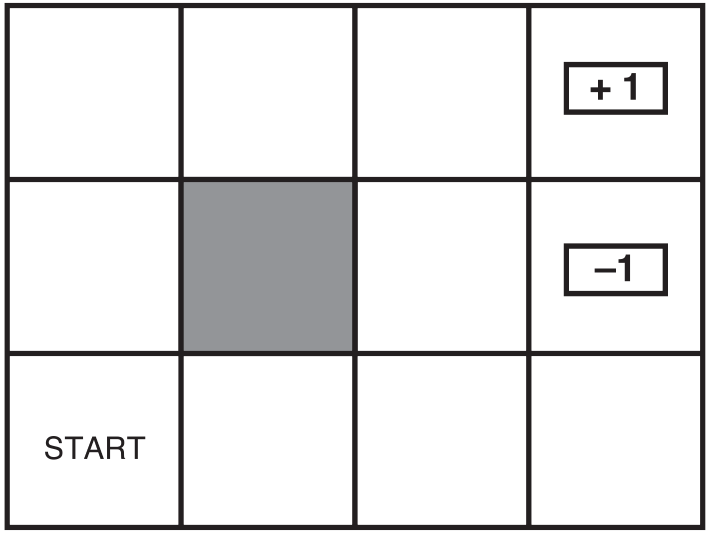
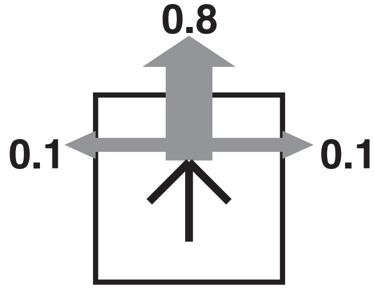
MDP Introduction

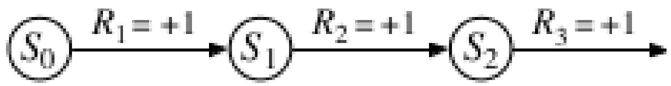
F. Gama, S. Dantu





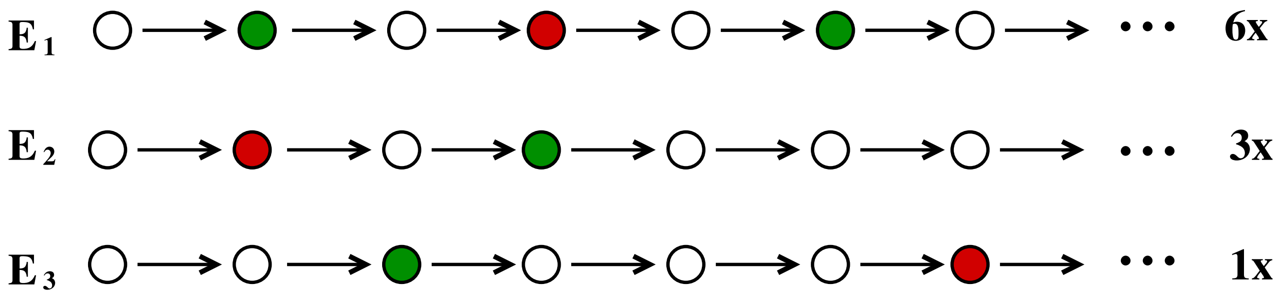
We have:

- State: S
- Action: A
- Transition model: $T(s, a, s') \equiv P(s, a, s')$, we are in state s, make action a, and arrive in state s'
- Reward: $r(s)$, $r(s, a), r(s, a, s')$ immediate reward/evaluation
- Policy: agent/robot behaviour strategy

