Zastosowanie metody UCT i drzewa strategii behawioralnej do aproksymacji stanu równowagowego Stackelberga w grach wielokrokowych

Jan Karwowski

Zakład Sztucznej Inteligencji i Metod Obliczeniowych
Wydział Matematyki i Nauk Informacyjnych PW

17 I 2018

Plan

Cele

Equilibrium Stackelberga: Istniejące rozwiązania

- Oparte o LP i MILP
- Uniwersalne ale wolne i pamięciożerne lub
- wykorzystujące własności konkretnych gier do przyspieszenia i aproksymacji wyniku
- Macierze rosną wykładniczo względem liczby rund

Nowe podejście

- Metoda dobrze działająca z dyskretnymi grami wielokrokowymi
- Metoda dobrze skalująca się z długością gry (szczególnie pamięć)
- Metoda łatwa do dostosowania do różnych rodzajów gier
- Rozwiązania przybliżone

Postać normalna gry o sumie (nie)zerowej

Macierz wypłat gracza P1

P1/P2	a	b	c	d
A	6	4	-2	1
B	0	0	2	3
C	-3	-2	-1	-1

- $U_{P1}(A, b) = 4$
- $U_{P2}(A, b) = -4$

Macierz wypłat gracza P2

P1/P2	a	b	c	d
A	-6	3	2	-1
B	0	0	-2	-3
C	-3	2	0	1

- $U_{P2}(A, b) = 3$

Postać normalna gry o sumie (nie)zerowej

Macierz wypłat gracza P1

P1/P2	a	b	c	d
A	6	4	-2	1
B	0	0	2	3
C	-3	-2	-1	-1

- $U_{P1}(A, b) = 4$
- $U_{P2}(A, b) = -4$

Macierz wypłat gracza P2

P1/P2	a	b	c	d
A	-6	3	2	-1
B	0	0	-2	-3
C	-3	2	0	1

-
- $U_{P2}(A, b) = 3$

Strategia mieszana

E	H	T
H	1	-1
T	-1	1

O	H	T
H	-1	1
T	1	-1

- Strategia podstawowa (wybór jednego ruchu) – π
- Strategia mieszana – rozkład prawdopodobieństwa nad π – σ

Równowaga Stackelberga

- Asymetryczni gracze: *Leader*, *Follower*
- *Follower* zna **strategię** *Leadera* w momencie wyboru swojej strategii. (Ale niekoniecznie wykonuje **ruch** po leaderze).
- *Follower* rozstrzyga remisy (swojej wypłaty) na korzyść *Leadera*

Gra

	F1	F2
L1	-15	3
L2	3	-15

	F1	F2
L1	30	0
L2	0	20

Equilibrium

Leader			Follower			
Pr.		U_l	Pr.		U_f	U_l
0.4	L1	-15	1	F1	12	-4.2
0.6	L2	3	0	F2	12	-7.8

Zaburzone equilirbium

Leader			Follower			
Pr.		U_l	Pr.		U_f	U_l
0.39	L1	3	0	F1	11.7	-4
0.61	L2	-15	1	F2	12.2	-8

Równowaga Stackelberga

- Asymetryczni gracze: *Leader*, *Follower*
- *Follower* zna **strategię** *Leadera* w momencie wyboru swojej strategii. (Ale niekoniecznie wykonuje **ruch** po leaderze).
- *Follower* rozstrzyga remisy (swojej wypłaty) na korzyść *Leadera*

