Algorithm 1: SAC-DRND (Online)

Input: Number of critic updates G, number of DRND updates per iteration K, target update coefficient ρ, entropy coefficient β, DRND coefficients $\lambda_{\text{actor}}, \lambda_{\text{critic}}$, replay buffer $\mathcal{D}$

Initialization: Initialize policy parameters ϕ, Q-function parameters φ_1, φ_2 and targets φ_1', φ_2', predictor network θ and targets $\{\theta_n\}_{n=1}^{N}$

Reset environment and observe initial state s_0;
for *each iteration* **do**
  Reset environment and observe initial state s_0;
  while *not terminal* **do**
    Sample action $a_t \sim \pi_\phi(\cdot|s_t)$;
    Execute a_t, observe r_t, s_{t+1};
    Store $(s_t, a_t, r_t, b(s_{t+1}, a_t), s_{t+1})$ in $\mathcal{D}$;
    $s_t \leftarrow s_{t+1}$;
  end
  for $k = 1$ **to** K **do**
    Sample minibatch $(s, a, r, b, s') \sim \mathcal{D}$;
    Sample $n \sim Uniform(1 \dots N)$;
    Update predictor network θ using:

$$\nabla_\theta \frac{2}{|B|} \sum (f_\theta(s,a) - f_{\theta_n}(s,a))$$

  end
  for $g = 1$ **to** G **do**
    Sample minibatch $(s, a, r, b, s') \sim \mathcal{D}$;
    Update actor using:

$$\nabla_\phi \frac{1}{|B|} \sum_{(s,a)} \left[\min_{i=1,2} Q_{\varphi_i}(s, \tilde{a}_\phi(s)) - \beta \log \pi_\phi(\tilde{a}_\phi(s)|s) - \lambda_{\text{actor}} b(s, \tilde{a}_\phi(s)) \right]$$

    where $\tilde{a}_\phi(s) \sim \pi_\phi(\cdot|s)$ via reparameterization;
    Update each critic Q_{φ_i} using:

$$\nabla_{\varphi_i} \frac{1}{|B|} \sum_{(s,a,r,s')} \Big[Q_{\varphi_i}(s,a) - \Big(r + \gamma E_{a' \sim \pi_\phi(\cdot|s')} \Big[Q_{\varphi_i'}(s',a') - \beta \log \pi_\phi(a'|s') - \lambda_{\text{critic}} b(s',a') \Big] \Big) \Big]^2$$

    Update target critics:

$$\varphi_i' \leftarrow (1-\rho)\varphi_i' + \rho \varphi_i$$

  end
end