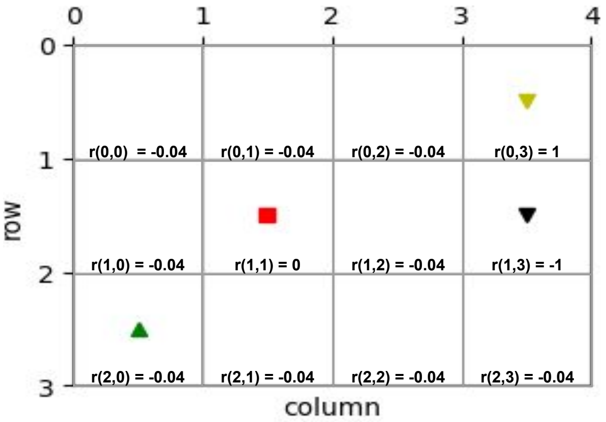

Before J=0 started

V(0,0)=0 N(0,0)=0	V(0,1)=0 N(0,1)=0	V(0,2)=0 N(0,2)=0	V(0,3)=0 N(0,3)=0
V(1,0)=0 N(1,0)=0	V(1,1)=0 N(1,1)=0	V(1,2)=0 N(1,2)=0	V(1,3)=0 N(1,3)=0
V(2,0)=0 N(2,0)=0	V(2,1)=0 N(2,1)=0	V(2,2)=0 N(2,2)=0	V(2,3)=0 N(2,3)=0

After J=0 finished

V(x,y) Update

V(0,0)=0.80298 N(0,0)=1	V(0,1)=0.851495 N(0,1)=1	V(0,2)=0.95 N(0,2)=1	V(0,3)=1 N(0,3)=1
V(1,0)=0.70740075 N(1,0)=1	V(1,1)=0 N(1,1)=0	V(1,2)=0 N(1,2)=0	V(1,3)=0 N(1,3)=0
V(2,0)=0.66032674 N(2,0)=1	V(2,1)=0 N(2,1)=0	V(2,2)=0 N(2,2)=0	V(2,3)=0 N(2,3)=0

Trajectory τ

There, are 12 states in this system and 4 possible actions. Each sample S_k in the trajectory $\tau = \{S_1, S_2, S_3, S_4, \dots, S_k\}$ consists (X_k, r_{k+1}) where k <= 20 (episode length)

At, J = 0, τ = {([2, 0], -0.04), ([1, 0], -0.04), ([1, 0], -0.04), ([0, 0], -0.04), ([0, 1], -0.04), ([0, 1], -0.04), ([0, 2], -0.04), ([0, 3], 1.0) }, where **k = 8**

$G_{k+1} = 0$, $\gamma = 0.99$
V(2,0), V(1,0), V(0,0), V(0,1), V(0,2), V(0,3) will be updated based on the criterion, $V_k(x,y) \leftarrow V_k(x,y) + [G_k - V_k(x,y)]/N_k(x,y)$

For length(τ):

- X, r = τ[-1] => [0, 3], 1.0
- $G_k = \gamma * G_{k+1} + R_{k+1} = 0.99 * 0 + 1.0 = 1.0 = \gamma * 0 + R_{k+1}$
- *If X=[0,3] is never visited before (not present in remaining list):*
  - N(X=[0,3]) += 1
  - V(X=[0,3])* = V(X=[0,3]) + [G - V(X=[0,3])]/N(X=[0,3]) = 0 + [1 - 0]/1 = 1

- X, r = τ[-2] => [0, 2], -0.04
- $G_{k-1} = \gamma * G_k + R_k = 0.99 * 1.0 + (-0.04) = 0.95 = \gamma * R_{k+1} + R_k$
- *If X=[0,2] is never visited before (not present in remaining list):*
  - N(X=[0,2]) += 1
  - V(X=[0,2])* = V(X=[0,2]) + [G - V(X=[0,2])]/N(X=[0,2]) = 0 + [0.95 - 0]/1 = 0.95

- X, r = τ[-3] => [0, 1], -0.04
- $G_{k-2} = \gamma * G_{k-1} + R_{k-1} = 0.99 * 0.95 + (-0.04) = 0.9005 = \gamma^2 * R_{k+1} + \gamma * R_k + R_{k-1}$
- *If X=[0,1] is never visited before (not present in remaining list):* ***False***

- X, r = τ[-4] => [0, 1], -0.04
- $G_{k-3} = \gamma * G_{k-2} + R_{k-2} = 00.99 * 0.9005 + (-0.04) = 0.851495 = \gamma^3 * R_{k+1} + \gamma^2 * R_k + \gamma * R_{k-1} + R_{k-2}$
- *If X=[0,1] is never visited before:*
  - N(X=[0,1]) += 1
  - V(X=[0,1])* = V(X=[0,1]) + [G - V(X=[0,1])]/N(X=[0,1]) = 0 + [0.9005 - 0]/1 = 0.851495

- X, r = τ[-5] => [0, 0], -0.04
- $G_{k-4} = \gamma * G_{k-3} + R_{k-3} = 0.99 * 0.851495 + (-0.04) = 0.80298 = \gamma^4 * R_{k+1} + \gamma^3 * R_k + \gamma^2 * R_{k-1} + \gamma * R_{k-2} + R_{k-3}$
- *If X=[0,0] is never visited before:*
  - N(X=[0,0]) += 1
  - V(X=[0,0])* = V(X=[0,0]) + [G - V(X=[0,0])]/N(X=[0,0]) = 0 + [0.80298 - 0]/1 = 0.80298