Gra

	F1	F2
L1	-15	3
L2	3	-15

	F1	F2
L1	30	0
L2	0	20

Dwupoziomowy problem optymalizacyjny

$$\arg\max_{\sigma_l \in \Sigma_l} U_l(\sigma_l, R_f(\sigma_l))$$

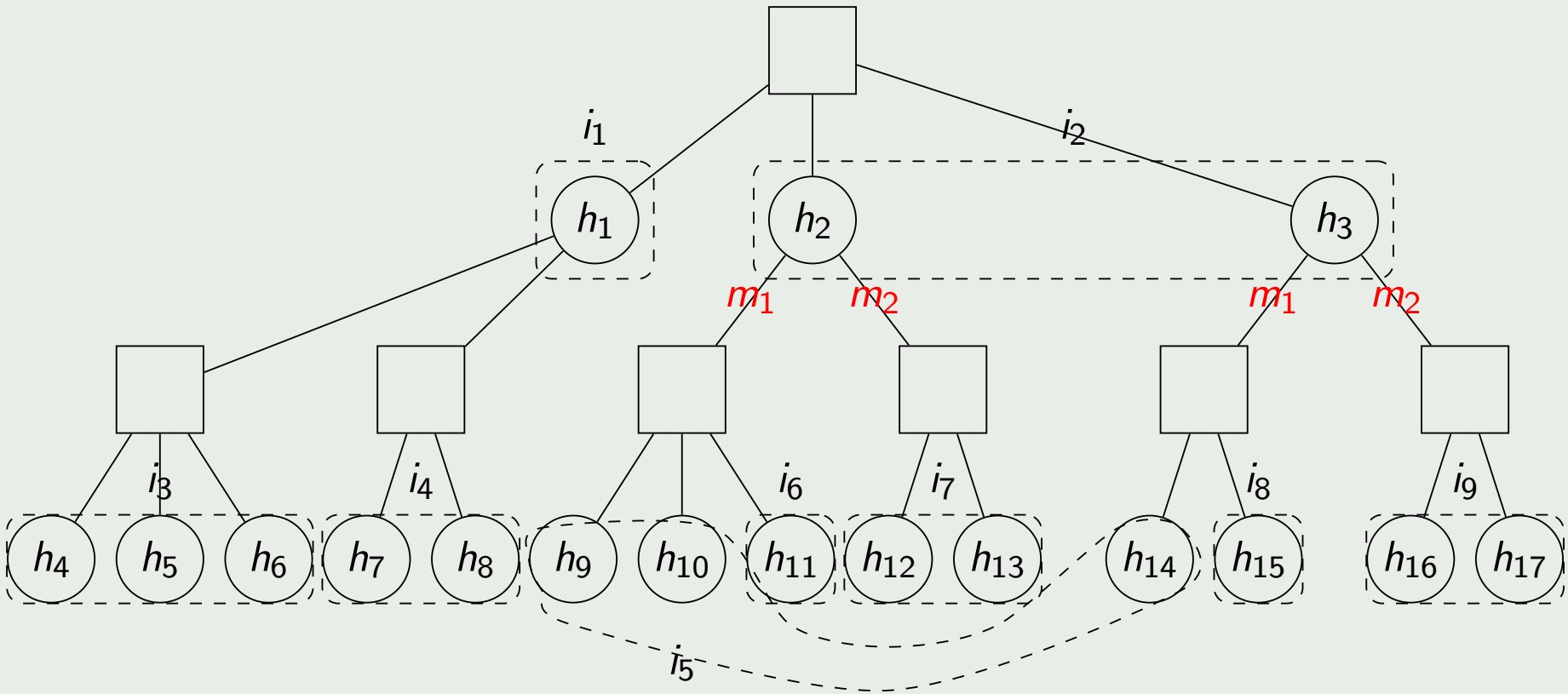
$R_f(\sigma_l) = \arg\max_{\sigma_f \in \Sigma_f} U_f(\sigma_l, \sigma_f)$ – funkcja schodkowa

Strategie followera

Można pokazać, że dla dowolnego equilirbium (σ_l, σ_f) układ (σ_l, π_f) dla dowolnego π_f mającego niezerowe prawdopodobieństwo w σ_f też jest equilibrium.
Wniosek: wystarczy rozważać tylko strategie podstawowe followera (ale mieszane leadera!!).

