Algorithm 1 RUN-EPISODE (Q, π^h, π^r)

INPUT: State space S, Action space A

Initialize environment to start state; $R \leftarrow 0$; $i \leftarrow 0$;
loop
  if goal state **then**
    return Q-functions, R
  end if
  Sample a_h from π^h and a_r from π^r
  Take action (a_h, a_r) and observe r, $\mathbf{s}'$
  Update Q-functions as in Eq. 2
  $R \leftarrow R + \gamma^i * r$
  $i \leftarrow i + 1$
end loop