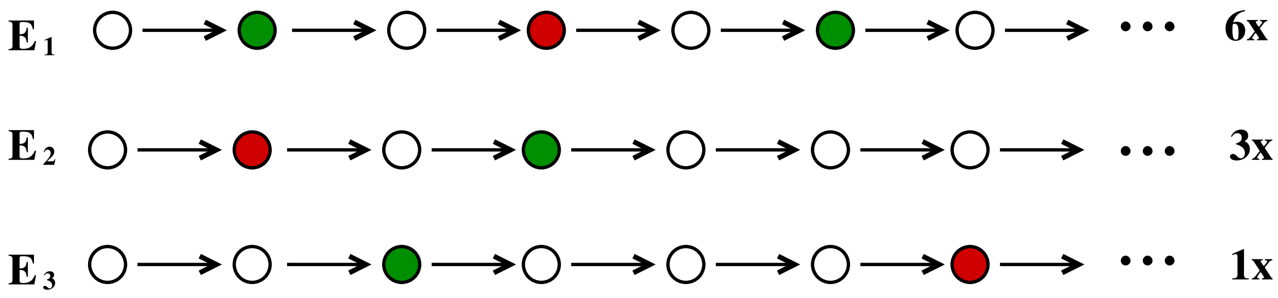


- Episode: sequence of states with rewards
- Return/Utility sequence: $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

Policy Evaluation: How good is the strategy?



Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence:

