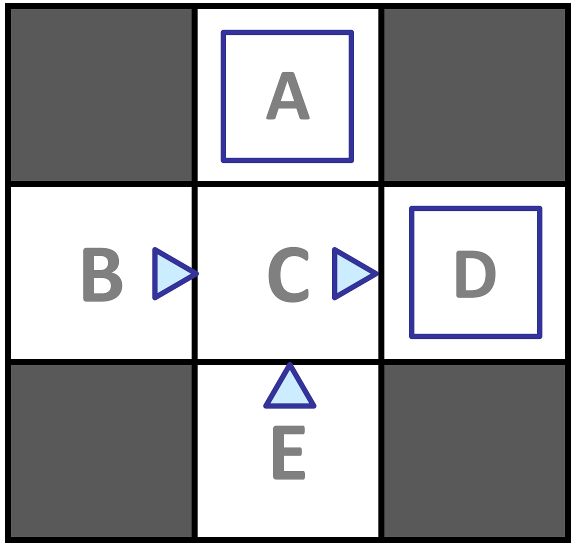
Example: Direct Evaluation

Input Policy π



Assume: γ = 1

Observed Episodes (Training)

Episode 1

B, east, C, -1
C, east, D, -1
D, exit, x, +10

+9 +8

Episode 2

B, east, C, -1
C, east, D, -1
D, exit, x, +10

+9 +8

Episode 3

E, north, C, -1
C, east, D, -1
D, exit, x, +10

+9

Episode 4

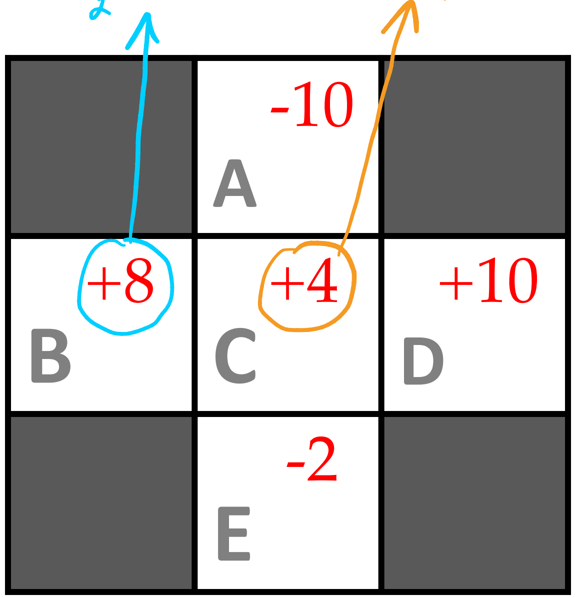
E, north, C, -1
C, east, A, -1
A, exit, x, -10

−11

Output Values

$\frac{8 \cdot 2}{2} = 8$

$\frac{9 \cdot 3 - 11}{4} = 4$



If B and E both go to C under this policy, how can their values be different?