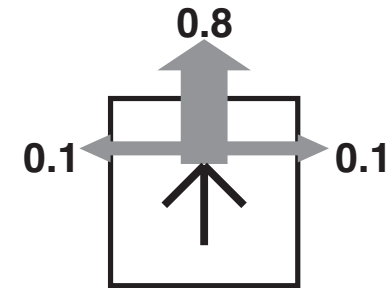
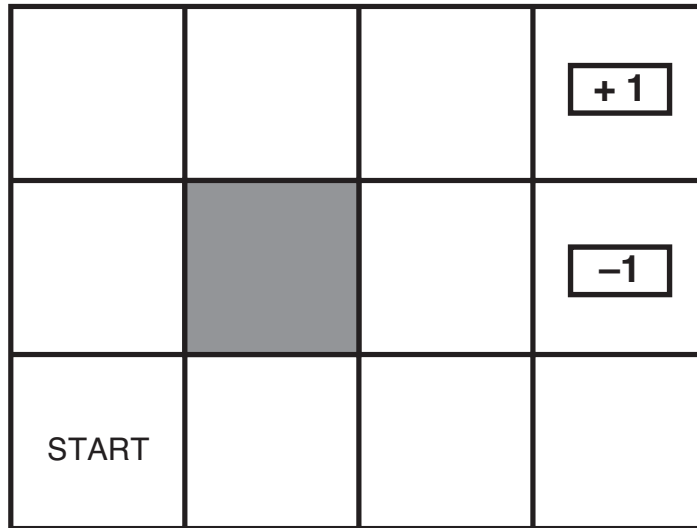


MDP úvod

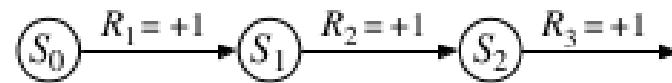
J. Kostlivá, Z. Straka, P. Švarný



Máme:

- Stavů: S
- Akce: A
- Přejchodový model: $T(s, a, s') \equiv P(s, a, s')$ pst, že ze stavu s při zvolení akce a skončíme ve stavu s'
- Reward: $r(s), r(s, a), r(s, a, s')$ okamžitá ocenění
- Policy: strategie chování robota/agenta

- Epizoda: sekvence stavů s rewards
- Return/Utility sekvence: $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$



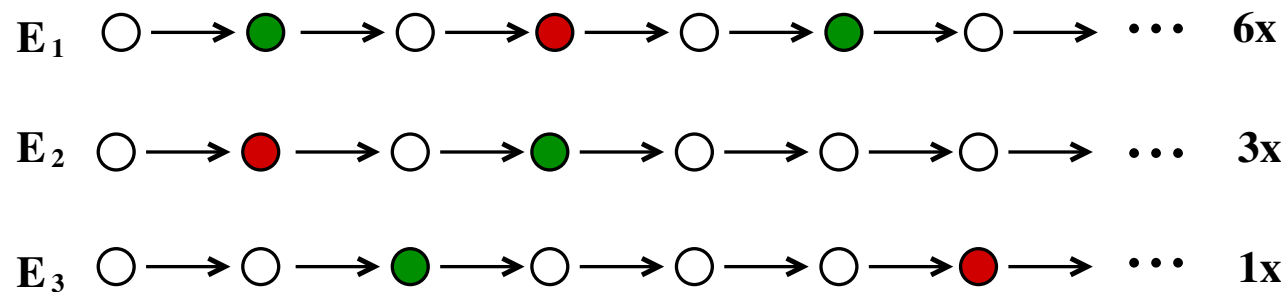
Policy Evaluation: Jak je dobrá daná strategie?

\mathbf{E}_1 ○ → ● → ○ → ● → ○ → ● → ○ → ... **6x**

\mathbf{E}_2 ○ → ● → ○ → ● → ○ → ○ → ○ → ... **3x**

\mathbf{E}_3 ○ → ○ → ● → ○ → ○ → ○ → ○ → ● → ... **1x**

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci:

