wulkanów

		-50	20
		-50	
2			

- Możliwe 4 akcje (UDLR)
- W normalnym przypadku efekt oczywisty (próba wyjścia poza planszę oznacza pozostanie na polu)
- Z prawdopodobieństwem p możemy się poślizgnąć, wówczas poruszamy się w losowym kierunku.

Mechanika świata wulkanów

		-50	20
		-50	
2			

- Możliwe 4 akcje (UDLR)
- W normalnym przypadku efekt oczywisty (próba wyjścia poza planszę oznacza pozostanie na polu)
- Z prawdopodobieństwem p możemy się poślizgnąć, wówczas poruszamy się w losowym kierunku.
- Dojście do pola z liczbą kończy grę (i odpowiednią dostajemy wypłatę).

Inny przykład. Gra w kości

Uwaga

Nagroda może być przydzielana w sposób ciągły, nie tylko w stanie końcowym.

Inny przykład. Gra w kości

Uwaga

Nagroda może być przydzielana w sposób ciągły, nie tylko w stanie końcowym.

- Mamy dwie opcje: pozostanie albo rezygnacja.

Inny przykład. Gra w kości

Uwaga

Nagroda może być przydzielana w sposób ciągły, nie tylko w stanie końcowym.

- Mamy dwie opcje: pozostanie albo rezygnacja.
- rezygnacja oznacza wypłatę **10$**

Inny przykład. Gra w kości

Uwaga

Nagroda może być przydzielana w sposób ciągły, nie tylko w stanie końcowym.

- Mamy dwie opcje: pozostanie albo rezygnacja.
- rezygnacja oznacza wypłatę **10$**
- pozostanie to wypłata **4$** po której rzucamy kostką.
- Interpretacja wyniku:

Inny przykład. Gra w kości

Uwaga

Nagroda może być przydzielana w sposób ciągły, nie tylko w stanie końcowym.

- Mamy dwie opcje: pozostanie albo rezygnacja.
- rezygnacja oznacza wypłatę **10$**
- pozostanie to wypłata **4$** po której rzucamy kostką.
- Interpretacja wyniku:
  - 1,2 – koniec gry
  - 3,4,5,6 – gramy dalej

Inny przykład. Gra w kości

Uwaga

Nagroda może być przydzielana w sposób ciągły, nie tylko w stanie końcowym.

- Mamy dwie opcje: pozostanie albo rezygnacja.
- rezygnacja oznacza wypłatę **10$**
- pozostanie to wypłata **4$** po której rzucamy kostką.
- Interpretacja wyniku:
  - 1,2 – koniec gry
  - 3,4,5,6 – gramy dalej

Pytanie

Ile mamy stanów?

Inny przykład. Gra w kości

Uwaga

Nagroda może być przydzielana w sposób ciągły, nie tylko w stanie końcowym.

- Mamy dwie opcje: pozostanie albo rezygnacja.
- rezygnacja oznacza wypłatę **10$**
- pozostanie to wypłata **4$** po której rzucamy kostką.
- Interpretacja wyniku:
  - 1,2 – koniec gry
  - 3,4,5,6 – gramy dalej

Pytanie

Ile mamy stanów? Odpowiedź: 2

Dla gry w kości

- 1 stan z decyzją, dwie polityki (czyli sensowne sposoby gry) (schemat na kolejnym slajdzie).

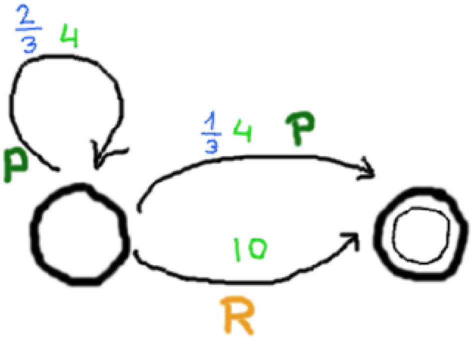
Dla gry w kości

- 1 stan z decyzją, dwie polityki (czyli sensowne sposoby gry) (schemat na kolejnym slajdzie).
- Możemy policzyć oczekiwaną wartość dla każdej:
  - rezygnacja – 10

Dla gry w kości

- 1 stan z decyzją, dwie polityki (czyli sensowne sposoby gry) (schemat na kolejnym slajdzie).
- Możemy policzyć oczekiwaną wartość dla każdej:
  - rezygnacja – 10
  - pozostanie – (na kolejnym slajdzie)

Schemat stanów



Wartość Pozostania

Oceniamy strategię pozostanie, czyli przedłużania gry.

Wartość Pozostania

Oceniamy strategię pozostanie, czyli przedłużania gry.

- Oznaczamy przez V wartość tej polityki (czyli ile średnio zarobimy, jak nie będziemy nigdy rezygnować z gry)

Wartość Pozostania

Oceniamy strategię pozostanie, czyli przedłużania gry.

- Oznaczamy przez V wartość tej polityki (czyli ile średnio zarobimy, jak nie będziemy nigdy rezygnować z gry)
- V spełnia równanie:

$$V = \frac{2}{3} \times (4 + V) + \frac{1}{3} \times 4$$

Wartość Pozostania

Oceniamy strategię pozostanie, czyli przedłużania gry.

