Algorithm 1 Finite Horizon Minimax Q-Learning

Notation:

Max player and Min player: Microgrid M1 and M2 respectively

$\mathbf{U}, \mathbf{V}$: Strategy/action sets of max and min player respectively

$\mathbf{u}, \mathbf{v}$: Action taken by max and min player respectively; $u \in U, v \in V$

$\mathbf{Q_n^m(i, u, v)}$: Q-value at state i, action pair u, v, stage n and recursion m.

$\mathbf{a(m)}$: step-size at recursion index m, which is set to $\lceil \frac{m+1}{10} \rceil$ in experiments.

$\mathbf{Q_N(i, u, v)}$: Q-value for state i and action pair u, v at terminal stage (N).

$\mathbf{r_n(i, u, v)}$: payoff matrix at stage n and state i indexed by u, v, same as $\mathbf{r(i, u, v)}$, we assume same function across all stages.

$\mathbf{r_N(i)}$: Payoff at the N^{th} stage (terminal stage) when terminal state is i (of size $|U| \times |V|$)

val: $A \in R^{m \times n}, val[A] = \min_y \max_x x^T A y$

Initialization: $Q_n^0(i, u, v) = 0$, $\forall (i, u, v)$,
$n = 0, \ldots, N-1$, and
$Q_N^0(i, u, v) = r_N(i), \forall (i, u, v)$

Input: Samples of the form,
$\big(n$ (current stage)$, i$ (current state), u, v (action pair), r (payoff), j (next state)$\big)$.

$\mathbf{Q^m}$: estimate of Q-values at current iteration m

Output: Updated Q-value $Q_n^{m+1}(i, u, v)$ estimated after $m+1$ iterations of the algorithm.

1: $Q_n^{m+1}(i, u, v) = \big(1 - a(m)\big)\Big(Q_n^m(i, u, v)\Big) + a(m)$

2: $\times\Big(r\big(i, u, v\big) + val[Q_{n+1}^m\big(j\big)]\Big), n = 0, 1, \ldots, N-1$

3: $Q_N^m(i, u, v) = r_N(i)$

4: **return** $Q_n^{m+1}(i, u, v)$