

Algorithm 1 Double DQN with proportional prioritization

- 1: **Input:** minibatch k , step-size η , replay period K and size N , exponents α and β , budget T .
- 2: Initialize replay memory $\mathcal{H} = \emptyset$, $\Delta = 0$, $p_1 = 1$
- 3: Observe S_0 and choose $A_0 \sim \pi_\theta(S_0)$
- 4: **for** $t = 1$ **to** T **do**
- 5: Observe S_t, R_t, γ_t
- 6: Store transition $(S_{t-1}, A_{t-1}, R_t, \gamma_t, S_t)$ in \mathcal{H} with maximal priority $p_t = \max_{i < t} p_i$
- 7: **if** $t \equiv 0 \pmod K$ **then**
- 8: **for** $j = 1$ **to** k **do**
- 9: Sample transition $j \sim P(j) = p_j^\alpha / \sum_i p_i^\alpha$
- 10: Compute importance-sampling weight $w_j = (N \cdot P(j))^{-\beta} / \max_i w_i$
- 11: Compute TD-error $\delta_j = R_j + \gamma_j Q_{\text{target}}(S_j, \arg \max_a Q(S_j, a)) - Q(S_{j-1}, A_{j-1})$
- 12: Update transition priority $p_j \leftarrow |\delta_j|$
- 13: Accumulate weight-change $\Delta \leftarrow \Delta + w_j \cdot \delta_j \cdot \nabla_\theta Q(S_{j-1}, A_{j-1})$
- 14: **end for**
- 15: Update weights $\theta \leftarrow \theta + \eta \cdot \Delta$, reset $\Delta = 0$
- 16: From time to time copy weights into target network $\theta_{\text{target}} \leftarrow \theta$
- 17: **end if**
- 18: Choose action $A_t \sim \pi_\theta(S_t)$
- 19: **end for**