- Oznaczamy przez V wartość tej polityki (czyli ile średnio zarobimy, jak nie będziemy nigdy rezygnować z gry)
- V spełnia równanie:

$$V = \frac{2}{3} \times (4 + V) + \frac{1}{3} \times 4 = 4 + \frac{2}{3} \times V$$

Wartość Pozostania

Oceniamy strategię pozostanie, czyli przedłużania gry.

- Oznaczamy przez V wartość tej polityki (czyli ile średnio zarobimy, jak nie będziemy nigdy rezygnować z gry)
- V spełnia równanie:

$$V = \frac{2}{3} \times (4 + V) + \frac{1}{3} \times 4 = 4 + \frac{2}{3} \times V$$

- Czyli $V = 12$, zatem opłaca się pozostawać w grze (bo $12 \geq 10$)

Wartość Pozostania

Oceniamy strategię pozostanie, czyli przedłużania gry.

- Oznaczamy przez V wartość tej polityki (czyli ile średnio zarobimy, jak nie będziemy nigdy rezygnować z gry)
- V spełnia równanie:

$$V = \frac{2}{3} \times (4 + V) + \frac{1}{3} \times 4 = 4 + \frac{2}{3} \times V$$

- Czyli $V = 12$, zatem opłaca się pozostawać w grze (bo $12 \geq 10$)

Uwaga

Tu była tylko jedna decyzja: 10 czy V, ale podobnie można rozwiązywać również (dużo) bardziej złożone MDP: rozwiązując równania i znajdując wartości stanów.

MDP – formalna definicja

Definicja

Markowowski proces decyzyjny (MDP) zawiera następujące składowe:

MDP – formalna definicja

Definicja

Markowowski proces decyzyjny (MDP) zawiera następujące składowe:

1. S – (skończony) zbiór stanów

MDP – formalna definicja

Definicja

Markowowski proces decyzyjny (MDP) zawiera następujące składowe:

1. S – (skończony) zbiór stanów
2. Stan startowy, $s_{\text{start}} \in S$

MDP – formalna definicja

Definicja

Markowowski proces decyzyjny (MDP) zawiera następujące składowe:

1. S – (skończony) zbiór stanów
2. Stan startowy, $s_{\text{start}} \in S$
3. Actions(s) – zbiór możliwych akcji w stanie s

Definicja

Markowowski proces decyzyjny (MDP) zawiera następujące składowe:

1. S – (skończony) zbiór stanów
2. Stan startowy, $s_{\text{start}} \in S$
3. Actions(s) – zbiór możliwych akcji w stanie s
4. T(s,a,s') – prawdopodobieństwo przejścia z s do s' w wyniku akcji a

MDP – formalna definicja

Definicja

Markowowski proces decyzyjny (MDP) zawiera następujące składowe:

1. S – (skończony) zbiór stanów
2. Stan startowy, $s_{\text{start}} \in S$
3. Actions(s) – zbiór możliwych akcji w stanie s
4. T(s,a,s') – prawdopodobieństwo przejścia z s do s' w wyniku akcji a
5. Reward(s,a,s') – nagroda (wypłata) związana z tym przejściem

MDP – formalna definicja

Definicja

Markowowski proces decyzyjny (MDP) zawiera następujące składowe:

1. S – (skończony) zbiór stanów
2. Stan startowy, $s_{\text{start}} \in S$
3. Actions(s) – zbiór możliwych akcji w stanie s
4. T(s,a,s') – prawdopodobieństwo przejścia z s do s' w wyniku akcji a
5. Reward(s,a,s') – nagroda (wypłata) związana z tym przejściem
6. IsEnd(s) – czy stan jest końcowy?

Definicja

Markowowski proces decyzyjny (MDP) zawiera następujące składowe:

1. S – (skończony) zbiór stanów
2. Stan startowy, $s_{\text{start}} \in S$
3. Actions(s) – zbiór możliwych akcji w stanie s
4. T(s,a,s') – prawdopodobieństwo przejścia z s do s' w wyniku akcji a
5. Reward(s,a,s') – nagroda (wypłata) związana z tym przejściem
6. IsEnd(s) – czy stan jest końcowy?
7. Discount factor, $0 < \gamma \leq 1$ – sprawia, że nagrody w przyszłości cieszą mniej.

MDP – komentarz do definicji

- Można też myśleć, że dla pary (s, a) mamy rozkład prawdopodobieństw po parach (nowy-stan, nagroda).

MDP – komentarz do definicji

- Można też myśleć, że dla pary (s, a) mamy rozkład prawdopodobieństw po parach (nowy-stan, nagroda).
- Nagroda może być pozytywna bądź negatywna

MDP – komentarz do definicji

- Można też myśleć, że dla pary (s, a) mamy rozkład prawdopodobieństw po parach (nowy-stan, nagroda).
- Nagroda może być pozytywna bądź negatywna

Uwaga

Oczywiście MDP jest ogólniejsze niż zadanie przeszukiwania (bo wystarczy przypisać niektórym rezultatom p-stwo 1, reszcie 0 i mamy zwykłe zadanie przeszukiwania)

Czym jest rozwiązanie MDP?