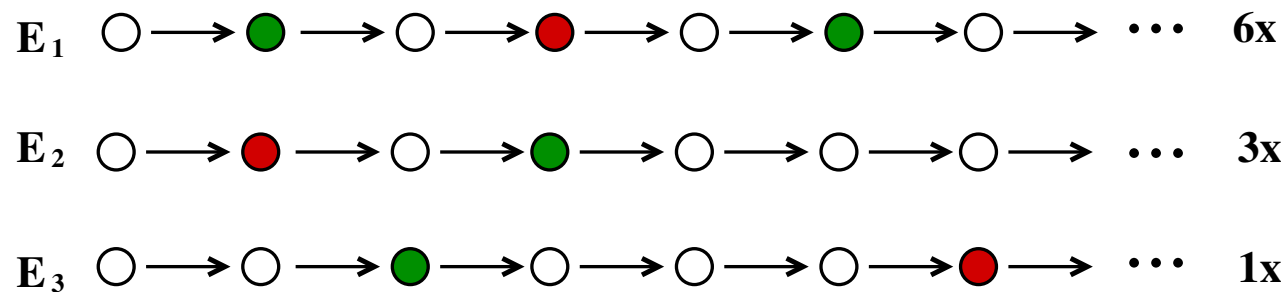
A: Hodnota stavu $V(s)$

B: Okamžitá odměna $r(s)$

C: Return/Utility G

D: Policy π

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$

○ **0** ● **1** ● **-0.3**

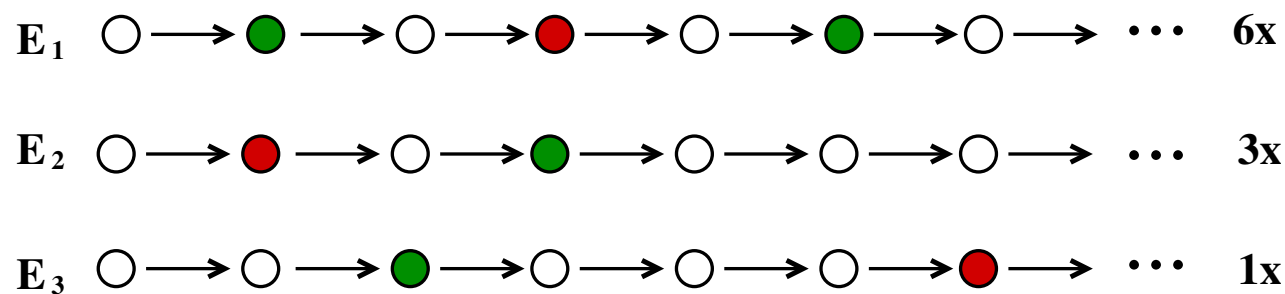
A: Hodnota stavu $V(s)$

B: Okamžitá odměna $r(s)$ ←

C: Return/Utility G

D: Policy π

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$

○ **0** ● **1** ● **-0.3**

2. Délka epizody:

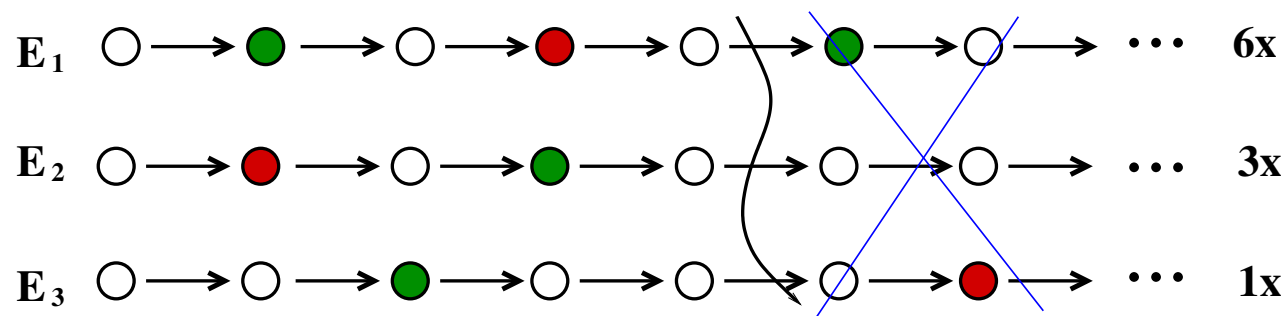
A: Nekonečná

B: Konečná

C: $T = 1000$

D: $T = 4$

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$

○ **0** ● **1** ● **-0.3**

2. Délka epizody: pro nás zvolme $T = 4$

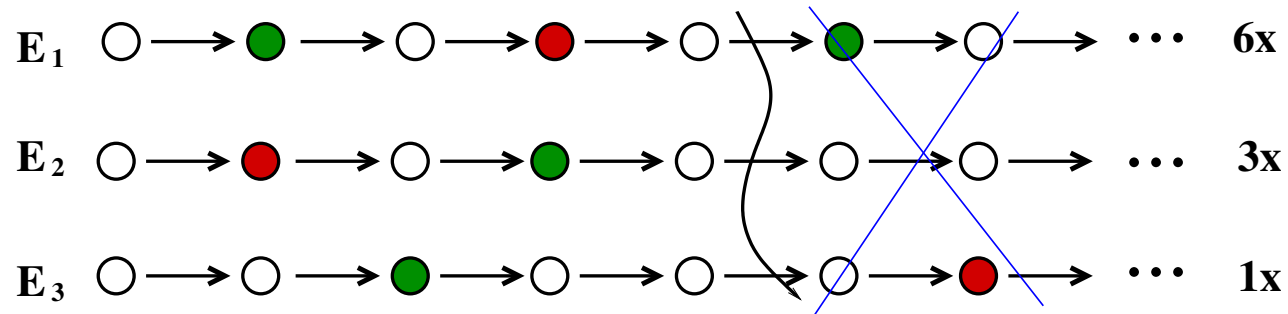
