Algorithm 1: ABR Model Training

```
Initialize critic networks $Q_{\theta_1}$, $Q_{\theta_2}$, actor network $\pi_\phi$
Initialize target networks $Q_{\theta_1^{tar}}$, $Q_{\theta_2^{tar}}$,$\pi_{\phi^{tar}}$
Initialize replay buffer $\mathcal{D}$, mini-batch $\Omega$
Initialize soft update factor $\tau$ ,delay update para $\delta$
for each video session do
    for segment $t$ = 1 to $T$ do
        Select action (bitrate): $a_t = \pi_\phi(s_t)$
        Observe reward (QoE) $r_t$, and done $d_t$
        Store tuple $(s_t, a_t, r_t, d_t)$ in $\mathcal{D}$
        Sample a mini-batch of $j$ sequences, each
          with $n$ consecutive steps from $\mathcal{D}$:
        $(s_{t'_0+i}, a_{t'_0+i}, r_{t'_0+i}, d_{t'_0+i})_{i=0,1,2,\ldots,n-1}$
        $(s_{t'_1+i}, a_{t'_1+i}, r_{t'_1+i}, d_{t'_1+i})_{i=0,1,2,\ldots,n-1}$
        $\vdots$
        $(s_{t'_{j-1}+i}, a_{t'_{j-1}+i}, r_{t'_{j-1}+i}, d_{t'_{j-1}+i})_{i=0,1,2,\ldots,n-1}$
        $R, m = \text{ComputeCumulativeReward}(n)$
        $Q_{tar} = \text{ComputeTargetQValue}(R, m, n)$
        Update $Q_{\theta_1}$, $Q_{\theta_2}$ by minimizing:
        $E_\Omega[(Q_{\theta_{i=1,2}}(s_{t'}, a_{t'}) - Q_{tar})^2]$
        if $t$ mod $\delta == 0$ then
            Update $\pi_\phi$ by maximizing:
            $E_\Omega[Q_{\theta_1}(s_{t'}, \pi_\phi(s_{t'}))]$
            Softly update target networks:
            $\phi^{tar} = \tau\phi^{tar} + (1-\tau)\phi$
            $\theta_1^{tar} = \tau\theta_1^{tar} + (1-\tau)\theta_1$
            $\theta_2^{tar} = \tau\theta_2^{tar} + (1-\tau)\theta_2$
Function ComputeCumulativeReward($n$):
    Initialize cumulative reward: $R = 0$
    Initialize video terminal tag: $m = n$
    for $k$ = 0 to $n$-1 do
        Update cumulative reward:
        $R = R + \gamma^k r_{t'+k}$
        if $d_{t'+k} == 1$ then
            Mark the termination step: $m = k$
            return $R, m$
    return $R, m$
Function ComputeTargetQValue($R, m, n$):
    if $m < n$ then
        return $R$
    else
        Calculate target bitrate with Gaussian noize:
        $a_{tar} = \pi_{\phi^{tar}}(s_{t'+n}) + \mathcal{N}(\mu, \sigma^2)$
        Calculate minimum target Q value:
        $Q_{tar_1} = Q_{\theta_1^{tar}}(s_{t'+n}, a_{tar})$
        $Q_{tar_2} = Q_{\theta_2^{tar}}(s_{t'+n}, a_{tar})$
        $min\_value = \min(Q_{tar_1}, Q_{tar_2})$
        return $R + \gamma^n \times min\_value$
```