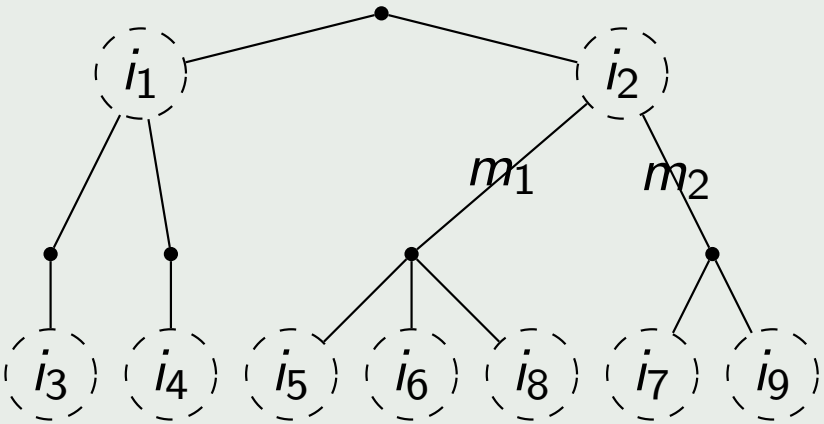
Gry wielokrokowe I

Gra

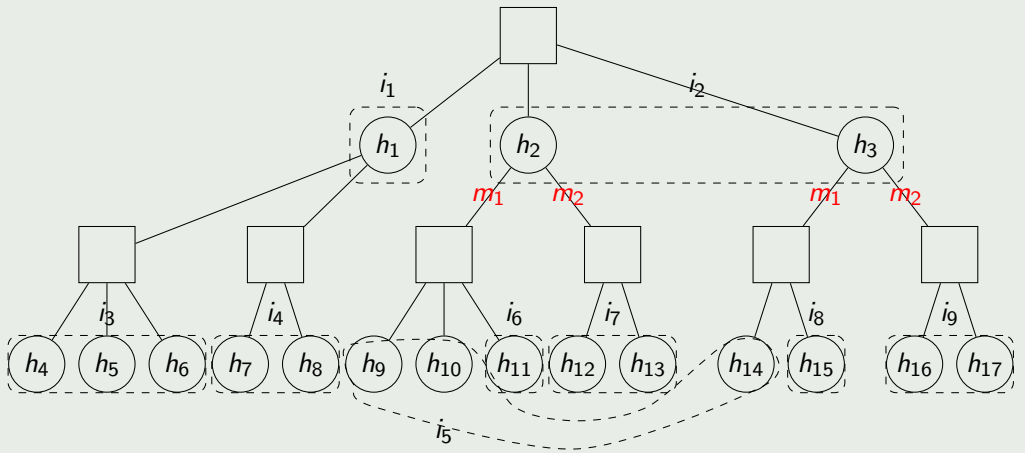


Gry wielokrokowe II

Gracz ○



Gra



- Strategia podstawowa — przypisanie akcji do każdego punktu decyzyjnego (information setu). Strategia zredukowana — tylko potencjalnie osiągalne IS.
- Strategia mieszana — rozkład prawdopodobieństwa nad strategiami podstawowymi
- Strategia behawioralna — przypisanie rozkładu prawdopodobieństwa akcji w każdym punkcie decyzyjnym

Security Games

- Obrońcy i atakujący
- Asymetria graczy
- Różne rodzaje przestrzeni
- Zwykle obrońca ma większą przestrzeń decyzyjną
- Brak jednolitej definicji
- Zazwyczaj równowaga Stackelberga

Typowe podejście

- Przegląd wszystkich strategii followera i rozwiązanie LP poszukującego dobrego leadera
- MILP, który łączy przegląd followera i poszukiwanie leadera

MILP

$$
\begin{array}{rl}
 & \max_{q,z,a} \sum_{i \in X} \sum_{j \in Q} R_{ij} z_{ij} \\
\text{s.t.} & \sum_{i \in X} \sum_{j \in Q} z_{ij} = 1 \\
(\forall_{i \in X}) & \sum_{j \in Q} z_{i,j} \leqslant 1 \\
(\forall_{j \in Q}) & q_j \leqslant \sum_{i \in X} z_{ij} \leqslant 1 \\
 & \sum_{j \in Q} q_j = 1 \\
(\forall_{j \in Q}) & 0 \leqslant (a - \sum_{i \in X} C_{ij} (\sum_{h \in Q} z_{ih})) \\
(\forall_{j \in Q}) & (a - \sum_{i \in X} C_{ij} (\sum_{h \in Q} z_{ih})) \leqslant (1 - q_j) M
\end{array}
$$

$$
z_{ij} \in [0,1], q_j \in \{0,1\}, a \in \mathbb{R}
$$

Praveen Paruchuri et al. "Efficient Algorithms to Solve Bayesian Stackelberg Games for Security Applications.". In: *AAAI*. 2008, pp. 1559–1562

Pomysł

Heurystyka wyboru interesujących strategii folllowera + przybliżone znalezienie strategii obrońcy.

Szukanie leadera

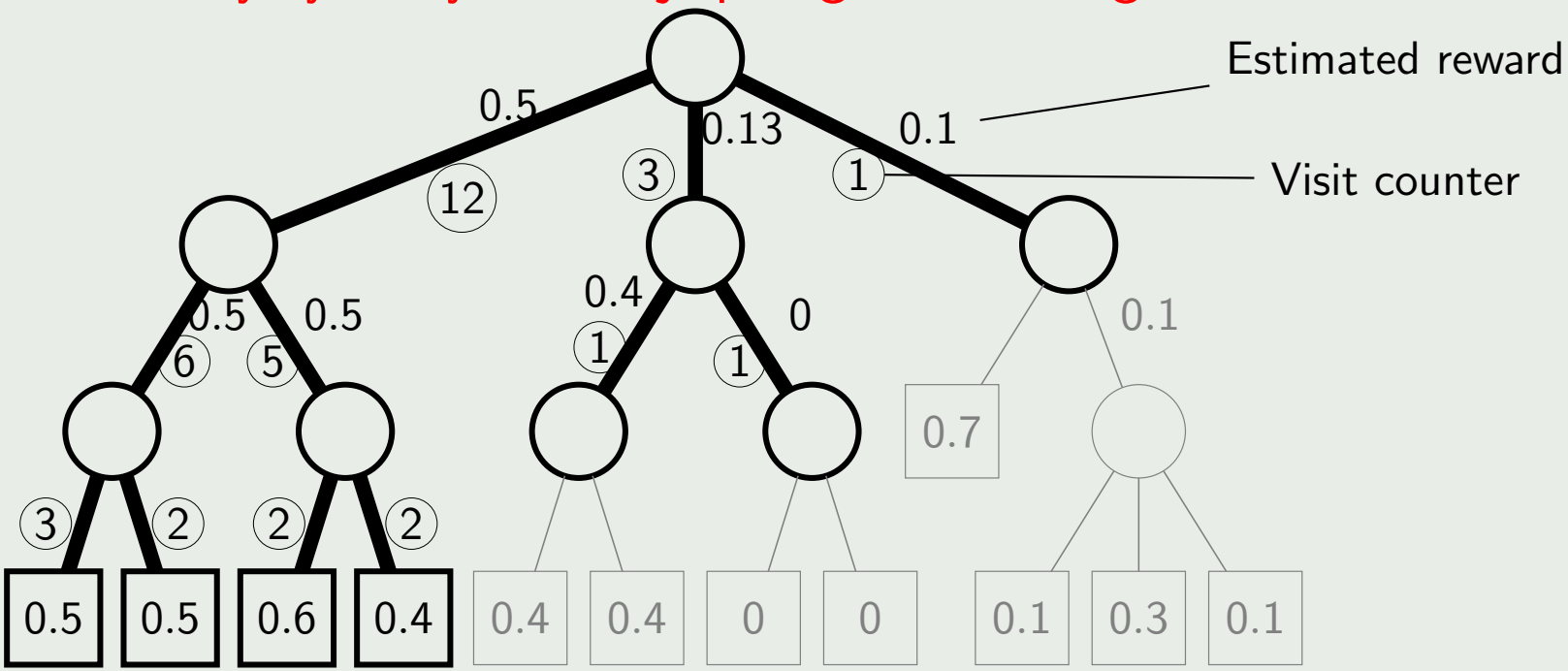
Strategia obrońcy musi spełniać ograniczenie wymuszające założoną odpowiedź followera.