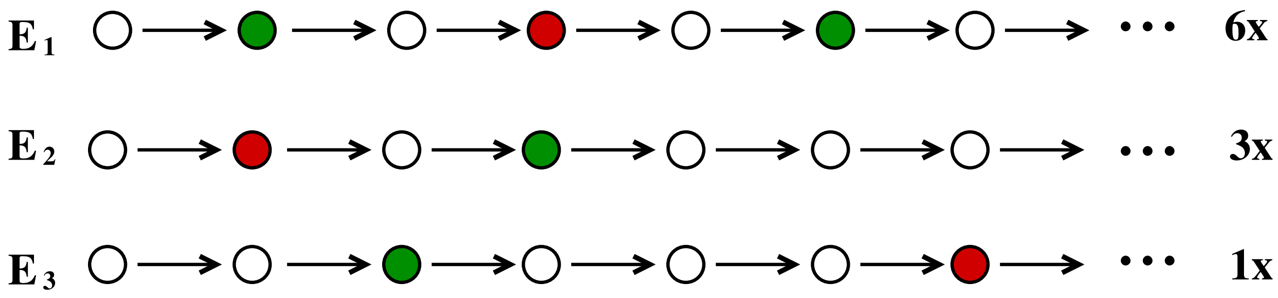
   A: State Value $V(s)$

   B: Immediate reward $r(s)$

   C: Return/Utility G

   D: Policy π

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$

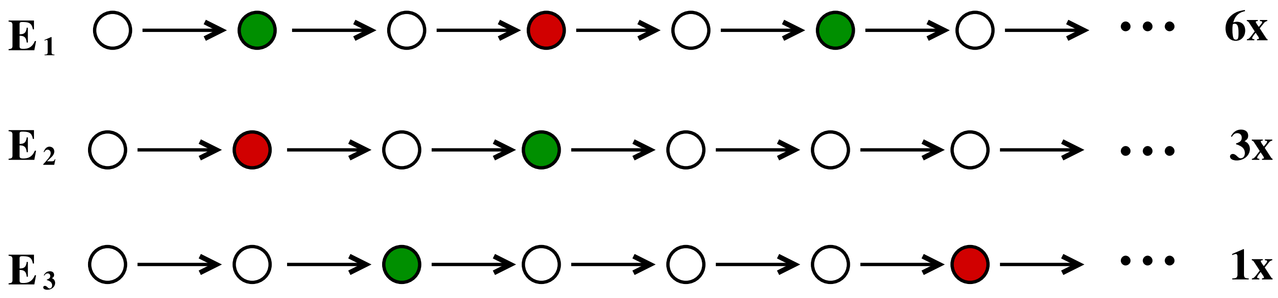
   A: State Value $V(s)$

   B: Immediate reward $r(s)$ $\Leftarrow$

   C: Return/Utility G

   D: Policy π

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$

2. Episode length::

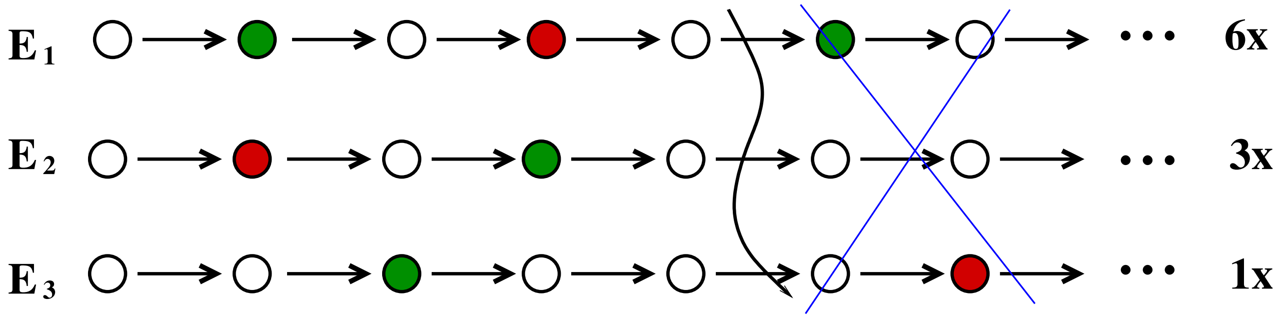
   A: Infinite

   B: Finite

   C: $T = 1000$

   D: $T = 4$

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$

2. Episode length: We'll chose $T = 4$

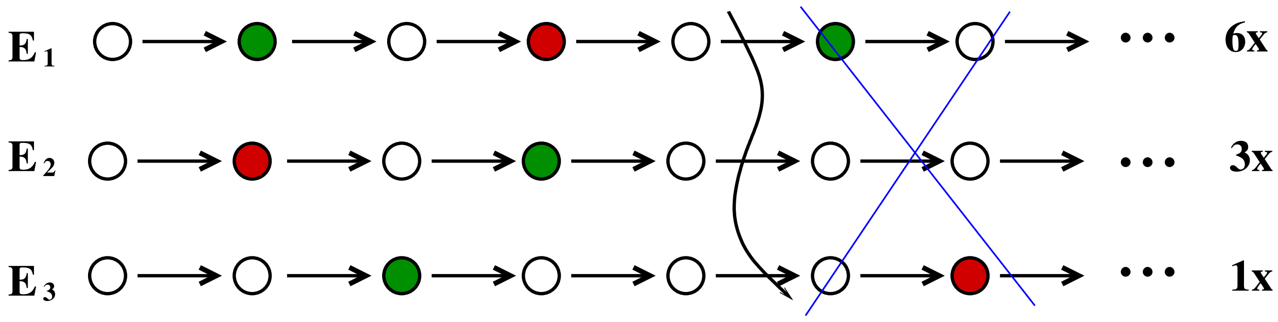
   A: Infinite $\Leftarrow$

   B: Finite $\Leftarrow$

   C: $T = 1000$ $\Leftarrow$

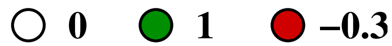
   D: $T = 4$ $\Leftarrow$

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$



2. Episode length: We chose $T = 4$
3. Discount factor: γ

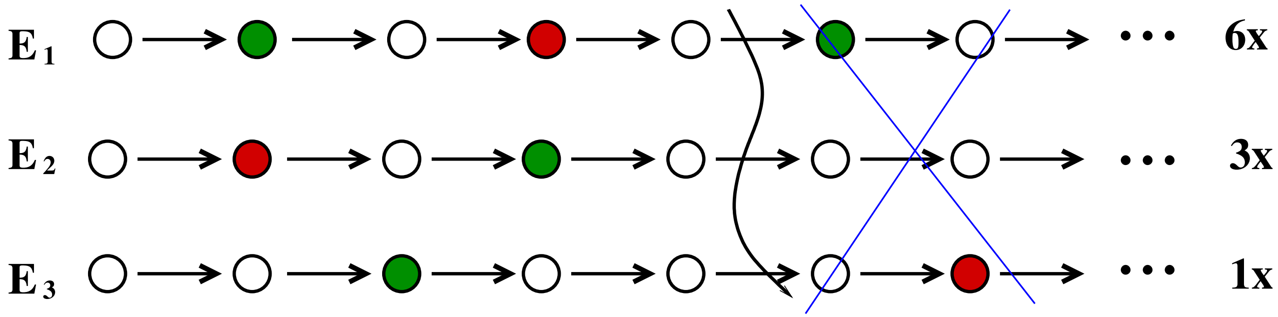
   A: 1

   B: 5

   C: 0.8

   D: 0.1

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$
2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$

