Define:

$S = \{(p_1, p_2, \theta_2)\}$ for player 1 position p_1, player 2 position p_2 and player 2 type θ_2

$S_R = \{(x_{red}, y_{red})\}$

$S_G = \{(x_{green}, y_{green})\}$

$A_1 = \{\phi, N, E, W, S, Z\}$

$A_2 = \{\phi, N, E, W, S\}$

$p_1 \coloneqq (x_1, y_1)$

$p_2 \coloneqq (x_2, y_2)$

Transition: $(x_i^{t+1}, y_i^{t+1}) = \begin{cases} (x_i^t, y_i^t) & a_i^t \in \{\phi, Z\} \\ (x_i^t, y_i^t - 1) & a_i^t = N \\ (x_i^t + 1, y_i^t) & a_i^t = E \\ (x_i^t - 1, y_i^t) & a_i^t = W \\ (x_i^t, y_i^t + 1) & a_i^t = S \end{cases}$ if not out of bound or obstacles

else (x_i^t, y_i^t) for $i \in \{1,2\}$ (North and South may not seem intuitive but we are using Pygames' indexing where the upper left cell is $(0,0)$)

Neighbors $\nu(x, y) = \{(x-1, y-1), (x, y-1), (x+1, y-1), (x-1, y), (x, y), (x+1, y), (x-1, y+1), (x, y+1), (x+1, y+1)\}$

Costs: $B \in [0,1)$, $C \in [0,1)$

Initialize:

$Q_1(h, a_1) = 0$ for all $h \in H$, $a_1 \in A_1$

$Q_2(s, a_2) = 0$ for all $s \in S$, $a_2 \in A_2$

Repeat for every episode:

Player 1's posterior in last episode is prior in this episode

Player 2 chooses type $\theta_2 \in \{0,1\}$

For $t = 0,1,2, \ldots,$, do:

\# State

$p_1^t \coloneqq (x_1^t, y_1^t)$

$p_2^t \coloneqq (x_2^t, y_2^t)$

Observe current state $s^t = (p_1^t, p_2^t, \theta_2)$

Observe current history $h^t = (s^0, a^0, s^1, \ldots s^{t-1}, a^{t-1}, s^t)$

Set: $r_1 = 0, r_2 = 0$

Player 1

Action choice

$a_1^t \in \arg\max_{a_1} VI_1(a_1|h^t)$

$VI_1(a_1|h^t) = \sum_{x\in\{0,1\}} \mathbb{P}(\theta_2 = x|h^t) \sum_{a_2\in A_2} Q_1\big(h^t, (a_1, a_2)\big)\mathbb{P}(a_2|\theta_2 = x)$

$\mathbb{P}(\theta_2 = x|h^t) = \frac{\mathbb{P}\left(h^t\middle|\theta_2 = x\right)\mathbb{P}(\theta_2=x)}{\sum_{y\in\{0,1\}} \mathbb{P}\left(h^t\middle|\theta_2 = y\right)\mathbb{P}(\theta_2=y)}$

$\mathbb{P}(h^t|\theta_2 = x) = \prod_{\tau=0}^{t-1} \mathbb{P}(a_2^\tau|\theta_2 = x)$

P1 Transition and rewards

If $a_1^t = Z$:

If $\theta_2 = 0$ and $p_2^t \in \nu(p_1^t)$: $r_1 = -C, r_2 = -C$

Episode Ends

Else:

$p_1^{t+1} \leftarrow p_1^t$

Player 2

With probability ϵ: choose random action $a_2^t \in A_2$

Otherwise: choose action $a_2^t \in \arg\max_{a_2} Q_2(s^t, a_2)$

P2 Transition and rewards

$p_2^{t+1} \leftarrow p_2^t$

If $\theta_2 = 0$ and $p_2^{t+1} \in S_G$: $r_1 = 1, r_2 = 1$

If $\theta_2 = 1$ and $p_2^{t+1} \in S_R$: $r_1 = -B, r_2 = B$

Q-value update

Joint action $a^t = (a_1^t, a_2^t)$

Transition $s^{t+1} = (p_1^{t+1}, p_2^{t+1}, \theta_2)$

$$Q_1\big(h^t, (a_1, a_2)\big) = \sum_{s' \in S} P\big(s' \big| s^t, (a_1, a_2)\big) \left[r_1 + \gamma \max_{a_1' \in A_1} VI_1(a_1' | \langle h^t, (a_1, a_2), s' \rangle\right]$$

$$Q_2(s^t, a_2^t) \leftarrow Q_2(s^t, a_2^t) + \alpha \left[r_2 + \gamma \max_{a_2'} Q_2(s^{t+1}, a_2') - Q_2(s^t, a_2^t)\right]$$

History update

$h^{t+1} = \langle h^t, a^t, s^{t+1} \rangle$