.

Georgios Frangias

August 2023

Algorithm 1 On-policy SLATEQ for Live Experiments

**Parameters**:

- $T$ : the number of iterations.
- $M$ : the interval to update label network.
- $\gamma$ : discount rate.
- $\theta_{\text{main}}$ : the parameter for the main neural network.
- $\bar{Q}_{\text{main}}$ : that predicts items' long-term value.
- $\theta_{\text{label}}$ : the parameter for the label neural network $\bar{Q}_{\text{label}}$ .
- $\theta_{\text{pctr}}$ : the parameter for the neural network that predicts items' pCTR.

**Input**:

- $D_{\text{training}} = \left(s, A, C, L_{\text{myopic}}, s', A'\right)$ : the training data set.
- $s$ : current state features
- $A = (a_1, \ldots, a_k)$ : recommended slate of items in current state; $a_i$ denotes item features
- $C = (c_1, \ldots, c_k)$ : $c_i$ denotes whether item $a_i$ is clicked
- $L_{\text{myopic}} = \left(l^1_{\text{myopic}}, \ldots, l^k_{\text{myopic}}\right)$ : myopic (immediate) labels
- $s'$ : next state features
- $A' = \left(a'_1, \ldots, a'_k\right)$ : recommended slate of items in next state.

**Output**: Trained Q-network $\bar{Q}_{\text{main}}$ that predicts items' long-term value.
**Initialization**: $\theta_{\text{label}} = 0, \theta_{\text{main}}$ randomly, $\theta_{pctr}$ randomly
**for** $i = 1 \ldots T$ **do**
  **if** $i \bmod M = 0$ **then**
    $\theta_{\text{label}} \leftarrow \theta_{\text{main}}$
  **for** each example $\left(s, A, C, L_{\text{myopic}}, s', A'\right) \in D_{\text{training}}$ **do**
    **for** each item $a_i \in A$ **do**
      update $\theta_{\text{pctr}}$ using click label $c_i$
      **if** $a_i$ is clicked **then**
        probability: $p\,\text{CTR}\left(s', a'_i, A'\right) \leftarrow p\,\text{CTR}\left(s', a'_i\right) / \sum_{a'_i \in A} p\,\text{CTR}\left(s', a'_i\right)$
        LTV label: $l^i_{\text{ltv}} \leftarrow l^i_{\text{myopic}} + \sum_{a'_i \in A'} p\,\text{CTR}\left(s', a'_i, A'\right) \bar{Q}_{\text{label}}\ \left(s', a'_i\right)$
        update $\theta_{\text{main}}$ using LTV label $l^i_{ltv}$

Taken from arXiv:1905.12767