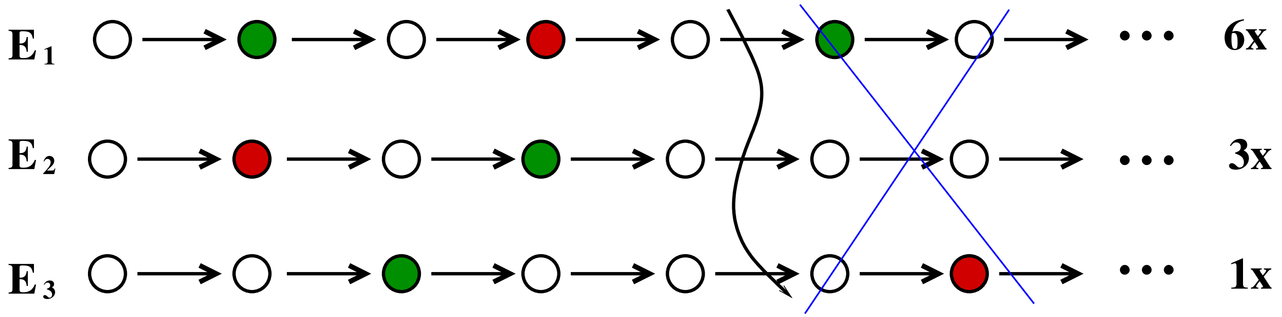
   A: 1 $\Leftarrow$

   B: 5

   C: 0.8 $\Leftarrow$

   D: 0.1 $\Leftarrow$

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$
2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: return/utility G_t

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

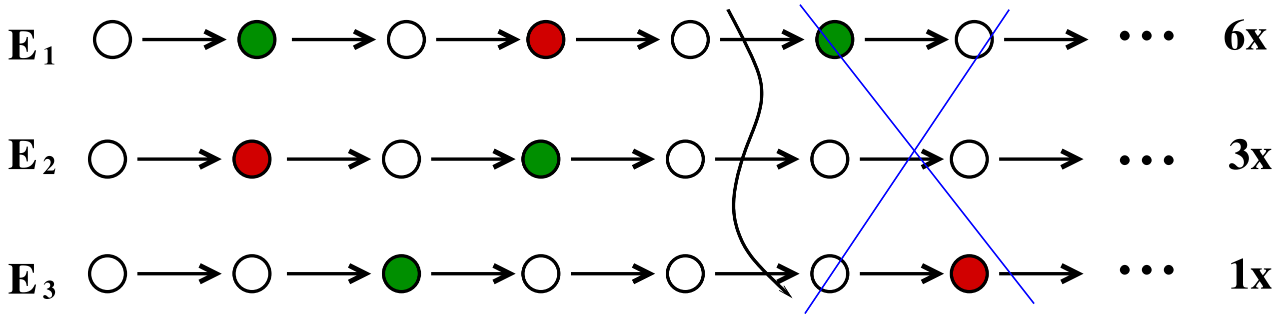
   B: $\prod_{n=1}^{T} \gamma^n$

   C: γr

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

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

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$
2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: return/utility $G_t = \sum_{n=0}^{T} \gamma^n r_n$

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

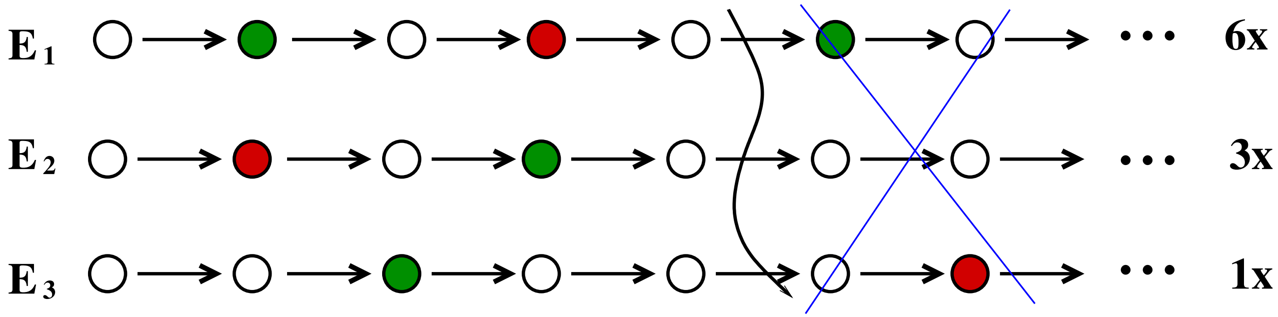
   B: $\prod_{n=1}^{T} \gamma^n$

   C: γr

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

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

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$

2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: return/utility $G_t = \sum_{n=0}^{T} \gamma^n r_n$