Czym jest rozwiązanie MDP?

- Przypominamy: rozwiązaniem zadania przeszukiwania jest ciąg akcji (ale to nie tu nie działa, bo?)

Czym jest rozwiązanie MDP?

- Przypominamy: rozwiązaniem zadania przeszukiwania jest ciąg akcji (ale to nie tu nie działa, bo?)
  - (wyniki akcji są niedeterministyczne, więc nie wystarczy podać jednego ciągu akcji)

Czym jest rozwiązanie MDP?

- Przypominamy: rozwiązaniem zadania przeszukiwania jest ciąg akcji (ale to nie tu nie działa, bo?)
  - (wyniki akcji są niedeterministyczne, więc nie wystarczy podać jednego ciągu akcji)
- Rozwiązanie: agent musi wiedzieć, co zrobić w każdym stanie.

Definicja 1

Polityką deterministyczną nazwiemy funkcję, która każdemu stanowi przypisuje akcję (możliwą w tym stanie).

Polityka

Definicja 1

Polityką deterministyczną nazwiemy funkcję, która każdemu stanowi przypisuje akcję (możliwą w tym stanie).

Definicja 2

Polityką nazwiemy funkcję, która każdemu stanowi przypisuje rozkład prawdopodobieństwa na akcjach (możliwych w tym stanie).

Polityka

Definicja 1

Polityką deterministyczną nazwiemy funkcję, która każdemu stanowi przypisuje akcję (możliwą w tym stanie).

Definicja 2

Polityką nazwiemy funkcję, która każdemu stanowi przypisuje rozkład prawdopodobieństwa na akcjach (możliwych w tym stanie).

Wartościowanie polityki

- Gdy używamy polityki, otrzymujemy ciąg stanów, akcji i nagród

Wartościowanie polityki

- Gdy używamy polityki, otrzymujemy ciąg stanów, akcji i nagród
- Dla takiej ścieżki możemy zsumować nagrody, otrzymując użyteczność dla tej ścieżki

Wartościowanie polityki

- Gdy używamy polityki, otrzymujemy ciąg stanów, akcji i nagród
- Dla takiej ścieżki możemy zsumować nagrody, otrzymując użyteczność dla tej ścieżki
- Wartością polityki jest oczekiwana użyteczność polityki (tzn. wartość oczekiwana zmiennej losowej wyrażającej użyteczność takiej ścieżki)

Discounting

- Realizując politykę, otrzymaliśmy ciąg stanów, nagród i akcji
  - $s_0, a_1, r_1, s_1, a_2, r_2, s_2 \ldots, s_n, a_{n+1}, r_{n+1}, s_{n+1} \ldots$

Discounting

- Realizując politykę, otrzymaliśmy ciąg stanów, nagród i akcji
  - $s_0, a_1, r_1, s_1, a_2, r_2, s_2 \ldots, s_n, a_{n+1}, r_{n+1}, s_{n+1} \ldots$
- Nagroda po uwzględnieniu zniżek:

$$r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \ldots$$

Discounting

- Realizując politykę, otrzymaliśmy ciąg stanów, nagród i akcji
  - $s_0, a_1, r_1, s_1, a_2, r_2, s_2 \ldots, s_n, a_{n+1}, r_{n+1}, s_{n+1} \ldots$
- Nagroda po uwzględnieniu zniżek:

$$r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \ldots$$

- Widzimy że:
  - Dla $\gamma = 1$ po prostu sumujemy nagrody cząstkowe

Discounting

- Realizując politykę, otrzymaliśmy ciąg stanów, nagród i akcji
  - $s_0, a_1, r_1, s_1, a_2, r_2, s_2 \ldots, s_n, a_{n+1}, r_{n+1}, s_{n+1} \ldots$
- Nagroda po uwzględnieniu zniżek:

$$r_1 + \gamma r_2 + \gamma^2 r_3 + \gamma^3 r_4 + \ldots$$

- Widzimy że:
  - Dla $\gamma = 1$ po prostu sumujemy nagrody cząstkowe
  - Dla $0 < \gamma < 1$ mamy możliwość mówienia o wartości nieskończonych ciągów akcji.

Uwaga o γ

- Zwróćmy uwagę, że discounting ma sens również w przypadku, gdy nagroda wypłacana jest jedynie w stanie końcowym.

Uwaga o γ

- Zwróćmy uwagę, że discounting ma sens również w przypadku, gdy nagroda wypłacana jest jedynie w stanie końcowym.
- Jeżeli wypłata jest tylko w ostatnim stanie, to:

Uwaga o γ

- Zwróćmy uwagę, że discounting ma sens również w przypadku, gdy nagroda wypłacana jest jedynie w stanie końcowym.
- Jeżeli wypłata jest tylko w ostatnim stanie, to:
  a) Agent, który wygrywa ($R > 0$) woli dostać ją wcześniej,
  b) agent, który przegrywa ($R < 0$) woli dostać ją później.

