Algorithm 1 Soft Q-learning

$\theta, \phi \sim$ some initialization distributions.
Assign target parameters: $\bar{\theta} \leftarrow \theta$, $\bar{\phi} \leftarrow \phi$.
$\mathcal{D} \leftarrow$ empty replay memory.

for each epoch **do**
  for each t **do**
    Collect experience
      Sample an action for $\mathbf{s}_t$ using f^{ϕ}:
        $\mathbf{a}_t \leftarrow f^{\phi}(\xi; \mathbf{s}_t)$ where $\xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$.
      Sample next state from the environment:
        $\mathbf{s}_{t+1} \sim p_{\mathbf{s}}(\mathbf{s}_{t+1} | \mathbf{s}_t, \mathbf{a}_t)$.
      Save the new experience in the replay memory:
        $\mathcal{D} \leftarrow \mathcal{D} \cup \{(\mathbf{s}_t, \mathbf{a}_t, r(\mathbf{s}_t, \mathbf{a}_t), \mathbf{s}_{t+1})\}$.
    Sample a minibatch from the replay memory
      $\{(\mathbf{s}_t^{(i)}, \mathbf{a}_t^{(i)}, r_t^{(i)}, \mathbf{s}_{t+1}^{(i)})\}_{i=0}^{N} \sim \mathcal{D}$.
    Update the soft Q-function parameters
      Sample $\{\mathbf{a}^{(i,j)}\}_{j=0}^{M} \sim q_{\mathbf{a}'}$ for each $\mathbf{s}_{t+1}^{(i)}$.
      Compute empirical soft values $\hat{V}_{\text{soft}}^{\bar{\theta}}(\mathbf{s}_{t+1}^{(i)})$ in (10).
      Compute empirical gradient $\hat{\nabla}_{\theta} J_Q$ of (11).
      Update θ according to $\hat{\nabla}_{\theta} J_Q$ using ADAM.
    Update policy
      Sample $\{\xi^{(i,j)}\}_{j=0}^{M} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ for each $\mathbf{s}_t^{(i)}$.
      Compute actions $\mathbf{a}_t^{(i,j)} = f^{\phi}(\xi^{(i,j)}, \mathbf{s}_t^{(i)})$.
      Compute Δf^{ϕ} using empirical estimate of (13).
      Compute empiricial estimate of (14): $\hat{\nabla}_{\phi} J_{\pi}$.
      Update ϕ according to $\hat{\nabla}_{\phi} J_{\pi}$ using ADAM.
  end for
  if epoch *mod* update_interval $= 0$ **then**
    Update target parameters: $\bar{\theta} \leftarrow \theta$, $\bar{\phi} \leftarrow \phi$.
  end if
end for