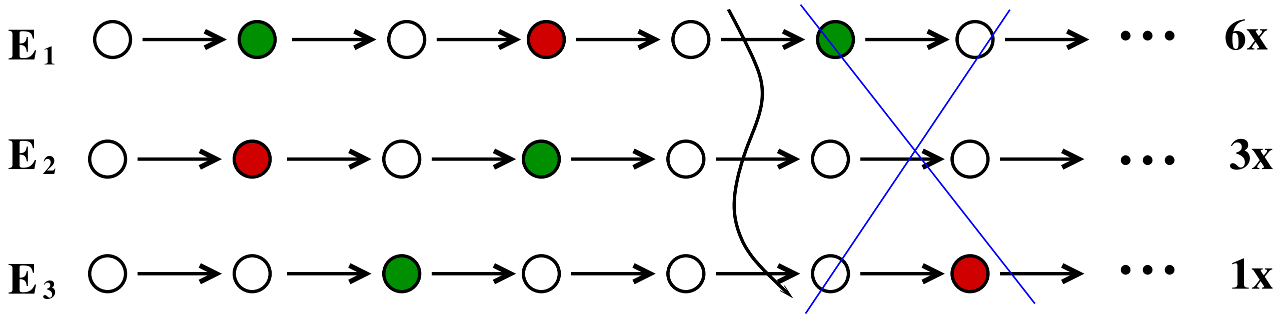
   A: $G(E_1) = 0.7$

   B: $G(E_1) = 0.65$

   C: $G(E_1) = 0.95$

   D: $G(E_1) = 0.8$

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$
2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: return/utility $G_t = \sum_{n=0}^{T} \gamma^n r_n$
   - $G(E_1) = 0.65$

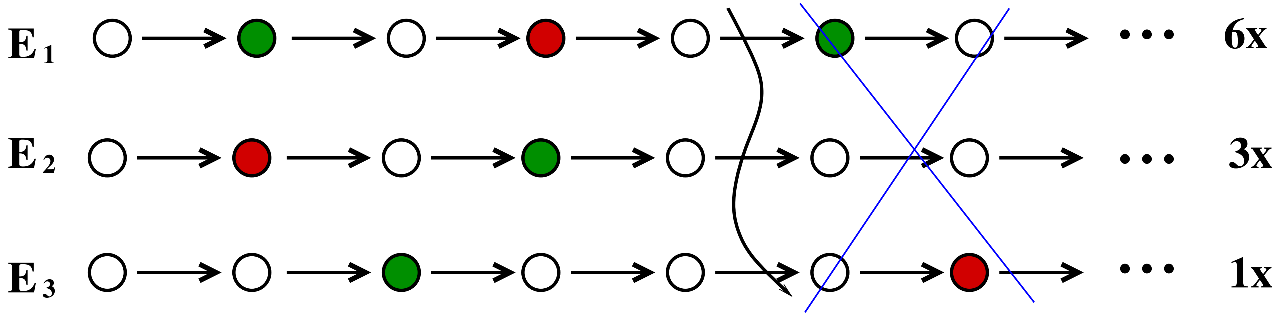
   A: $G(E_1) = 0.7$

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

   C: $G(E_1) = 0.95$

   D: $G(E_1) = 0.8$

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$

2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: return/utility $G_t = \sum_{n=0}^{T} \gamma^n r_n$
   - $G(E_1) = 0.65$