A: Nekonečná ⇐

B: Konečná ⇐

C: $T = 1000$ ⇐

D: $T = 4$ ⇐

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$

○ 0 ● 1 ● -0.3

2. Délka epizody: pro nás zvolme $T = 4$

3. Discount factor: γ

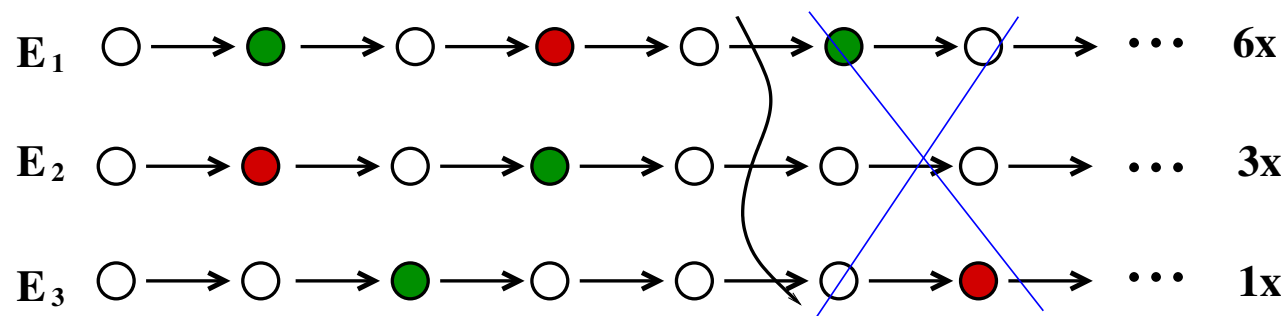
A: 1

B: 5

C: 0.8

D: 0.1

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$

○ **0** ● **1** ● **-0.3**

2. Délka epizody: pro nás zvolme $T = 4$

3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$

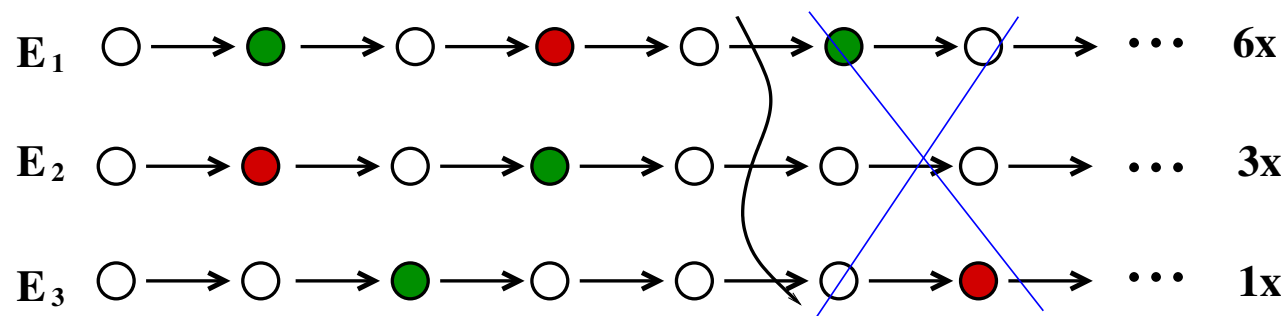
A: 1 ←

B: 5

C: 0.8 ←

D: 0.1 ←

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$

○ **0** ● **1** ● **-0.3**

2. Délka epizody: pro nás zvolme $T = 4$

3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$