- X, r = τ[-6] => [1, 0], -0.04
- $G_{k-5} = \gamma * G_{k-4} + R_{k-4} = 0.99 * 0.80298 + (-0.04) = 0.70740075$
- *If X=[1,0] is never visited before (not present in remaining list):* ***False***

Before J=0 started

V(0,0)=0 N(0,0)=0	V(0,1)=0 N(0,1)=0	V(0,2)=0 N(0,2)=0	V(0,3)=0 N(0,3)=0
V(1,0)=0 N(1,0)=0	V(1,1)=0 N(1,1)=0	V(1,2)=0 N(1,2)=0	V(1,3)=0 N(1,3)=0
V(2,0)=0 N(2,0)=0	V(2,1)=0 N(2,1)=0	V(2,2)=0 N(2,2)=0	V(2,3)=0 N(2,3)=0

After J=0 finished

V(0,0)=0.80298 N(0,0)=1	V(0,1)=0.851495 N(0,1)=1	V(0,2)=0.95 N(0,2)=1	V(0,3)=1 N(0,3)=1
V(1,0)=0.70740075 N(1,0)=1	V(1,1)=0 N(1,1)=0	V(1,2)=0 N(1,2)=0	V(1,3)=0 N(1,3)=0
V(2,0)=0.66032674 N(2,0)=1	V(2,1)=0 N(2,1)=0	V(2,2)=0 N(2,2)=0	V(2,3)=0 N(2,3)=0

After J=1 finished

V(0,0)=0.827237525 N(0,0)=2	V(0,1)=0.875997 5 N(0,1)=2	V(0,2)=0.95 N(0,2)=2	V(0,3)=1 N(0,3)=2
V(1,0)=0.7551904 N(1,0)=2	V(1,1)=0 N(1,1)=0	V(1,2)=0 N(1,2)=0	V(1,3)=0 N(1,3)=0
V(2,0)=0.707638495 N(2,0)=2	V(2,1)=0 N(2,1)=0	V(2,2)=0 N(2,2)=0	V(2,3)=0 N(2,3)=0

At, J = 1, τ = {([2, 0], -0.04), ([1, 0], -0.04), ([0, 0], -0.04), ([0, 1], -0.04), ([0, 2], -0.04), ([0, 3], 1.0) }

# G = 0, γ = 0.99
# V(2,0), V(1,0), V(0,0), V(0,1), V(0,2), V(0,3) will be updated based on the criterion,
V(x,y) ← V(x,y) + [G - V(x,y)]/N(x,y)

For length(τ):

- X, r = τ[-1] => [0, 3], 1.0
- G = γ * G + r = 0.99 * 0 + 1.0 = 1.0
- *If X=[0,3] is never visited before (not present in remaining list):*
  - N(X=[0,3]) += 1
  - V(X=[0,3])* = V(X=[0,3]) + [G - V(X=[0,3])]/N(X=[0,3]) = 1 + [1 - 1]/2 = 1

- X, r = τ[-2] => [0, 2], -0.04
- G = γ * G + r = 0.99 * 1.0 + (-0.04) = 0.95
- *If X=[0,2] is never visited before (not present in remaining list):*
  - N(X=[0,2]) += 1
  - V(X=[0,2])* = V(X=[0,2]) + [G - V(X=[0,2])]/N(X=[0,2]) = 0.95 + [0.95 - 0.95]/2 = 0.95

- X, r = τ[-3] => [0, 1], -0.04
- G = γ * G + r = 0.99 * 0.95 + (-0.04) = 0.9005
- *If X=[0,1] is never visited before:*
  - N(X=[0,1]) += 1
  - V(X=[0,1])* = V(X=[0,1]) + [G - V(X=[0,1])]/N(X=[0,1]) = 0.851495 + [0.9005 - 0.851495]/2 = 0.8759975

- X, r = τ[-4] => [0, 0], -0.04
- G = γ * G + r = 0.99 * 0.9005 + (-0.04) = 0.851495
- *If X=[0,0] is never visited before:*
  - N(X=[0,0]) += 1
  - V(X=[0,0])* = V(X=[0,0]) + [G - V(X=[0,0])]/N(X=[0,0]) = 0.80298005 + [0.851495 - 0.80298005]/2 = 0.827237525

- X, r = τ[-5] => [1, 0], -0.04
- G = γ * G + r = 0.99 * 0.851495 + (-0.04) = 0.80298005
- *If X=[1,0] is never visited before:*
  - N(X=[1,0]) += 1
  - V(X=[1,0])* = V(X=[1,0]) + [G - V(X=[1,0])]/N(X=[1,0]) = 0.70740075 + [0.80298005 - 0.70740075]/2 = 0.7551904

- X, r = τ[-6] => [2, 0], -0.04
- G = γ * G + r = 0.99 * 0.80298005 + (-0.04) = 0.75495025
- *If X=[2,0] is never visited before:*
  - N(X=[2,0]) += 1
  - V(X=[2,0])* = V(X=[2,0]) + [G - V(X=[2,0])]/N(X=[2,0]) = 0.66032674 + [0.75495025 -