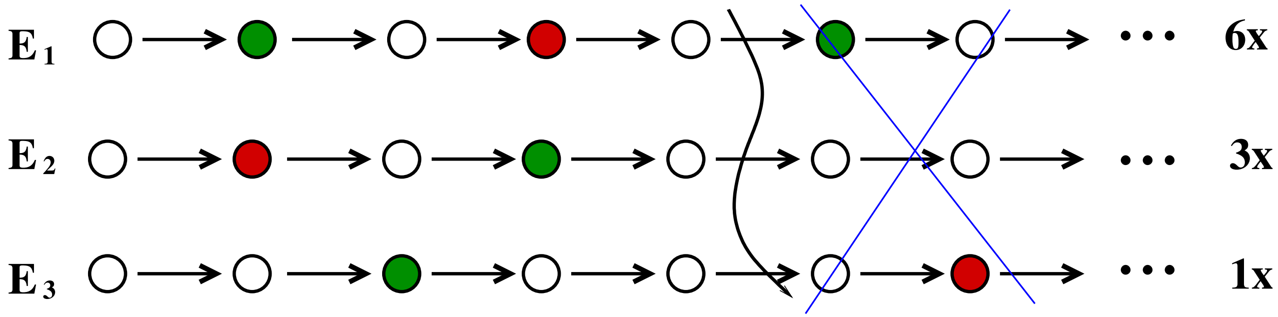
   A: $G(E_2) = 0.272$

   B: $G(E_2) = 0.4$

   C: $G(E_2) = 0.7$

   D: $G(E_2) = 0.99$

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$ — ○ 0, ● 1, ● −0.3
2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: return/utility $G_t = \sum_{n=0}^{T} \gamma^n r_n$
   - $G(E_1) = 0.65$, $G(E_2) = 0.272$

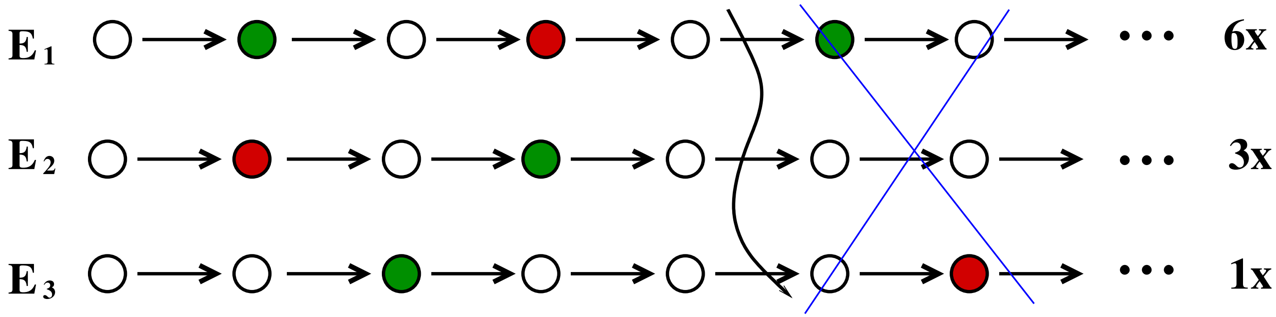
   A: $G(E_2) = 0.272 = 0.8^1 \cdot (-0.3) + 0.8^3 \cdot 1 \quad \Leftarrow$

   B: $G(E_2) = 0.4$

   C: $G(E_2) = 0.7$

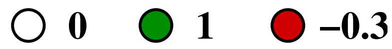
   D: $G(E_2) = 0.99$

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$



2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: return/utility $G_t = \sum_{n=0}^{T} \gamma^n r_n$
   - $G(E_1) = 0.65$, $G(E_2) = 0.272$