Przegląd strategii followera z użyciem UCT I

UCT

Levente Kocsis and Csaba Szepesvári. "Bandit based monte-carlo planning". In: *Machine Learning: ECML 2006*. Springer, 2006, pp. 282–293

Metaheurystyka wyboru najlepszego ruchu w grze.



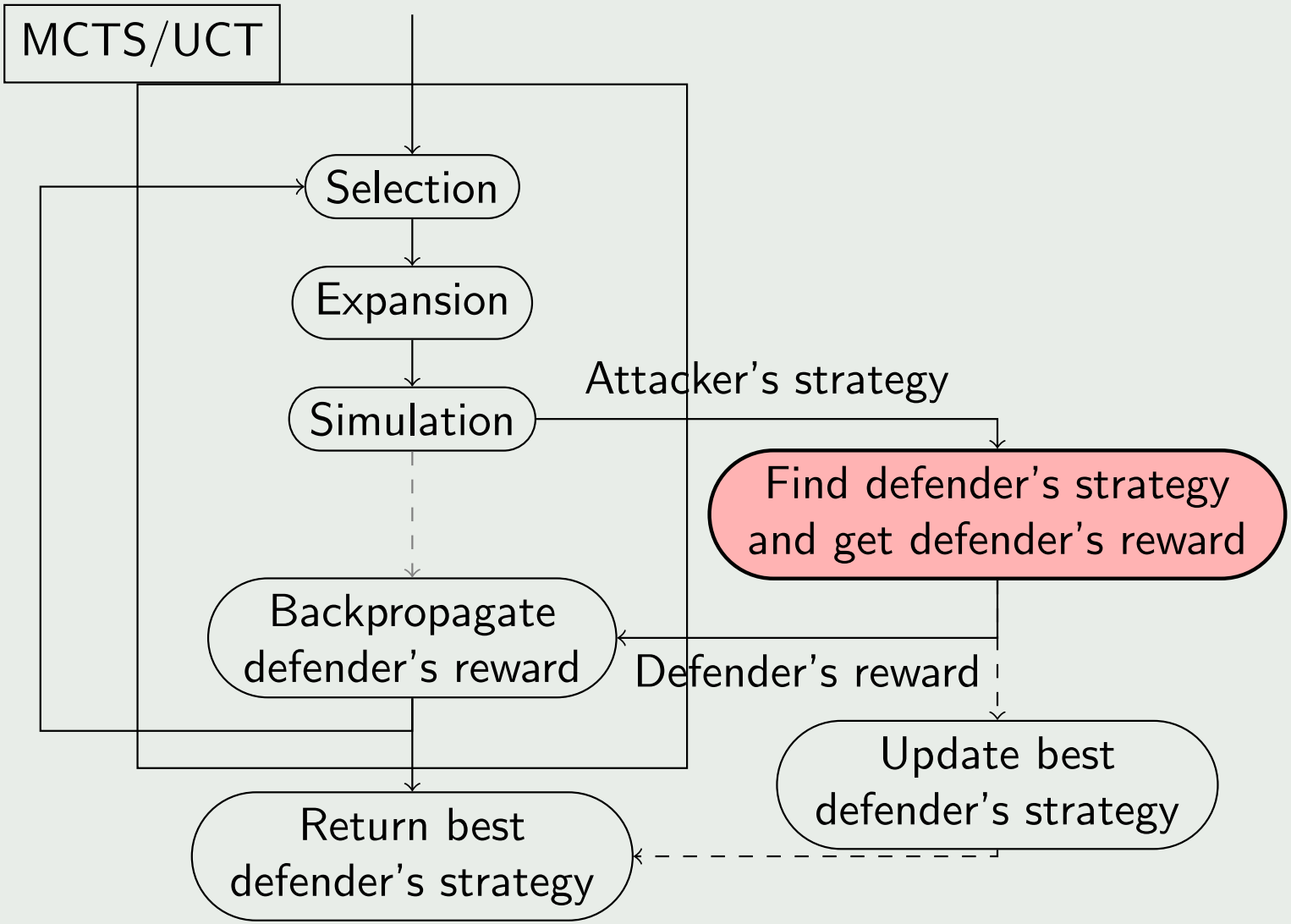
Przegląd strategii followera z użyciem UCT II