Uwaga o γ

- Zwróćmy uwagę, że discounting ma sens również w przypadku, gdy nagroda wypłacana jest jedynie w stanie końcowym.
- Jeżeli wypłata jest tylko w ostatnim stanie, to:
  a) Agent, który wygrywa ($R > 0$) woli dostać ją wcześniej,
  b) agent, który przegrywa ($R < 0$) woli dostać ją później.

Przyśpieszanie zwycięstwa i opóźnianie porażki jest „sensownym" zachowaniem.

Wartość polityki

Definicja

Wartość $V_\pi(s)$ jest oczekiwaną użytecznością dla agenta startującego w stanie s i działającego zgodnie z polityką π

Wartość polityki

Definicja

Wartość $V_\pi(s)$ jest oczekiwaną użytecznością dla agenta startującego w stanie s i działającego zgodnie z polityką π

Definicja

Wartość $Q_\pi(s, a)$ jest oczekiwaną użytecznością dla agenta startującego w stanie s, wykonującego w tym stanie akcję a i **dalej** działającego zgodnie z polityką π

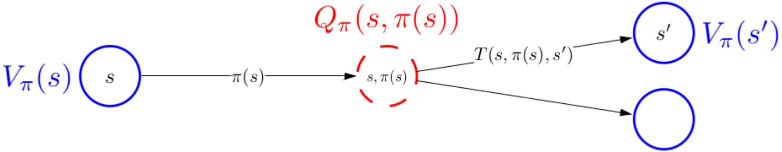
Wartość polityki

Definicja

Wartość $V_\pi(s)$ jest oczekiwaną użytecznością dla agenta startującego w stanie s i działającego zgodnie z polityką π

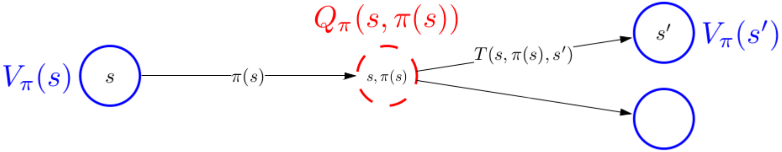
Definicja

Wartość $Q_\pi(s, a)$ jest oczekiwaną użytecznością dla agenta startującego w stanie s, wykonującego w tym stanie akcję a i **dalej** działającego zgodnie z polityką π



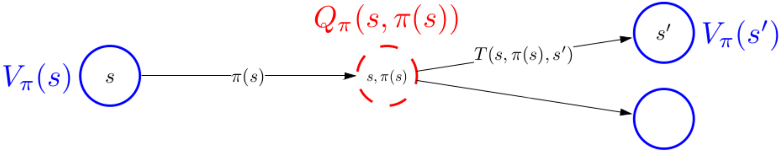
Źródło: CS221 / Autumn 2017 / Liang & Ermon

Zależności pomiędzy V i Q



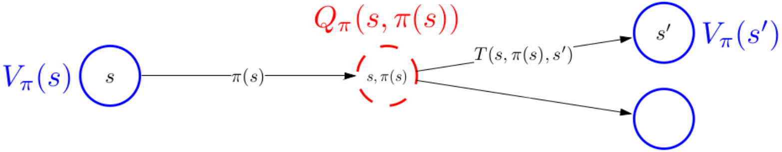
$Q_\pi(s, a) =$

Zależności pomiędzy V i Q



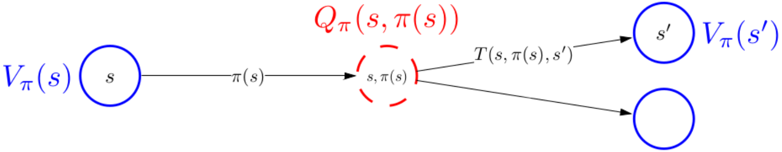
$$Q_\pi(s,a) = \sum_{s'}$$

Zależności pomiędzy V i Q



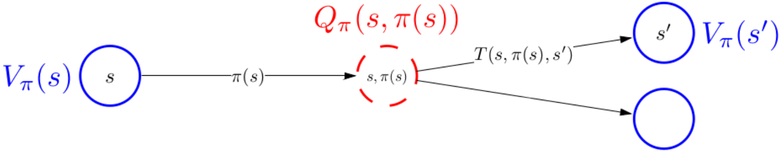
$$Q_\pi(s, a) = \sum_{s'} T(s, a, s')$$

Zależności pomiędzy V i Q



$$Q_\pi(s,a) = \sum_{s'} T(s,a,s')(\text{Reward}(s,a,s') +$$

Zależności pomiędzy V i Q



$$Q_\pi(s,a) = \sum_{s'} T(s,a,s')(\mathrm{Reward}(s,a,s') + \gamma V_\pi(s'))$$

Algorytm: policy evaluation

- Napiszmy rekurencyjny wzór dla wartości V (przy zadanej polityce)

Algorytm: policy evaluation

- Napiszmy rekurencyjny wzór dla wartości V (przy zadanej polityce)
-
$$V_\pi(s) = \sum_{s'}$$

Algorytm: policy evaluation

- Napiszmy rekurencyjny wzór dla wartości V (przy zadanej polityce)
-

$$V_\pi(s) = \sum_{s'} T(s, \pi(s), s')(\text{Reward}(s, \pi(s), s') + \gamma V_\pi(s'))$$

Algorytm: policy evaluation

- Napiszmy rekurencyjny wzór dla wartości V (przy zadanej polityce)
- $$V_\pi(s) = \sum_{s'} T(s, \pi(s), s')(\text{Reward}(s, \pi(s), s') + \gamma V_\pi(s'))$$
- Mamy układ równań (liniowych), który można rozwiązywać standardowymi metodami.