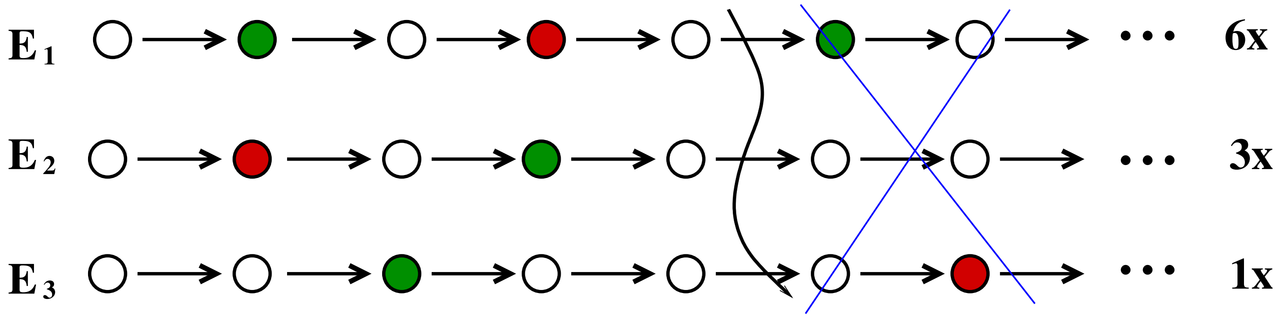
   A: $G(E_3) = -0.3$

   B: $G(E_3) = 0.7$

   C: $G(E_3) = 0.64$

   D: $G(E_3) = 0.8$

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$
2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: return/utility $G_t = \sum_{n=0}^{T} \gamma^n r_n$
   - $G(E_1) = 0.65$, $G(E_2) = 0.272$, $G(E_3) = 0.64$

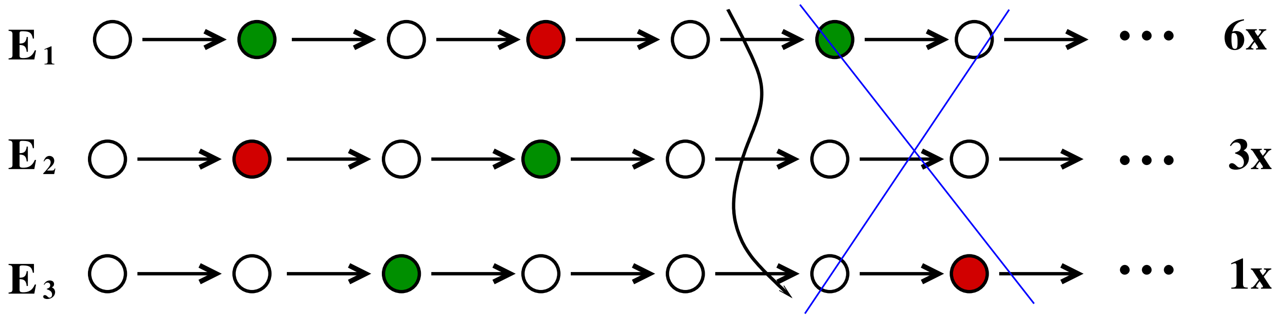
   A: $G(E_3) = -0.3$

   B: $G(E_3) = 0.7$

   C: $G(E_3) = 0.64 = 0.8^2 \cdot 1$ $\Leftarrow$

   D: $G(E_3) = 0.8$

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$

2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: 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. Calculation for the whole policy:

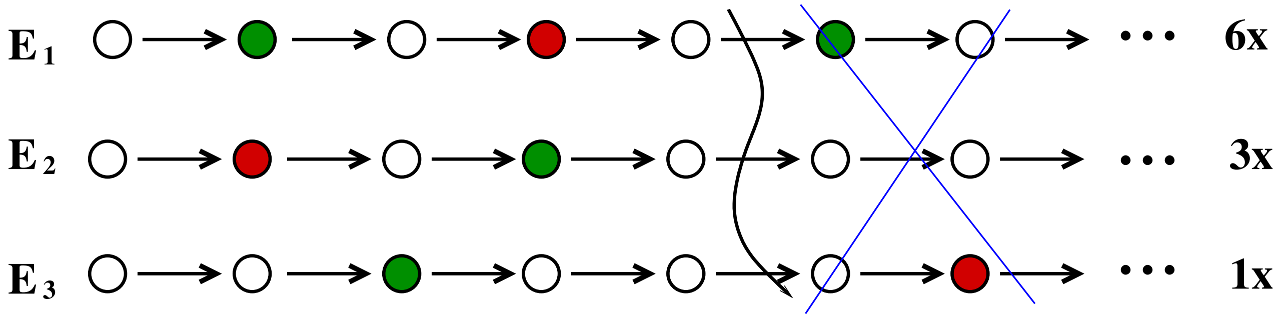
   A: $\sum_{e=1}^{E} \sum_{n=0}^{T} \gamma^n r_n$