Gra

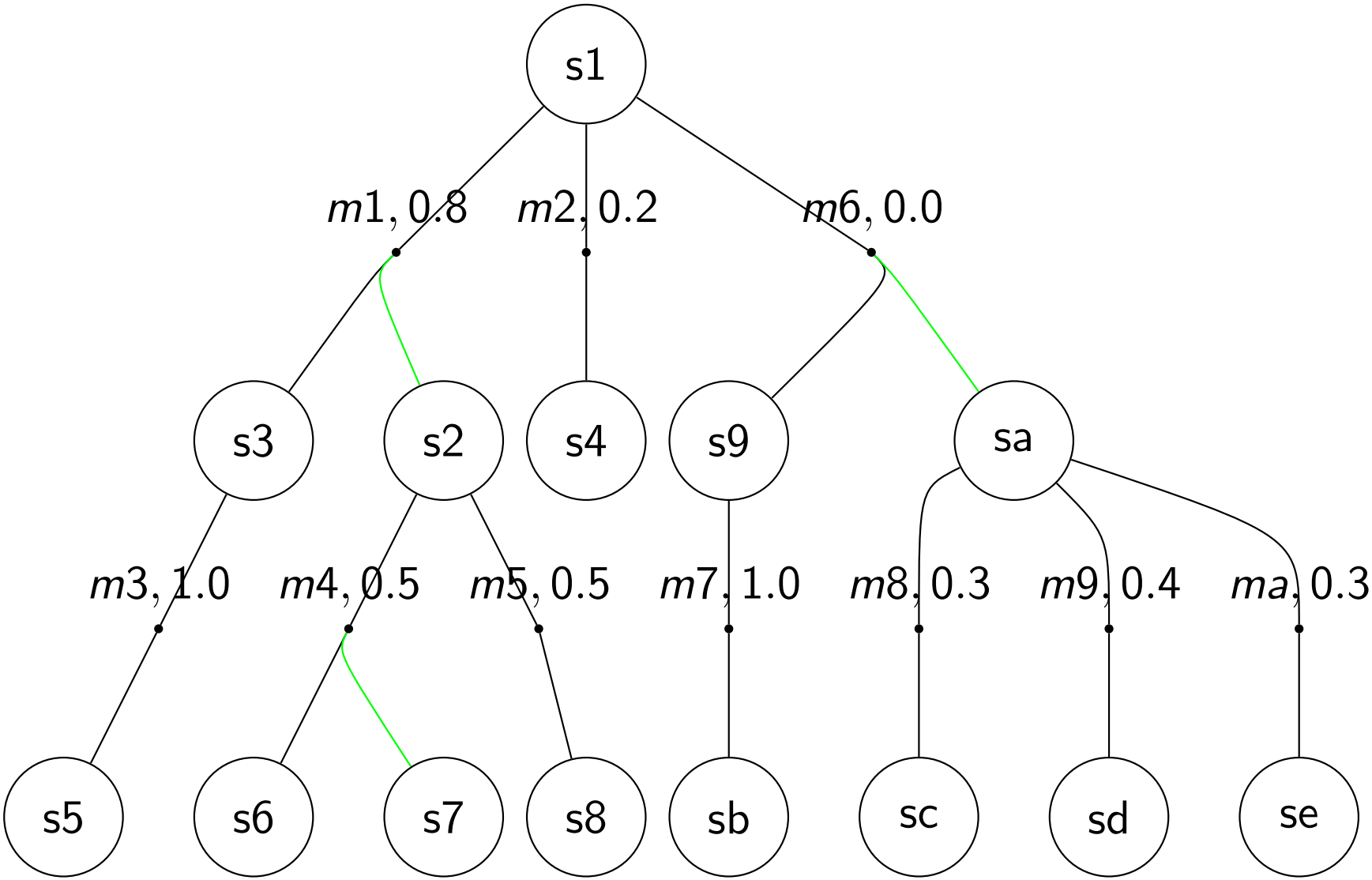
- Stany etykietowane potencjalnie osiągalnymi stanami followera
- Wykonanie ruchu oznacza ustalenie decyzji w strategii podstawowej.
- Wynikiem gry jest zredukowana strategia podstawowa