Algorytm: policy evaluation

- Napiszmy rekurencyjny wzór dla wartości V (przy zadanej polityce)
-
$$V_\pi(s) = \sum_{s'} T(s, \pi(s), s')(\text{Reward}(s, \pi(s), s') + \gamma V_\pi(s'))$$
- Mamy układ równań (liniowych), który można rozwiązywać standardowymi metodami.
- Równań jest tyle co stanów (czyli potencjalnie sporo)

Algorytm: policy evaluation (2)

Możemy ten wzór zmodyfikować, mówiąc: zamiast nieznanego V po prawej stronie weźmiemy poprzednie przybliżenie V:

Algorytm: policy evaluation (2)

Możemy ten wzór zmodyfikować, mówiąc: zamiast nieznanego V po prawej stronie weźmiemy poprzednie przybliżenie V:

$$V_{\pi}^{(t+1)}(s) = \sum_{s'}$$

Algorytm: policy evaluation (2)

Możemy ten wzór zmodyfikować, mówiąc: zamiast nieznanego V po prawej stronie weźmiemy poprzednie przybliżenie V:

$$V_\pi^{(t+1)}(s) = \sum_{s'} T(s, \pi(s), s')(\text{Reward}(s, \pi(s), s') + \gamma V_\pi^{(t)}(s'))$$

Algorytm: policy evaluation (3)

1. Zainicjuj $V_\pi^{(0)}(s) \leftarrow 0$, dla wszystkich s

Algorytm: policy evaluation (3)

1. Zainicjuj $V_\pi^{(0)}(s) \leftarrow 0$, dla wszystkich s
2. Powtarzaj dla $t = 1, \ldots, t_{PE}$

Algorytm: policy evaluation (3)

1. Zainicjuj $V_\pi^{(0)}(s) \leftarrow 0$, dla wszystkich s
2. Powtarzaj dla $t = 1, \ldots, t_{PE}$
   - Powtarzaj dla każdego stanu s

$$V_\pi^{(t+1)}(s) \leftarrow \sum_{s'}$$

Algorytm: policy evaluation (3)

1. Zainicjuj $V_\pi^{(0)}(s) \leftarrow 0$, dla wszystkich s
2. Powtarzaj dla $t = 1, \ldots, t_{PE}$
   - Powtarzaj dla każdego stanu s

$$V_\pi^{(t+1)}(s) \leftarrow \sum_{s'} T(s, \pi(s), s')(\mathrm{Reward}(s, \pi(s), s') + \gamma V_\pi^{(t)}(s'))$$

Uwagi implementacyjne

- Kończymy, gdy dla każdego stanu zmiana mniejsza niż ε

Uwagi implementacyjne

- Kończymy, gdy dla każdego stanu zmiana mniejsza niż ε
- Oczywiście nie musimy pamiętać całej historii, tylko dwa ostatnie jej elementy (stany zmieniane i poprzednie)

Uwagi implementacyjne

- Kończymy, gdy dla każdego stanu zmiana mniejsza niż ε
- Oczywiście nie musimy pamiętać całej historii, tylko dwa ostatnie jej elementy (stany zmieniane i poprzednie)

Złożoność

Uwagi implementacyjne

- Kończymy, gdy dla każdego stanu zmiana mniejsza niż ε
- Oczywiście nie musimy pamiętać całej historii, tylko dwa ostatnie jej elementy (stany zmieniane i poprzednie)

Złożoność

$O(t_{PE}SS')$, gdzie S to liczba stanów, a S' (maksymalna) liczba stanów z niezerową $T(s, a, s')$.

Polityka optymalna

- Interesuje nas wyznaczanie polityki (a nie tylko ocenianie jej wartości).

Polityka optymalna

- Interesuje nas wyznaczanie polityki (a nie tylko ocenianie jej wartości).

Definicja

Optymalną wartością stanu $V_{opt}(s)$ jest maksymalna wartość stanu (ze względu na wszystkie polityki).

Rekurencja dla polityki optymalnej

Jaka polityka jest optymalna?