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

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

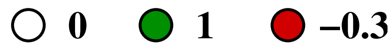
   D: $\max p_e \sum_{n=0}^{T} \gamma^n r_n$

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$



2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: 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. Calculation for the whole policy: $\sum_{e=1}^{E} p_e \sum_{n=0}^{T} \gamma^n r_n$

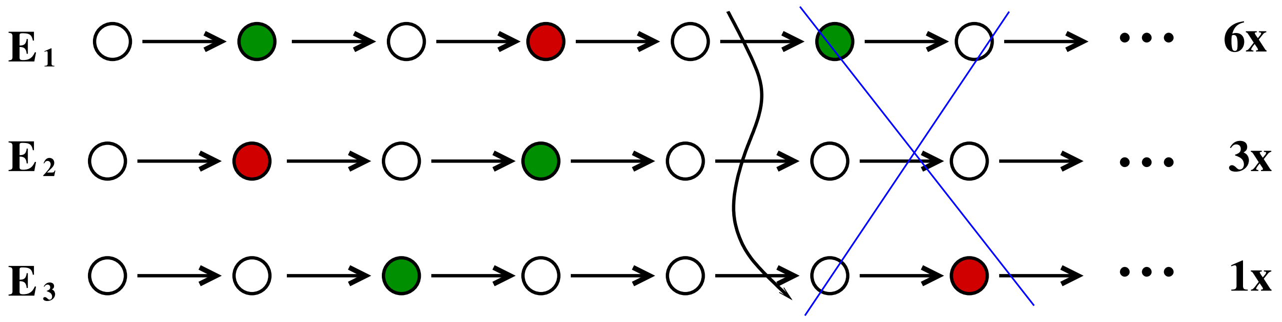
   A: $\sum_{e=1}^{E} \sum_{n=0}^{T} \gamma^n r_n$

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

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

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

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$

2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: 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. Calculation for the whole policy: $\sum_{e=1}^{E} p_e \sum_{n=0}^{T} \gamma^n r_n$

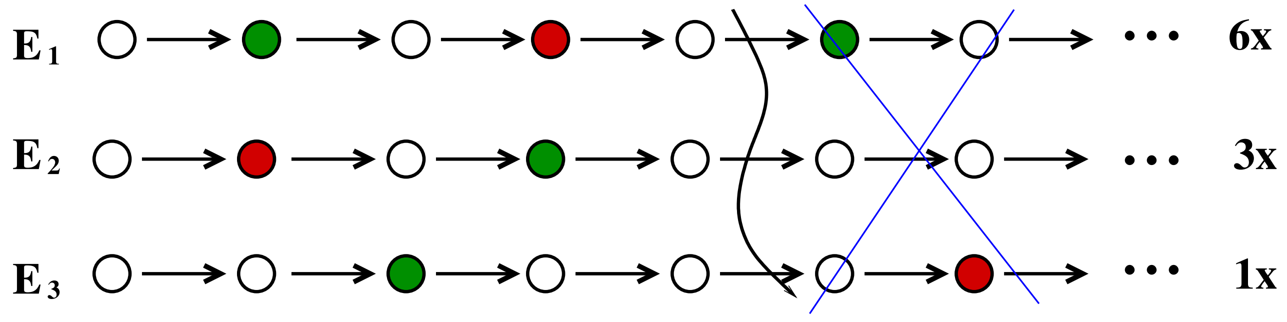
   A: 0.535

   B: 1.562

   C: 1

   D: 0.86

Policy Evaluation: How good is the strategy?



What do we need?

1. Evaluation of the state in the sequence: Immediate reward, reward function $r(s)$
2. Episode length: We chose $T = 4$
3. Discount factor: $0 \leq \gamma \leq 1$, we'll chose $\gamma = 0.8$
4. Episode value calculation: 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. Calculation for the whole policy: $\sum_{e=1}^{E} p_e \sum_{n=0}^{T} \gamma^n r_n = \mathbf{0.535}$

   A: $0.535 = 0.6 \cdot 0.65 + 0.3 \cdot 0.272 + 0.1 \cdot 0.64$ $\Leftarrow$

   B: 1.562

   C: 1

   D: 0.86