4. Výpočet hodnoty epizody: return/utility G_t

A: $\sum_{n=1}^T \gamma^n$

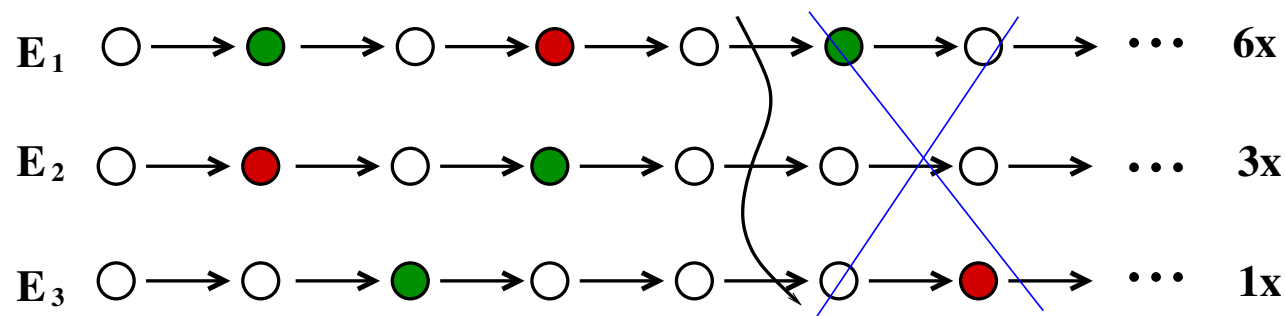
B: $\prod_{n=1}^T \gamma^n$

C: γr

D: $\prod_{n=1}^T \gamma^n r_n$

E: $\sum_{n=0}^T \gamma^n r_n$

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$

○ **0** ● **1** ● **-0.3**

2. Délka epizody: pro nás zvolme $T = 4$

3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$

4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$

A: $\sum_{n=1}^T \gamma^n$

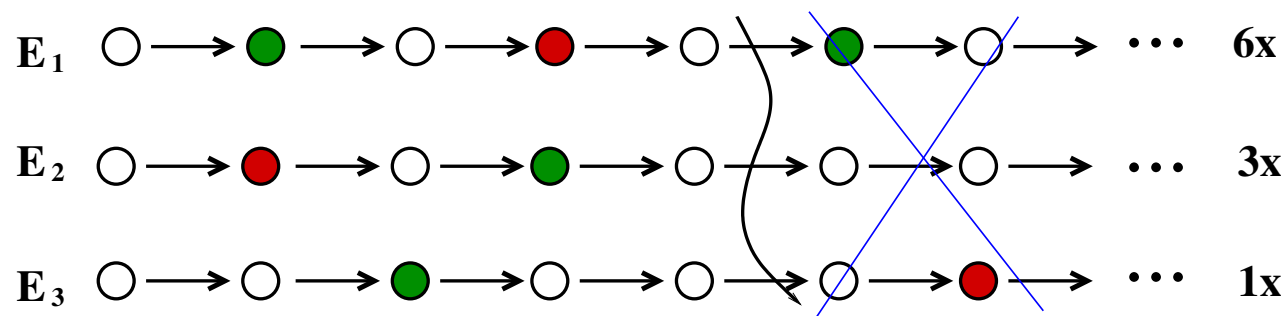
B: $\prod_{n=1}^T \gamma^n$

C: γr

D: $\prod_{n=1}^T \gamma^n r_n$

E: $\sum_{n=0}^T \gamma^n r_n$ \Leftarrow

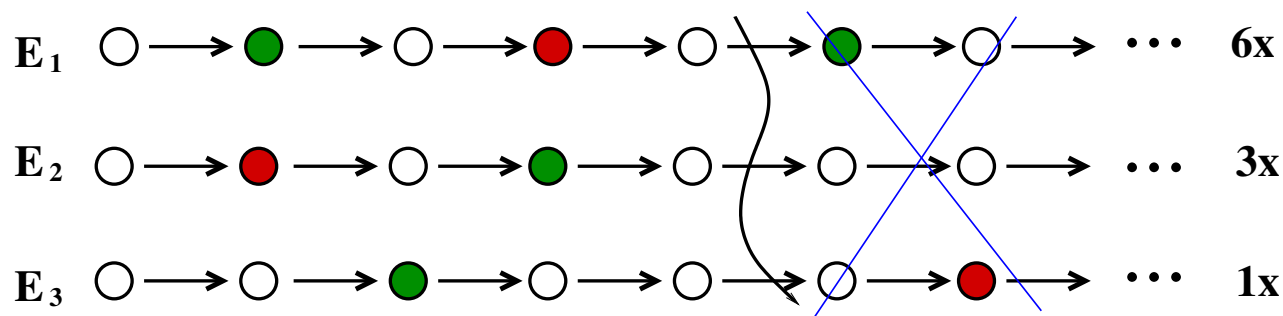
Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$ ○ **0** ● **1** ● **-0.3**
2. Délka epizody: pro nás zvolme $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$
4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$
 - A: $G(E_1) = 0.7$
 - B: $G(E_1) = 0.65$
 - C: $G(E_1) = 0.95$
 - D: $G(E_1) = 0.8$

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$ ○ **0** ● **1** ● **-0.3**

2. Délka epizody: pro nás zvolme $T = 4$

3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$

4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$

- $G(E_1) = 0.65$

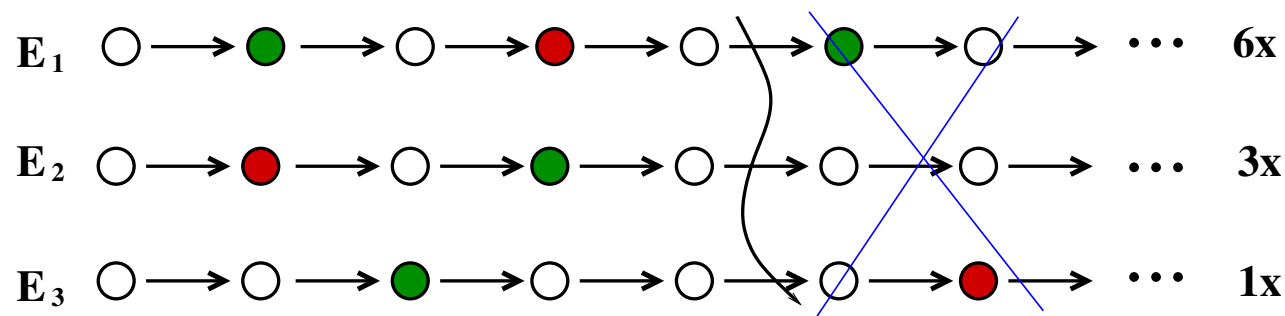
A: $G(E_1) = 0.7$

B: $G(E_1) = 0.65 = 0.8^0 * 0 + 0.8^1 * 1 + 0.8^2 * 0 + 0.8^3 * (-0.3) + 0.8^4 * 0 \quad \Leftarrow$