Rekurencja dla polityki optymalnej

Jaka polityka jest optymalna? Taka, która wybiera stany o optymalnej wartości

Jaka polityka jest optymalna? Taka, która wybiera stany o optymalnej wartości

- Przypominamy, dla **każdej** polityki mamy:

$$Q_\pi(s,a) = \sum_{s'} T(s,a,s')(\text{Reward}(s,a,s') + \gamma V_\pi(s'))$$

Rekurencja dla polityki optymalnej

Jaka polityka jest optymalna? Taka, która wybiera stany o optymalnej wartości

- Przypominamy, dla **każdej** polityki mamy:

$$Q_\pi(s,a) = \sum_{s'} T(s,a,s')(\mathrm{Reward}(s,a,s') + \gamma V_\pi(s'))$$

- Dla polityki optymalnej: $V_{\mathrm{opt}}(s) = \max_{a \in \mathrm{Actions(s)}} Q_{\mathrm{opt}}(s,a)$

Rekurencja dla polityki optymalnej

Jaka polityka jest optymalna? Taka, która wybiera stany o optymalnej wartości

- Przypominamy, dla **każdej** polityki mamy:

$$Q_\pi(s,a) = \sum_{s'} T(s,a,s')(\text{Reward}(s,a,s') + \gamma V_\pi(s'))$$

- Dla polityki optymalnej: $V_{\text{opt}}(s) = \max_{a \in \text{Actions(s)}} Q_{\text{opt}}(s,a)$

Możemy podstawić do drugiego wzoru wzór na Q_π dla $\pi = \text{opt}$.

Rekurencja dla polityki optymalnej

Jaka polityka jest optymalna? Taka, która wybiera stany o optymalnej wartości

- Przypominamy, dla **każdej** polityki mamy:

$$Q_\pi(s,a) = \sum_{s'} T(s,a,s')(\mathrm{Reward}(s,a,s') + \gamma V_\pi(s'))$$

- Dla polityki optymalnej: $V_{\mathrm{opt}}(s) = \max_{a \in \mathrm{Actions(s)}} Q_{\mathrm{opt}}(s,a)$

Możemy podstawić do drugiego wzoru wzór na Q_π dla $\pi = \mathrm{opt}$.

$$V_{\mathrm{opt}}(s) = \max_{a \in \mathrm{Actions(s)}} \sum_{s'} T(s,a,s')(\mathrm{Reward}(s,a,s') + \gamma V_{\mathrm{opt}}(s'))$$

Algorytm Iteracji wartości – Bellman, 1957

Polityka optymalna (do poprzedniego slajdu)

$\pi_{\text{opt}}(s) = \arg\max_{a \in \text{Actions(s)}} Q_{\text{opt}}(s, a)$

Algorytm Iteracji wartości – Bellman, 1957

Polityka optymalna (do poprzedniego slajdu)

$\pi_{\text{opt}}(s) = \arg\max_{a \in \text{Actions(s)}} Q_{\text{opt}}(s, a)$

Nasz wzorek zmieniony na wersję do iterowania

$$V_{\text{opt}}^{(t+1)}(s) = \max_{a \in \text{Actions(s)}} \sum_{s'} T(s, a, s')(\text{Reward}(s, a, s') + \gamma V_{\text{opt}}^{(t)}(s'))$$

Algorytm Iteracji wartości – Bellman, 1957

Polityka optymalna (do poprzedniego slajdu)

$\pi_{\text{opt}}(s) = \arg\max_{a \in \text{Actions(s)}} Q_{\text{opt}}(s, a)$

Nasz wzorek zmieniony na wersję do iterowania

$$V_{\text{opt}}^{(t+1)}(s) = \max_{a \in \text{Actions(s)}} \sum_{s'} T(s, a, s')(\text{Reward}(s, a, s') + \gamma V_{\text{opt}}^{(t)}(s'))$$

Algorytm Bellmana (value iteration)

- Mamy dodatkową pętlę wybierającą optymalną akcję (zamiast akcji danej przez politykę)
- Reszta bez zmian, tak jak w policy evaluation.

Warunki zbieżności

Algorytm jest zbieżny, jeżeli zachodzi któryś z warunków

- $\gamma < 1$
- Graf MDP jest acykliczny

Warunki zbieżności

Algorytm jest zbieżny, jeżeli zachodzi któryś z warunków

- $\gamma < 1$
- Graf MDP jest acykliczny

Uwaga

W tym ostatnim przypadku wymagana jest jedna iteracja, w której stany przeglądane są w odwrotnym porządku topologicznym (wyjaśnienie na ćwiczeniach)

Warunki zbieżności

Algorytm jest zbieżny, jeżeli zachodzi któryś z warunków

- $\gamma < 1$
- Graf MDP jest acykliczny

Uwaga

W tym ostatnim przypadku wymagana jest jedna iteracja, w której stany przeglądane są w odwrotnym porządku topologicznym (wyjaśnienie na ćwiczeniach)

Uwaga

Zwróćmy uwagę na to ci się dzieje, jeżeli $\gamma = 1$ i mamy cykl.

Warunki zbieżności

Algorytm jest zbieżny, jeżeli zachodzi któryś z warunków

- $\gamma < 1$
- Graf MDP jest acykliczny

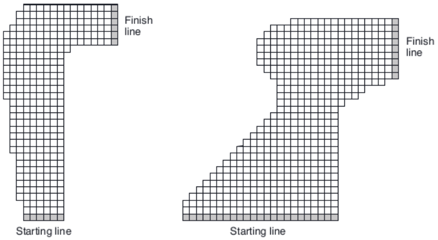
Uwaga

W tym ostatnim przypadku wymagana jest jedna iteracja, w której stany przeglądane są w odwrotnym porządku topologicznym (wyjaśnienie na ćwiczeniach)

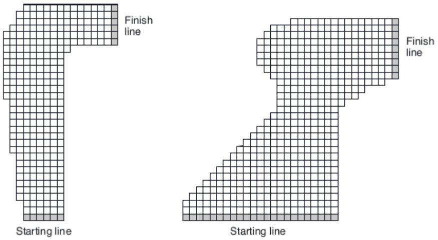
Uwaga

Zwróćmy uwagę na to ci się dzieje, jeżeli $\gamma = 1$ i mamy cykl.
Dla niezerowych nagród na krawędziach cyklu wartość oczekiwana może być nieokreślona

Przykład. Wyścigi samochodzików.

