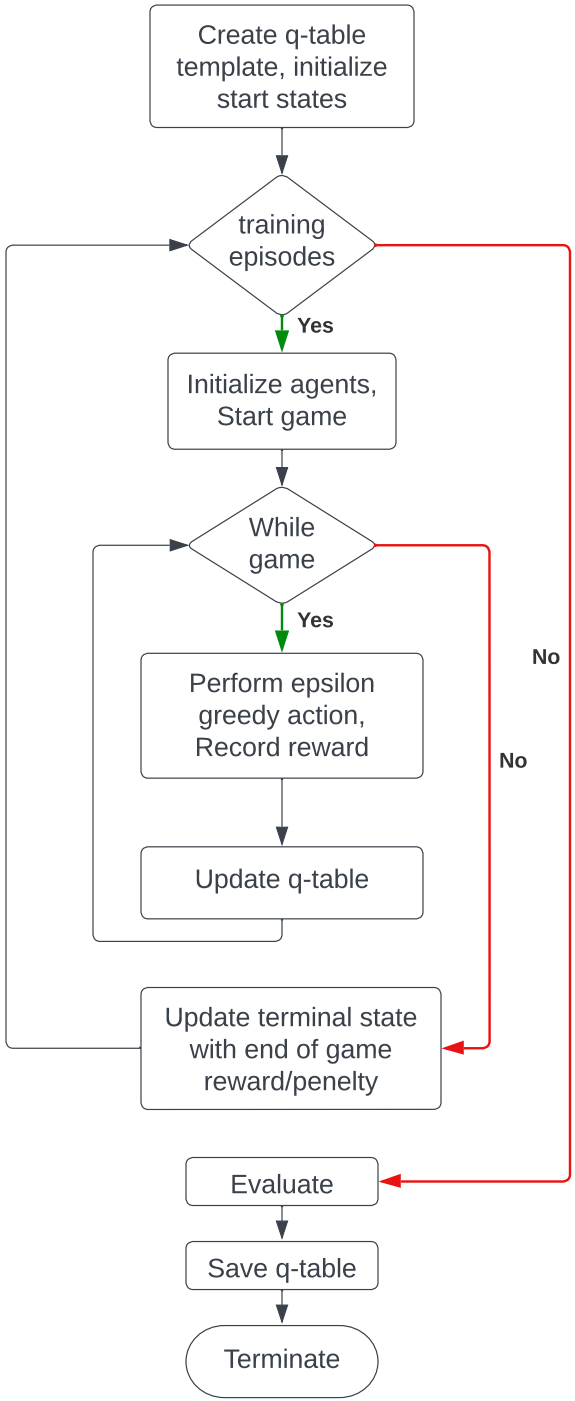

Update Rule: $q(s_t, a) = (1 - \alpha)q(s_t, a) + \alpha\gamma max_{a'}[q(s_{t+2}, a')] + \alpha \sum_{i=0}^{1} r_{t+i}$

Exploration decay:

Let p = epsilon at the end of training
N = training episodes
initial exploration = 1
n = current episode

$$\epsilon = \epsilon_{initial}\, d^n \quad d = \exp\left(\frac{\ln(p)}{N}\right)$$

Rewards and Penelties:

During gameplay:

+1 reward for correct firework played

At terminal state:

+10 reward if all 5 fireworks have been played

-10 penelty of 0 correct cards have been played

State representation:

i = # of hint tokens
l = # of life tokens
f = # current firework
d = up to 10D vector of discarded cards
h1 = 2D vector of your own hand knowledge
h2 = 2D vector of opponent hand knowledge
o2 = 2D vector of opponent hand

Note that the total length of the state vector is 19, however, some symmetries can be exploited to reduce the total number of states (i.e, d is ordered)

$$s^{19} = [i, l, f, d^{10}, h_1^2, o_2^2, h_2^2]$$