Przegląd strategii followera z użyciem UCT III

Przegląd

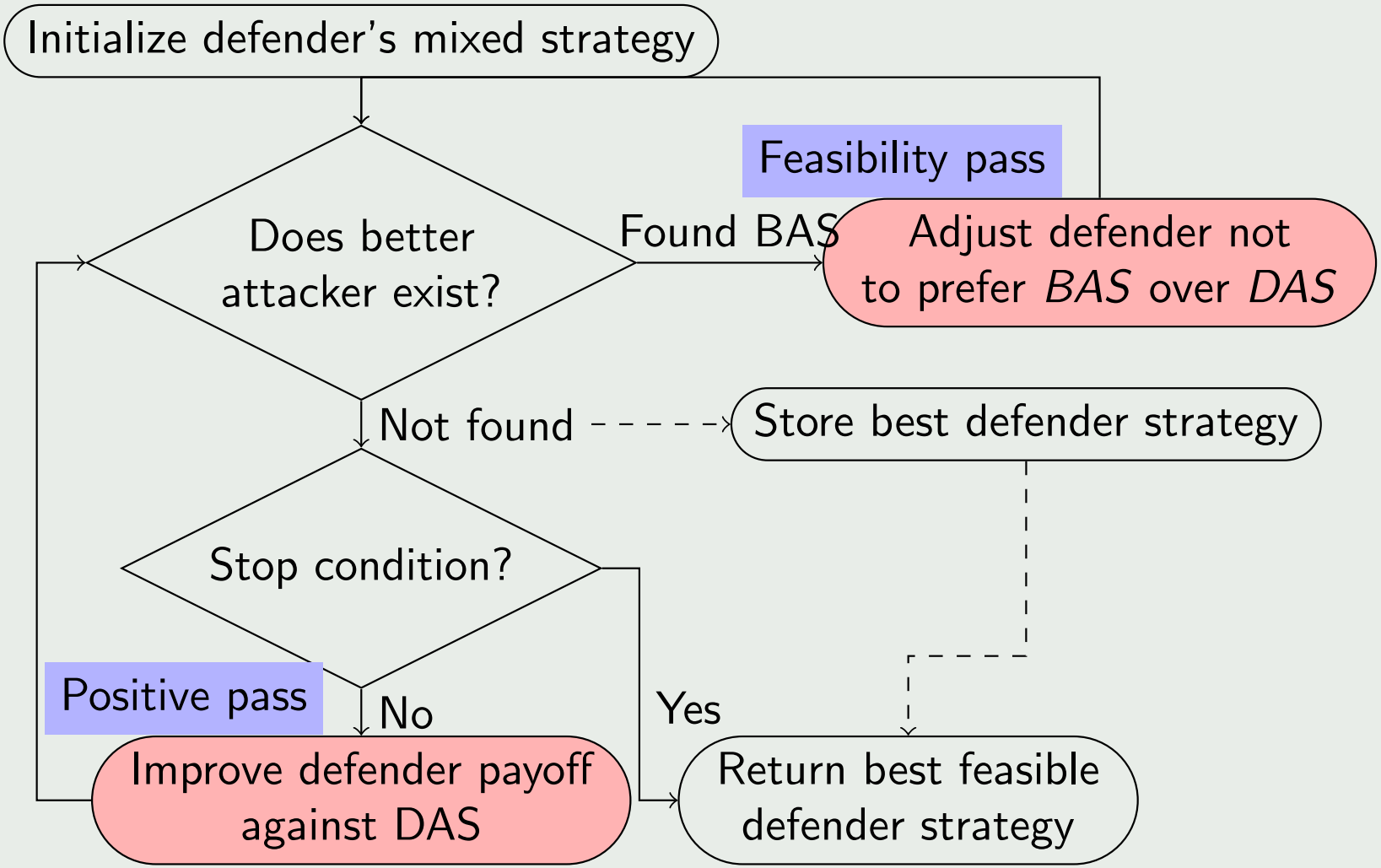


Drzewo strategii obrońcy



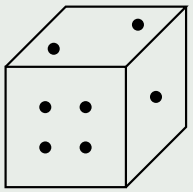
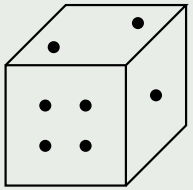
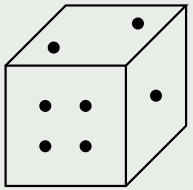
Dostosowanie strategii I

Zarys metody



Dostosowanie strategii II

Dostosowanie strategii w punkcie

1.  Dodaj nowy ruch wraz ze ścieżką do liścia wychodzący z węzła
2.  Zejdź do potomków
3.  Dostosuj strategię w węźle

Metoda dostosowywania węzła

Struktury danych

- Moment – wektor o długości równej liczbie ruchów
- Współczynnik normalizacji momentu

Data: $prob \in [0,1]^M$ probabilities vector, $mom \in (R^+ \cup \{0\})^M$ momentum vector, w – momentum normalization factor, $as \in \mathbb{R}^M$ assessments vector. All vectors contain values for i-th move at i-th position