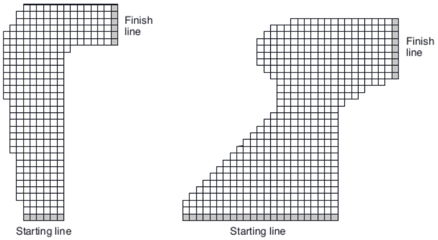


Przykład. Wyścigi samochodzików.



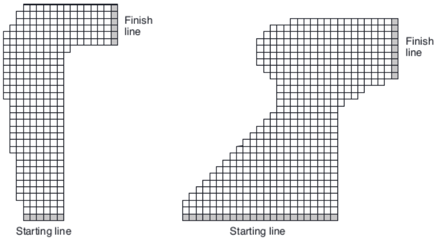
- Prędkość autka jest wektorem $(dx, dy) \in \{-3, -2, \ldots, 2, 3\} \times \{-3, -2, \ldots, 2, 3\}$

Przykład. Wyścigi samochodzików.



- Prędkość autka jest wektorem $(dx, dy) \in \{-3, -2, \ldots, 2, 3\} \times \{-3, -2, \ldots, 2, 3\}$
- Akcja: zmiana prędkości (każda składowa o co najwyżej 1)
- Celem jest (przejechać) przez metę (możemy to uprościć poszerzając metę i mówiąc, że celem jest dotarcie do piksela mety)

Przykład. Wyścigi samochodzików.

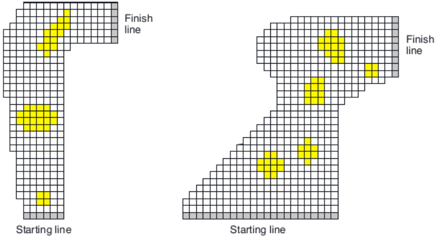


- Prędkość autka jest wektorem $(dx, dy) \in \{-3, -2, \ldots, 2, 3\} \times \{-3, -2, \ldots, 2, 3\}$
- Akcja: zmiana prędkości (każda składowa o co najwyżej 1)
- Celem jest (przejechać) przez metę (możemy to uprościć poszerzając metę i mówiąc, że celem jest dotarcie do piksela mety)
- W pełni deterministyczny świat (BFS, A^*?)

Wyścigi samochodzików. (2)

- Dodajemy plamy po oleju

Wyścigi samochodzików. (2)



- Dodajemy plamy po oleju
- Ruch z pola oleju dodaje dodatkową składową losową do prędkości (znamy rozkład).

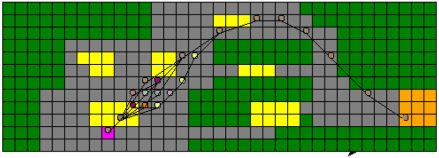
Wyścigi samochodzików. (2)

- Dodajemy plamy po oleju
- Ruch z pola oleju dodaje dodatkową składową losową do prędkości (znamy rozkład).

W tym momencie klasyczne MDP + algorytm Bellmana (czyli iteracji wartości) powinny dać dobry wynik.

Wynik algorytmu Value Iteration

Wynik algorytmu Value Iteration



Zwróćmy uwagę, że bez żadnych dodatkowych obliczeń można umieszczać w innych miejscach punkt startowy.

Jeszcze o autach i oleju

- Fajnie jest dojechać na metę. (+100)
- Ale jeszcze fajniej nie dać się zabić. (-100?)

Jeszcze o autach i oleju

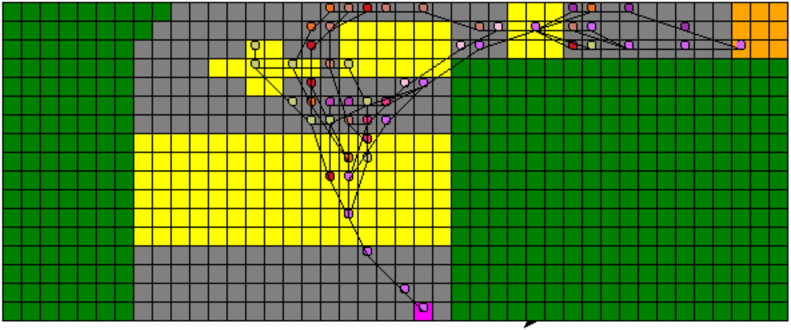
- Fajnie jest dojechać na metę. (+100)
- Ale jeszcze fajniej nie dać się zabić. (-100?)