C: $G(E_1) = 0.95$

D: $G(E_1) = 0.8$

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$

○ **0** ● **1** ● **-0.3**

2. Délka epizody: pro nás zvolme $T = 4$

3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$

4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$

• $G(E_1) = 0.65$

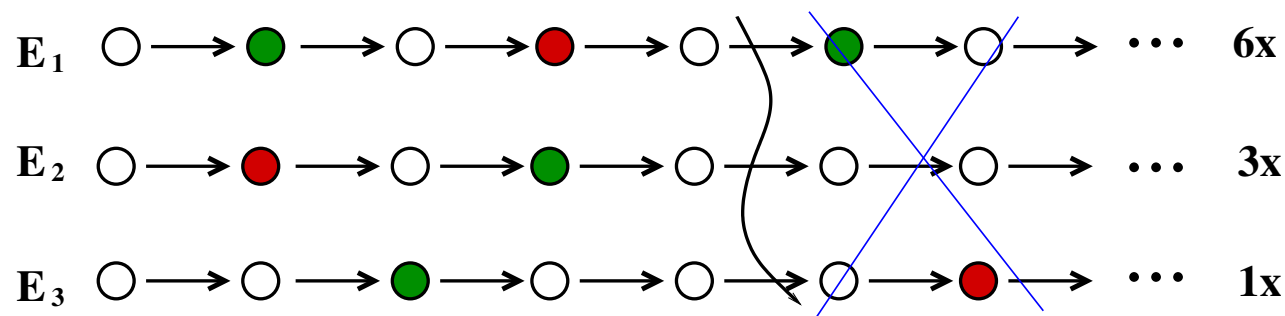
A: $G(E_2) = 0.272$

B: $G(E_2) = 0.4$

C: $G(E_2) = 0.7$

D: $G(E_2) = 0.99$

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$ ○ **0** ● **1** ● **-0.3**

2. Délka epizody: pro nás zvolme $T = 4$

3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$

4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$

• $G(E_1) = 0.65$, $G(E_2) = 0.272$

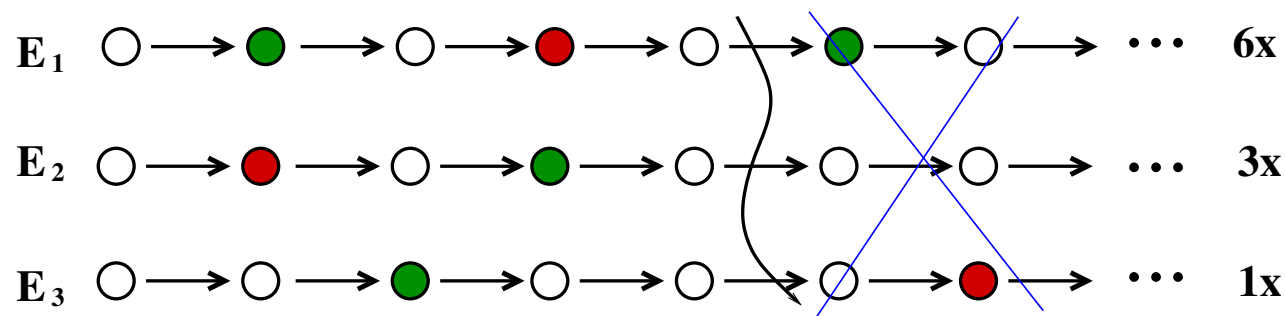
A: $G(E_2) = 0.272 = 0.8^1 * (-0.3) + 0.8^3 * 1$ \Leftarrow

B: $G(E_2) = 0.4$

C: $G(E_2) = 0.7$

D: $G(E_2) = 0.99$

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$

○ **0** ● **1** ● **-0.3**

2. Délka epizody: pro nás zvolme $T = 4$

3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$

4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$

• $G(E_1) = 0.65$, $G(E_2) = 0.272$

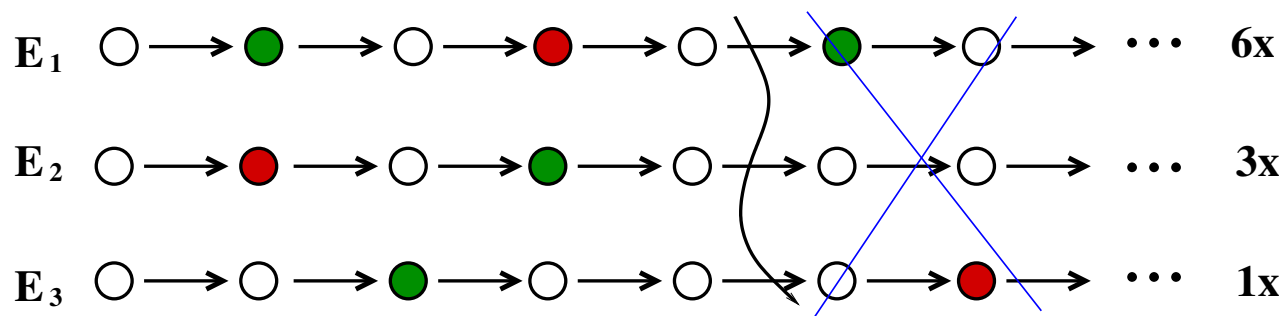
A: $G(E_3) = -0.3$

B: $G(E_3) = 0.7$

C: $G(E_3) = 0.64$

D: $G(E_3) = 0.8$

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$ ○ **0** ● **1** ● **-0.3**

