**First idea:**

**States**: For the agent, the definition of the state will coincide with the jobs that have finished and the order in which they were done, that is, the sequence that has been built. This means that at the beginning of an episode, the state of the environment is empty, since no work has been added to the sequence; therefore, at the end of the episode the state of the environment will be given by all the sequenced works.

**Actions**:

* Action 1: pick job 1
* Action 2: pick job 2
* Action 3: pick job 3
* …

**Rewards**: As described, the state is represented by the sequence that has been built, an action constitutes inserting a new job in said sequence and taking into account that our final objective is to minimize the makespan; We define reward 1/𝑚𝑎𝑘𝑒𝑠𝑝𝑎𝑛, note that the greater the makespan of the sequence, the lower the reward for the selected action. It should be noted then that the highest Q-value determines the best action in any state.

**Example**:

The problem solving process is illustrated below through a small example of only three jobs and two machines.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Machine 1 | Machine 2 | Total processing time |
| Job 1 | 6 | 1 | 7 |
| Job 2 | 2 | 4 | 6 |
| Job 3 | 3 | 2 | 5 |
| Job 4 | 5 | 2 | 7 |
| Job 5 | 4 | 6 | 10 |
| Job 6 | 3 | 5 | 8 |

Action 1: pick job 1

Action 2: pick job 2

Action 3: pick job 3

…

**Episode 1:**

* The agent observes the state of the environment 𝑠 = {}.
* Evaluate the set of possible actions
* Suppose that a random action, in this case the option 4 𝑎 = 𝐽4:1.
* Perform the selected action:
* 𝑠′ = {𝐽4}
* 𝑟 = 1⁄makespan 𝑠 ′ = 1⁄5 = 0.2

**Iteration 2:**

Action 1: pick job 1

Action 2: pick job 2

Action 3: pick job 3

…

* 𝑠 = {𝐽4}, actions = {𝐽1: 2,𝐽2: 2, J3:2, 𝐽5:2, 𝐽6:2}.
* Suppose that a random action, in this case the action 3.
* Perform the selected action:

𝑠′ = {𝐽4, 𝐽3}𝑟 = 1⁄makespan 𝑠′ = 1⁄10 = 0.1

**Second Idea:**

**States**: For the agent, the definition of the state will coincide with the jobs that have finished and the order in which they were done, that is, the sequence that has been built. This means that at the beginning of an episode, the state of the environment is empty, since no work has been added to the sequence; therefore, at the end of the episode the state of the environment will be given by all the sequenced works.

**Actions**: the taking of an action by the agent is equivalent to deciding in which position of the already constructed sequence the work will be introduced.

**Rewards**: As described, the state is represented by the sequence that has been built, an action constitutes inserting a new job in said sequence and taking into account that our final objective is to minimize the makespan; We define reward 1/𝑚𝑎𝑘𝑒𝑠𝑝𝑎𝑛, note that the greater the makespan of the sequence, the lower the reward for the selected action. It should be noted then that the highest Q-value determines the best action in any state.

**Example**:

The problem solving process is illustrated below through a small example of only three jobs and two machines.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Machine 1 | Machine 2 | Total processing time |
| Job 1 | 6 | 1 | 7 |
| Job 2 | 2 | 4 | 6 |
| Job 3 | 3 | 2 | 5 |
| Job 4 | 5 | 2 | 7 |
| Job 5 | 4 | 6 | 10 |
| Job 6 | 3 | 5 | 8 |

# initial sequence is 3,2,1,4,6,5

**Episode 1:**

* The agent observes the state of the environment 𝑠 = {}.
* Evaluate the set of possible actions
* Suppose that a random value less than ε is generated and therefore a random action, in this case there is only the option 𝑎 = 𝐽3:1.
* Perform the selected action:
* 𝑠′ = {𝐽3}
* 𝑟 = 1⁄makespan 𝑠 ′ = 1⁄8 = 0.125

**Iteration 2:**

* 𝑠 = {𝐽3}, actions = {𝐽2: 1,𝐽2: 2}.
* Suppose that a random value less than ε is generated and therefore a random action, in this case the option 1.
* Perform the selected action:

𝑠′ = {𝐽2, 𝐽3}𝑟 = 1⁄makespan 𝑠′ = 1⁄10 = 0.1

**Iteration 3:**

* 𝑠 = { 𝐽2, 𝐽3}, actions = {𝐽1: 1,𝐽1: 2, 𝐽1: 3}.
* Suppose that a random value less than ε is generated and therefore a random action, in this case the option 2.
* Perform the selected action:

𝑠′ = {𝐽2, J1, 𝐽3}𝑟 = 1⁄makespan 𝑠′ = 1⁄20 = 0.05