$mom \leftarrow mom + as$;

$w \leftarrow w + L_1(as)$;

$prob \leftarrow max\{prob + mom/w, 0\}$// `max at each position`

$prob \leftarrow normalizeOrEqual(prob)$// `Normalize vector values so their sum is 1 or as a fallback assign uniform probability`

Problemy

- Czy zbiega do optymalnej strategii

MILP: metoda do porównań

MILP

$$
\begin{array}{rl}
 & \max_{q,z,a} \sum_{i\in X} \sum_{j\in Q} R_{ij} z_{ij} \\
\text{s.t.} & \sum_{i\in X} \sum_{j\in Q} z_{ij} = 1 \\
(\forall_{i\in X}) & \sum_{j\in Q} z_{i,j} \leqslant 1 \\
(\forall_{j\in Q}) & q_j \leqslant \sum_{i\in X} z_{ij} \leqslant 1 \\
 & \sum_{j\in Q} q_j = 1 \\
(\forall_{j\in Q}) & 0 \leqslant (a - \sum_{i\in X} C_{ij} (\sum_{h\in Q} z_{ih})) \\
(\forall_{j\in Q}) & (a - \sum_{i\in X} C_{ij} (\sum_{h\in Q} z_{ih})) \leqslant (1 - q_j) M
\end{array}
$$

$$
\begin{array}{l}
z_{ij} \in [0,1] \\
q_j \in \{0,1\} \\
a \in \mathbb{R}
\end{array}
$$

Złożoność

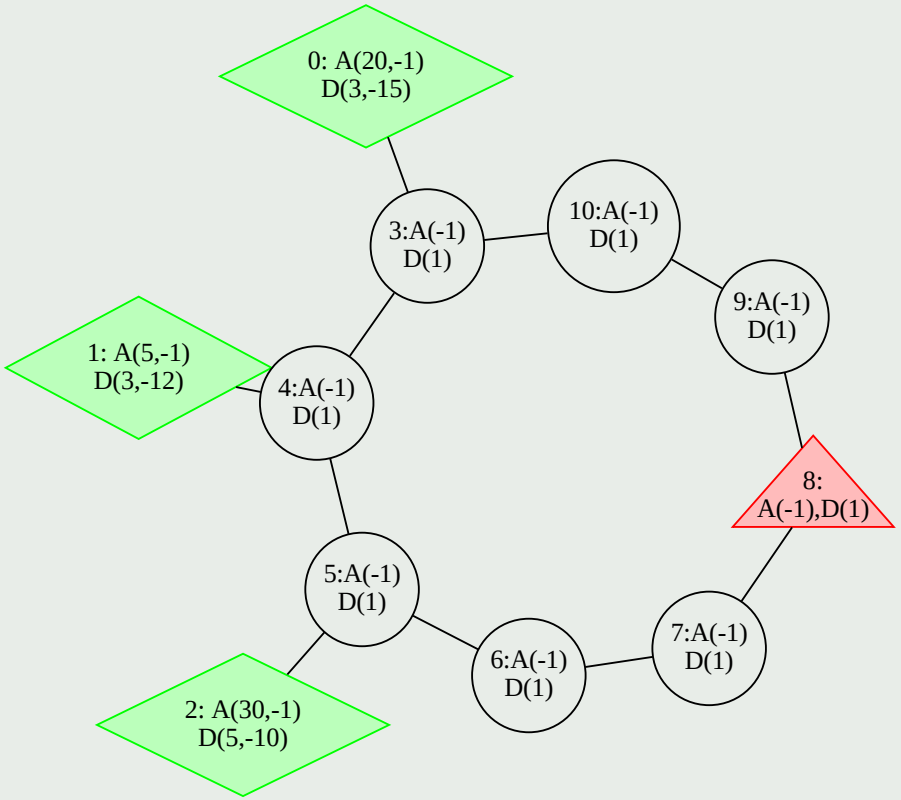
$|X|$ – liczba strategii (sekwencji) leadera, $|Q|$ – liczba strategii (sekwencji) followera

- $O(|X| \cdot |Q|)$ zmiennych, w tym $|Q|$ binarnych.
- $O(|X| + |Q|)$ ograniczeń