Uwaga

Pamiętamy, że monotoniczna zmiana funkcji wypłaty:

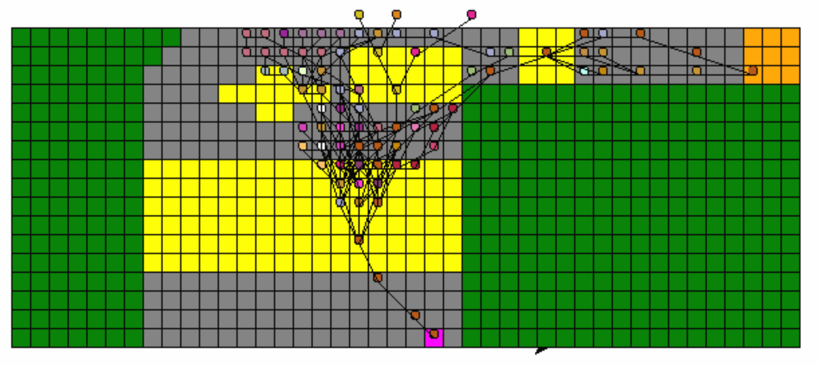
1. nie zmienia wartości MiniMax-owej gry,
2. może zmienić ExpectMinMax

Rozwiązanie podstawowe

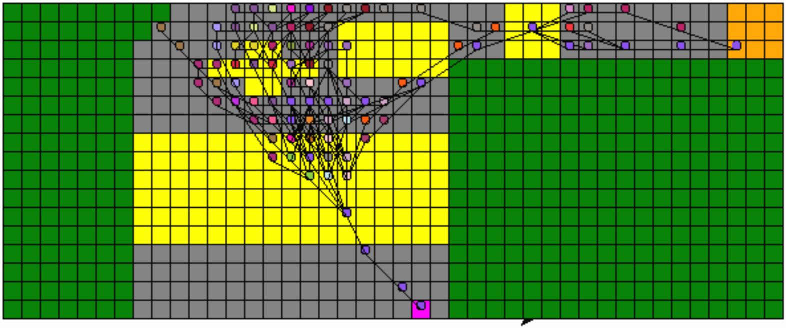


Pytanie: Czego spodziewamy się, jeżeli zamienimy karę na wypadek na 10000?

Kara=100



Kara=10000



Wyścigi samochodzików. (3)

- Problem: dużo większa plansza, dużo większa liczba stanów.

Wyścigi samochodzików. (3)

- Problem: dużo większa plansza, dużo większa liczba stanów.
- **Pomysł 1**: położenie „rozmyte”, na przykład w kwadracie 10×10 pikseli.

Wyścigi samochodzików. (3)

- Problem: dużo większa plansza, dużo większa liczba stanów.
- **Pomysł 1**: położenie „rozmyte”, na przykład w kwadracie 10×10 pikseli.
- **Pomysł 2**: dodatkowo informacja, czy jestem 1, 2, czy 3 raz w takim kwadracie $(3 < 100)$

Wyścigi samochodzików. (3)

- Problem: dużo większa plansza, dużo większa liczba stanów.
- **Pomysł 1**: położenie „rozmyte”, na przykład w kwadracie 10×10 pikseli.
- **Pomysł 2**: dodatkowo informacja, czy jestem 1, 2, czy 3 raz w takim kwadracie ($3 < 100$)

Fundamentalny problem: nie znamy mechaniki takiego świata (i wielu innych)

Wyścigi samochodzików. Float

- Prędkość autka jest wektorem $(v \cos(d), v \sin(d))$,

Wyścigi samochodzików. Float

- Prędkość autka jest wektorem $(v \cos(d), v \sin(d))$,
- Możemy zmieniać d (skręcać), oraz v (przyśpieszać, hamować)
- Celem jest meta.

Wyścigi samochodzików. Float

- Prędkość autka jest wektorem $(v \cos(d), v \sin(d))$,
- Możemy zmieniać d (skręcać), oraz v (przyśpieszać, hamować)
- Celem jest meta.
- W pełni deterministyczny świat, ale

Wyścigi samochodzików. Float

- Prędkość autka jest wektorem $(v\cos(d), v\sin(d))$,
- Możemy zmieniać d (skręcać), oraz v (przyśpieszać, hamować)
- Celem jest meta.
- W pełni deterministyczny świat, ale bardzo duża liczba stanów, zawierających liczby `float`)

Autka float. Rozwiązanie

- Możemu stworzyć **stan abstrakcyjny** i opisać mechanikę świata dla takich stanów

Autka float. Rozwiązanie

- Możemu stworzyć **stan abstrakcyjny** i opisać mechanikę świata dla takich stanów
- Oczywiście będzie ona niedeterministyczna, bo nigdy nie będziemy wiedzieć, czy zmiana w świecie float-ów przenosi się na zmianę w świecie int-ów.

Autka float. Rozwiązanie

- Możemu stworzyć **stan abstrakcyjny** i opisać mechanikę świata dla takich stanów
- Oczywiście będzie ona niedeterministyczna, bo nigdy nie będziemy wiedzieć, czy zmiana w świecie float-ów przenosi się na zmianę w świecie int-ów.

Uwaga

Możemy myśleć o tym, że modelujemy błędy pomiarowe (int zamiast float) za pomocą losowości.