2. Délka epizody: pro nás zvolme $T = 4$

3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$

4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$

• $G(E_1) = 0.65$, $G(E_2) = 0.272$, $G(E_3) = 0.64$

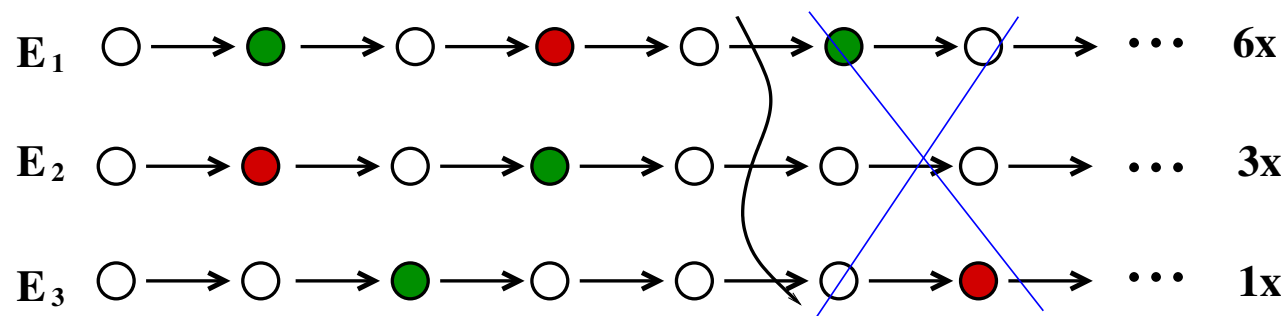
A: $G(E_3) = -0.3$

B: $G(E_3) = 0.7$

C: $G(E_3) = 0.64 = 0.8^2 * 1$ ←

D: $G(E_3) = 0.8$

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$ ○ **0** ● **1** ● **-0.3**

2. Délka epizody: pro nás zvolme $T = 4$

3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$

4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$

• $G(E_1) = 0.65$, $G(E_2) = 0.272$, $G(E_3) = 0.64$

5. Výpočet pro celou policy:

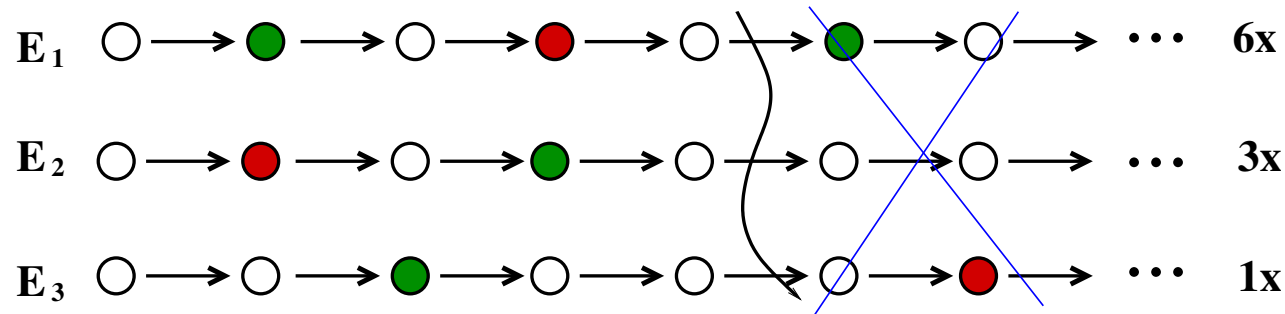
A: $\sum_{e=1}^E \sum_{n=0}^T \gamma^n r_n$