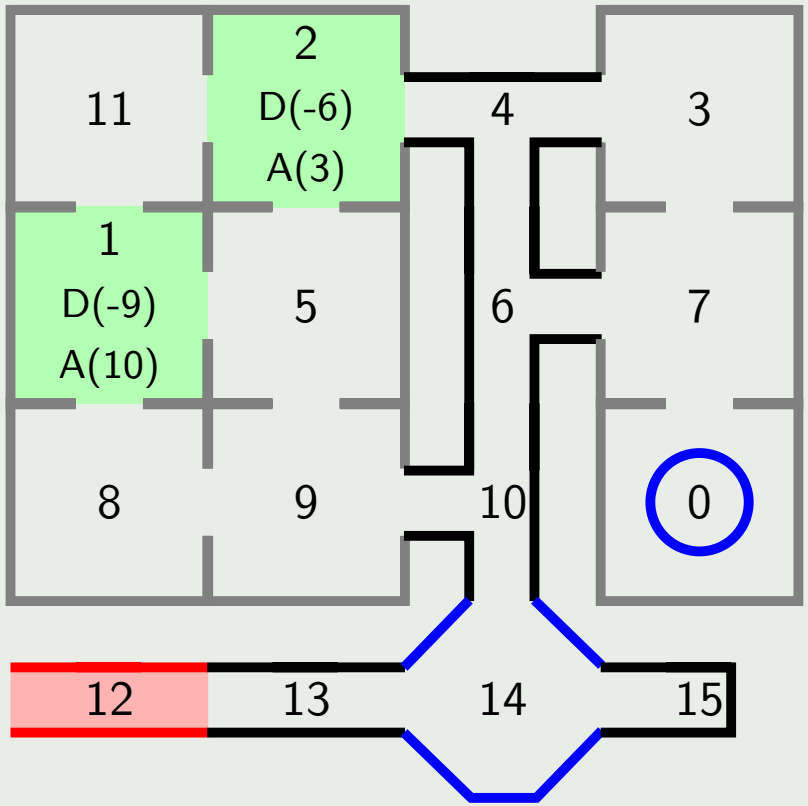
Gra na grafie i generator budynków

- Każdy z graczy operuje jednostkami chodzącymi po grafie
- Obaj wykonują ruch jednocześnie
- Nie widzą siebie nawzajem

Przykładowa gra



Generator budynków



game	New					MixedUCT		R	MILP		Sc	Sc
	payoff	p. max	sd	time	sd	payoff	time	payoff	payoff	time	MU	new
game1-5	0.54	0.54	0	3108	136	0.51	1778	-10.46	0.54	28053	1	1
game2-5	0.08	0.08	0	922	13	0.07	1661	-7.21	0.08	107	1	1
game3a-5	-4.5	-4.5	0	1299	33	-4.51	1480	-13.97	-4.5	31926	1	1
game3d-5	0.9	0.9	0	1047	15	0.9	1053	-4.61	0.9	37870	1	1
game6-5	-4.87	-4.87	0	1197	53	-4.87	1211	-13.91	-4.87	5313	1	1
game6a-5	-6	-6	0	1251	106	-6	1403	-13.84	-6	5949	1	1
game6b-5	0.79	0.79	0	992	33	0.76	851	-0.81	0.79	5547	0.98	1
sb-1-5	-0.76	-0.08	1.33	2888	85	-0.2	1151	-2.79	0.03	232	0.92	0.72
sb-2-5	0	0	0	1574	38	0	1724	-15.08	0	255	1	1
sb-20-5	1.29	1.29	0	1063	51	1	1050	-1.28	1.29	0	0.89	1
sb-24-5	1.54	1.54	0.01	1889	44	0.92	1874	0.08	1.55	0	0.57	0.99
sb-3-5	0.47	0.49	0.03	1166	58	0.5	1002	-8.55	0.5	0	1	1
sb-35-5	0.5	0.5	0	1027	22	0.5	1582	-0.89	0.5	339	1	1
sb-39-5	0.05	0.06	0.01	1423	77	0.03	2100	-17.06	0.09	0	1	1
sb-41-5	-2.88	-2.88	0	1255	28	-2.88	2240	-5.07	-2.85	0	0.99	0.99
sb-42-5	0	0	0	1947	44	0	1440	-16.53	0	0	1	1
sb-56-5	1.09	1.1	0.01	2638	48	1.01	1676	0.52	1.6	0	0.45	0.53
sb-59-5	0.6	0.6	0	2718	44	0.6	1321	-7.39	0.62	0	1	1
sb-64-5	0.12	0.13	0.01	1568	64	0.12	1698	-11.81	0.16	0	1	1
sb-7-5	-0.77	-0.77	0	1909	54	-1	1053	-9.7	-0.77	0	0.97	1
sb-78-5	0.5	0.5	0	2017	167	0.31	1208	-9.28	0.5	0	0.98	1
sb-82-5	0	0	0	3338	82	0	2498	-10.66	0	0	1	1
sb-85-5	-0.89	-0.89	0	1472	27	-0.95	1456	-1.79	-0.89	0	0.93	1
sb-87-5	0.8	0.8	0	936	23	0.7	1202	-1.66	0.8	0	0.96	1
sb-91-5	-5.62	-5.62	0	2325	51	-5.62	1852	-5.97	-5.62	0	1	1
sb-96-5	0.12	0.19	0.08	788	15	0.13	1876	-10.15	0.19	0	0.99	0.99
sb-98-5	0	0	0	930	8	0	1102	-13.02	-0.3	0	1.02	1.02

Dalsze badania

- Gwarancje zbieżności
- Skalowanie się wraz ze wzrostem liczby kroków w grze
- Porównanie skalowania z innymi metodami

*

(Koniec)