Episodic Semi-gradient Sarsa for Estimating $\hat{q} \approx q_*$

Input: a differentiable action-value function parameterization $\hat{q} : \mathcal{S} \times \mathcal{A} \times \mathbb{R}^d \to \mathbb{R}$
Algorithm parameters: step size $\alpha > 0$, small $\varepsilon > 0$
Initialize value-function weights $\mathbf{w} \in \mathbb{R}^d$ arbitrarily (e.g., $\mathbf{w} = \mathbf{0}$)

Loop for each episode:
   $S, A \leftarrow$ initial state and action of episode (e.g., ε-greedy)
   Loop for each step of episode:
      Take action A, observe R, S'
      If S' is terminal:
         $\mathbf{w} \leftarrow \mathbf{w} + \alpha \big[R - \hat{q}(S, A, \mathbf{w}) \big] \nabla \hat{q}(S, A, \mathbf{w})$
         Go to next episode
      Choose A' as a function of $\hat{q}(S', \cdot, \mathbf{w})$ (e.g., ε-greedy)
      $\mathbf{w} \leftarrow \mathbf{w} + \alpha \big[R + \gamma \hat{q}(S', A', \mathbf{w}) - \hat{q}(S, A, \mathbf{w}) \big] \nabla \hat{q}(S, A, \mathbf{w})$
      $S \leftarrow S'$
      $A \leftarrow A'$