B: $\prod_{e=1}^E \sum_{n=0}^T \gamma^n r_n$

C: $\sum_{e=1}^E p_e \sum_{n=0}^T \gamma^n r_n$

D: $\max p_e \sum_{n=0}^T \gamma^n r_n$

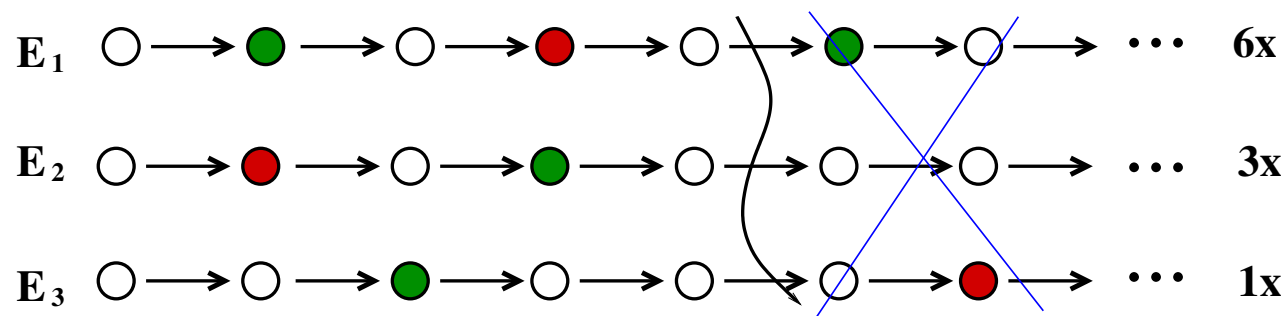
Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$ ○ **0** ● **1** ● **-0.3**
2. Délka epizody: pro nás zvolme $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$
4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$
 - $G(E_1) = 0.65$, $G(E_2) = 0.272$, $G(E_3) = 0.64$
5. Výpočet pro celou policy: $\sum_{e=1}^E p_e \sum_{n=0}^T \gamma^n r_n$
 - A: $\sum_{e=1}^E \sum_{n=0}^T \gamma^n r_n$
 - B: $\prod_{e=1}^E \sum_{n=0}^T \gamma^n r_n$
 - C: $\sum_{e=1}^E p_e \sum_{n=0}^T \gamma^n r_n \quad \Leftarrow$
 - D: $\max p_e \sum_{n=0}^T \gamma^n r_n$

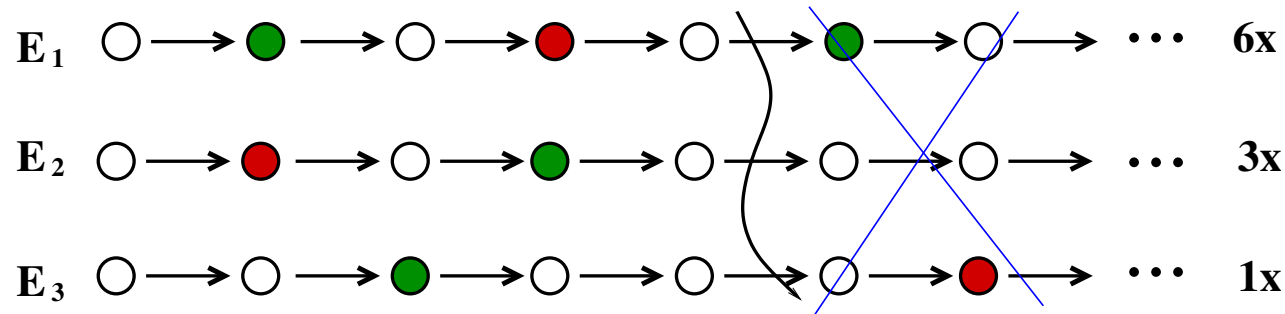
Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$ ○ **0** ● **1** ● **-0.3**
2. Délka epizody: pro nás zvolme $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$
4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$
 - $G(E_1) = 0.65$, $G(E_2) = 0.272$, $G(E_3) = 0.64$
5. Výpočet pro celou policy: $\sum_{e=1}^E p_e \sum_{n=0}^T \gamma^n r_n$
 - A: 0.535
 - B: 1.562
 - C: 1
 - D: 0.86

Policy Evaluation: Jak je dobrá daná strategie?



Co potřebujeme?

1. Ohodnocení stavu v sekvenci: Okamžitá odměna, reward function $r(s)$ ○ **0** ● **1** ● **-0.3**
2. Délka epizody: pro nás zvolme $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, pro nás zvolme $\gamma = 0.8$
4. Výpočet hodnoty epizody: return/utility $G_t = \sum_{n=0}^T \gamma^n r_n$
 - $G(E_1) = 0.65$, $G(E_2) = 0.272$, $G(E_3) = 0.64$
5. Výpočet pro celou policy: $\sum_{e=1}^E p_e \sum_{n=0}^T \gamma^n r_n = \mathbf{0.535}$
 - A: $0.535 = 0.6 * 0.65 + 0.3 * 0.272 + 0.1 * 0.64$ ←
 - B: 1.562
 - C: 1
 - D: 0.86